Under Sampling Techniques for Handling Unbalanced Data with Various Imbalance Rates: A Comparative Study

Esraa Abu Elsoud[1], Mohamad Hassan[2], Omar Alidmat[3], Esraa Al Henawi[4],
Nawaf Alshdaifat[5], Mosab Igtait[6], Ayman Ghaben[7], Anwar Katrawi[8], Mohmmad Dmour[9]
Department of Computer Science-Faculty of Information Technology, Zarqa University, Zarqa, Jordan[1,2,3,4]
Faculty of Information Technology, Applied Science Private University, Amman, Jordan[5]
Department of Data Science and Artificial Intelligence, Zarqa University, Zarqa, Jordan[6,8]
Department of Cyber Security-Faculty of Information Technology, Zarqa University, Zarqa, Jordan[7]
Department of Computer Science-Faculty of Information Technology, Zarqa University, Zarqa, Jordan[9]

***Abstract*—Unbalanced data sets represent data sets that contain an unequal number of examples for different classes. This dataset represents a problem faced by machine learning tools; as in datasets with high imbalance ratios, false negative rate percentages will be increased because most classifiers will be affected by the major class. Choosing specific evaluation metrics that are most informative and sampling techniques represent a common way to handle this problem. In this paper, a comparative analysis between four of the most common under-sampling techniques is conducted over datasets with various imbalance rates (IR) range from low to medium to high IR. Decision Tree classifier and twelve imbalanced data sets with various IR are used for evaluating the effects of each technique depending on Recall, F1-measure, gmean, recall for minor class, and F1-measure for minor class evaluation metrics. Results demonstrate that Clusters Centroid outperformed Neighborhood Cleaning Rule (NCL) based on recall for all low IR datasets. For both medium, and high IR datasets NCL, and Random Under Sampling (RUS) outperformed the rest techniques, while Tomek Link has the worst effect.**

Keywords*—*Clusters centroid; decision tree; neighborhood cleaning rule; random under sampling; Tomek Link under sampling; unbalanced datasets

I. Introduction

Machine learning and statistics have been used for classification in many fields such as security [1]–[8], medical [9]–[15], text classification [16]–[18], and others [19], [20]. Classifications are defined as building a training model based on previous experiences or examples. Recently, there has been a rapid growth in data that has been collected from different environments but, unfortunately, there is a lack of quality data.

The quality of the data means a balanced distribution for all classes, the range of values is normalized, and no missing values, and so on. The point is, that several traditional machine learning techniques assumed that the target classes have a balanced distribution in the data [21]–[23].

Multi-class classifiers such as Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), and others are sensitive for imbalanced class distribution problems [24] while one class classifiers such as Isolation forest, local outlier factor (LOF), OC-SVM and other were not.

Due to the fact that, it is rare to find balanced datasets, that contain equal or nearly equal numbers of instances for each class in real-life classification problems; Building classification models under highly imbalanced datasets is an issue in machine learning algorithms.

In unbalanced datasets the most important class (class of interest) has fewer examples than other classes like rare disease datasets [25], therefore, the classification performance of the classifier will be affected by skewing to the majority class instances, which is usually not class of interest [26].

There are two main approaches for handling unbalanced dataset problems: "algorithm-driven approach", and "data-driven approach". The first approach concerns adjusting the classifier to improve its learning from the minority class samples [27].

On the other hand, the second approach concentrates on changing the data distribution in two ways either by adding new minor class examples (over-sampling) or removing some major class instances (under-sampling). Each way has its own advantages and disadvantages, where over-sampling is considered more overhead than under-sampling and can lead to overfitting problems while under-sampling causes the loss of important information [27].

In the literature, most researchers either propose some under-sampling techniques and commonly use these techniques in research, or compare these techniques using a specific dataset.

To our knowledge, there does not exist any research in the literature comparing the effects of specific under sampling techniques in the classification performance over datasets with low, medium, and high IR. This notice represents the rationale behind this paper where the main goal from this paper is to compare the influence of four common under sampling techniques called Tomek Link, NCL, Clusters Centroid, and Random Under Sampling (RUS) in the classification results for different datasets with various IR ranges from low to high IR datasets.

To accomplish this goal twelve datasets from the Keel collection with different IR variate between low, medium, and high IR have been used for comparing the performance of the decision tree classifier using each one of the previous four under-sampling techniques based on average recall, average F1 measure, gmean, minor class recall, and minor class F1 measure.

The rest of this paper is organized as follows: Section II displays the most recent studies about handling imbalance datasets in the literature. Section III-A describes Tomek Link, NCL, Clusters Centroid, and RUS under-sampling techniques that are compared. methodology has been demonstrated in Section III. In Section III-B the datasets are displayed. In Section IV results are presented. Finally, Section V contains the conclusion.

II. Related Work

A large number of domains with significant environmental, vital, or commercial importance encounter the class imbalance problem. The class imbalance problem means that there is a majority of one or more class spreads in the datasets [28], [29]. Moreover, it has been shown in some instances to significantly impede the performance achievable by conventional learning techniques that assume a balanced distribution of the classes and produce biased classifiers. Also, it degrades the performance of machine learning classifiers [30].

Many proposals have been presented in the literature to solve the imbalanced dataset. One of the most well-known techniques is the cluster-based under-sampling approach. It has been widely used to solve the imbalance of class distribution. In [31]–[35] a cluster-based under-sampling approach has been used to select the representative data as training data. Thus, the classification accuracy for minority classes will be improved. The experimental results show that the proposed approach outperforms other under-sampling techniques in the previous studies.

Random Under Sampling (RUS) is considered an under-sampling technique that is used for class imbalance problems. Many proposals used RUS to maintain a balanced class distribution [36]–[38]. In [39] a combination of T-Link and, Synthetic Minority Technique (SMOTE) and another sampling method such as RUS, and ROS in order to produce balance data.

Additionally, RUS has been used with different ratios to detect the performance of some of the machine learning classifiers as [40] eight random undersampling (RUS) ratios which are no sampling, 999:1, 99:1, 95:5, 9:1, 3:1, 65:35, and 1:1 have been used. Moreover, to show the performance of these ratios seven different classifiers are employed which are LightGBM (LGB), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR) CatBoost (CB), and XGBoost (XGB).

In [41] an ensemble feature selection has been proposed to classify the attack using the BoT-IoT dataset. The proposed approach is centered on the building of predictive models that are based on different classifiers. RUS has been used to solve the imbalance BoT-IoT dataset. The results show that the best RUS ratio was 1:1 or 1:3.

In [42] a new hybrid under sampling-based ensemble approach (HUSBoost) has been proposed. The main objective of HUSBoost is to handle imbalanced data using three main steps which are data cleaning, data balancing, and classification. At first, we remove the noisy data using Tomek-Links. RUS has been applied to create several balanced subsets.

The neighborhood cleaning rule (NCL) method has been used in many proposals in literature to deal with imbalance data [43]–[45] while other studies used hybrid approaches instead of NCL alone [46]. In [47] a combination of under-sampling and oversampling methods have been used to solve imbalance cases. Their proposal used is NCL under-sampling method and Adaptive Semi unsupervised Weighted Oversampling (A-SUWO) for the oversampling method.

Tomek link Tomek link technique is used in many studies to overcome the challenges of data imbalances that affect the performance of supervised learning-based [48]. In [49] Synthetic minority oversampling technique (SMOTE) and T-link have been used for imbalanced data. In addition, a Naïve Bayes classifier, support vector machine, and k-nearest neighbors together have been used for performance evaluation.

In [50] Cluster Based, Tomek Link, and Condensed Nearest neighbours have been used to handle the class imbalance problem by equalizing the number of instances. This is done by under-sampling the majority class based on some particular criteria [51]–[54]. The performance evaluation was done based on applied different machine learning classifiers such as K-Nearest Neighbor, Decision Tree, and Naive Bayes. The results showed that Decision Tree outperformed other machine learning techniques using the proposed technique.

Up to our best knowledge and based on an extensive literature review search, we noticed that most of the previous works exist in the literature compare the performance of specific under sampling techniques versus a hybrid version of these techniques over specific datasets with specific IR. However, in this paper, four of the most common under sampling techniques were applied over three categories of datasets that were categorized based on IR into three categories (low IR, medium IR, and high IR), and the effects of each technique on the classifier performance were compared.

The main purpose of this paper is to conclude a standard relationship between the compared under sampling techniques and dataset IR value in order to guide researchers in choosing the most suitable under sampling technique from the compared ones based on the dataset IR. In this paper, twelve imbalanced data sets with various IR have been used for evaluation in addition to DT for classification.

III. Methodology

For each dataset Work starts by dividing the original file into two parts: 70% for learning the classifier (training data set) and 30% for evaluating the effectiveness of the constructed model (testing set) based on average recall, average F1 measure, gmean, minor class recall, and minor class F1 measure [55].

For the training dataset, four under sampling techniques have been applied in order to make it balanced for learning decision tree classifier then using the generated model for classifying the test set and evaluating its performance, as shown in Fig. 1.

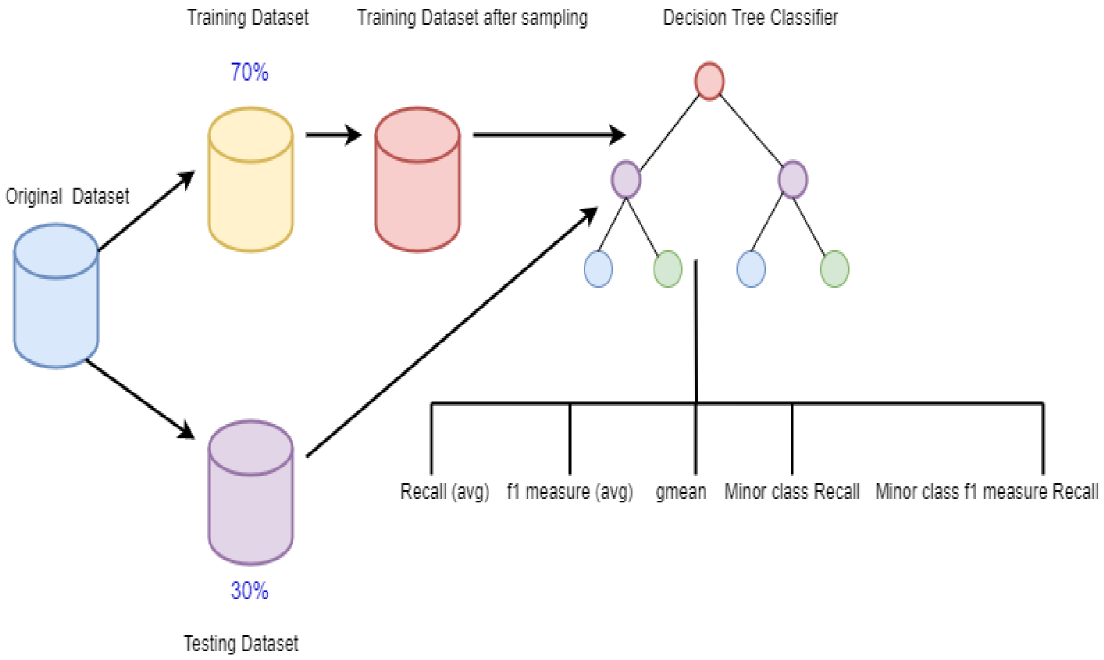

Fig. 1. Workflow for the used methodology

For each dataset, the best results that have been generated by DT classifier with each under sampling technique after adjusting its parameters are documented and compared.

In this paper, python is used for conducting all experiments through the PyCharm integrated development environment (IDE). Python libraries sci-kit-learn and imbalanced-learn are deployed for generating the results [56].

A. *Under Sampling Techniques*

1) Tomek link Tomek link refers to a pair of nearest neighbors where each one belongs to different classes. Under-sampling is done by removing all majority class samples from all Tomek Links [57].
2) Neighborhood Cleaning Rule (NCL) It was proposed by Laurikkala in 2001, where in this method for each sample in the training set three nearest neighbors for it must be defined then if it belongs to the major class while all of its selected neighbours belong to the minor class then it will be removed as a noise sample but if this sample belongs to the minor class and its three nearest neighbours belongs to the major class then these neighbours must be removed now. This method needs numerous computations with large-size datasets [58].
3) Clusters Centroid This method undersamples the majority class by replacing a cluster of majority samples as it finds the clusters of the majority class with the K-mean algorithm then it keeps the Cluster Centroids of the N clusters as the new majority samples [5].
4) Random Under Sampling (RUS) RUS works by removing some of the majority class samples randomly to change the distribution of data in the imbalanced dataset in order to convert it to a more balanced one for improving the classifier learning process in machine learning but this method sometimes means losing important information which considered as one drawback according to using this technique [59].

B. *Datasets*

Twelve datasets from Keel [8] collection with different IR are used in this paper. Table I summarizes these datasets properties.

Datasets are divided into three groups based on their IR, where the first five datasets with IR smaller than 9 represent a low IR group while the medium IR group contains datasets with IR greater than 9 and smaller than 50. Finally, datasets with IR greater than 50 are members of the high IR group in this paper.

C. *Evaluation Metrics*

This section is devoted to displaying the evaluation metrics that are used for evaluating the effects of TL, RUS, NCL, and CC under sampling techniques in the classifier performance for different datasets from various IR.

- Recall or TPR: It measures how often the classifier correctly detects the positive instances from all positive instances [60], [61], as shown in Eq. (1)

$$Recall = \frac{TP}{(TP + FN)} \tag{1}$$

- F1-Score: It combines the effects of precision and recall together [62], as shown in Eq. (2).

TABLE I. Properties of Twelve Imbalanced Datasets with Different IR from Keel Repository

Imbalance Category	Dataset Num.	Dataset Name	features	examples	Minor class	Major class	imbalance rate
Low	D1	glass1	9	214	76	138	1.82
	D2	ecoli_0_vs_1	7	220	77	139	1.86
	D3	vehicle0	18	846	199	647	3.25
	D4	ecoli3	7	336	35	301	8.6
	D5	page-blocks0	10	5472	559	4913	8.79
Medium	D6	glass4	9	214	13	201	15.47
	D7	car-good	6	1728	69	1659	24.04
	D8	kr-vs-k-one_vs_fifteen	6	2244	78	2166	27.77
High	D9	kr-vs-k-zero_vs_eight	6	1460	27	1433	53.07
	D10	Winquality	11	691	10	681	68.1
	D11	kr-vs-k-one_vs_fifteen	6	2193	27	2166	80.22
	D12	abalone19	8	4174	32	4142	129.44

$$F1 - Score = \frac{(2TP)}{(2TP + FP + FN)} \qquad (2)$$

- gmean: It is a combination of TPR, and TNR metrics, as shown in Eq. (3).

$$gmean = \sqrt{(TPR * TNR)} \qquad (3)$$

All metrics depend on four main parameters explained below:

- True Positive (TP): represents a number of actually positive instances classified as "positive".
- True Negative (TN): represents a number of actually negative instances classified as "negative".
- False Positive (FP): represents a number of actually negative instances classified as "positive".
- False Negative (FN): represents a number of actually positive instances classified as "negative".

IV. Experimental Results and Analysis

This section discusses the results that are generated from comparing the effects of Tomek Link, NCL, Clusters Centroid, and RUS in the performance of Decision Tree (DT) classifier for all groups of low, medium, and high IR datasets based on average Recall in subsection IV-A, then based on average F1 measure in subsection IV-B. Later the results of gmean, recall of minor class, and F1 measure of minor class results were discussed and analyzed in subsections IV-C, IV-D, and IV-E, consequently.

A. Recall

From Table II, we can conclude the following results by comparing the effects of the selected under sampling techniques in decision tree classifier average recall value for low, medium and high IR groups of datasets. From Fig. 2 we can concludes the following points for all low IR dataset groups recall value

- NCL provides better performance based on recall than Tomek link for all datasets.
- Also we concluded that NCL provides better recall value than RUS for first three datasets.

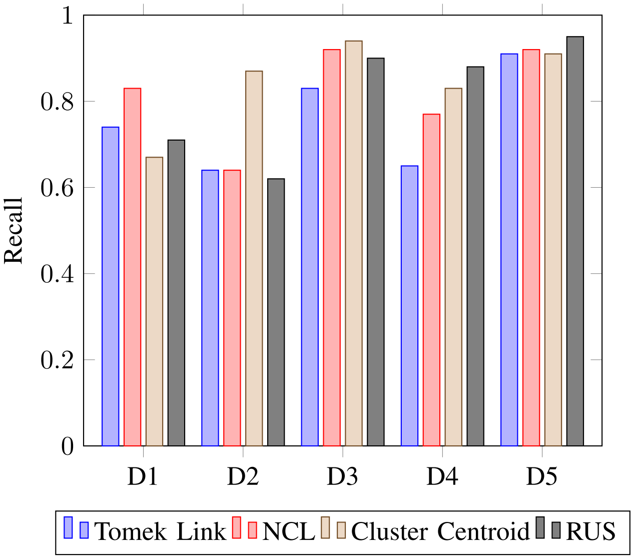

Fig. 2. Recall results for all low imbalance rate datasets.

- Cluster Centroid has better performance than Tomek link and NCL for all datasets except the first one with imbalance rate = 1.82

From Fig. 4 we can conclude the following points for all medium IR dataset groups recall values.

- From Fig. 5 we show that Tomek link provides the worst performance for all datasets.
- NCL provides better performance than Cluster Centroid for all datasets
- RUS outperformed all other techniques for all datasets

From Fig. 5 we can concludes the following points for all high IR dataset groups recall value

- NCL provides better performance based on recall than Tomek link for all datasets.
- RUS outperformed all other techniques for all datasets except the last one with imbalance rate = 129.44.
- Tomek link provides the worst performance for all datasets

TABLE II. Recall Results for All Examined Datasets Using the Selected Undersampling Techniques

Dataset	Tomek Link	NCL	Cluster Centroids	RUS
glass1	.74	.83	.67	.71
ecoli_0_vs_1	.64	.64	.87	.62
vehicle0	.83	.92	.94	.9
ecoli3	.65	.77	.83	.88
page-blocks0	.91	.92	.91	.95
glass4	.49	.98	.91	.99
car-good	.92	.97	.95	.99
kr-vs-k-one_vs_fifteen	1	1	.98	1
kr-vs-k-zero_vs_eight	.92	.96	.93	.99
Winquality	.49	.5	.76	.81
kr-vs-k-zero_vs_fifteen	1	1	1	1
abalone19	.56	.68	.77	.71

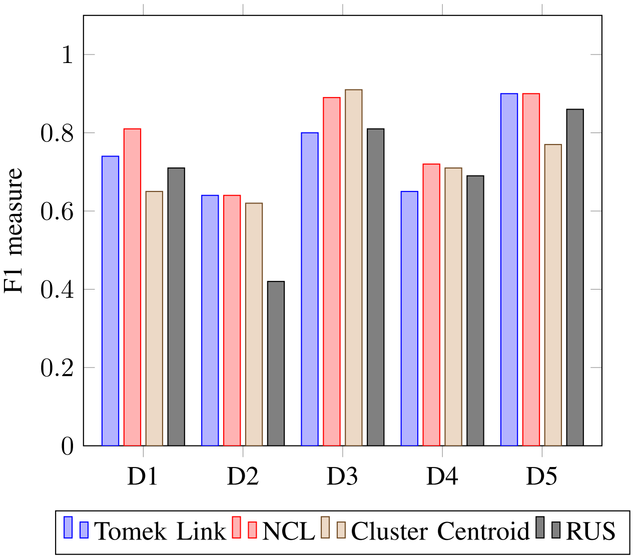

Fig. 3. F1 measure results for all low imbalance rate datasets.

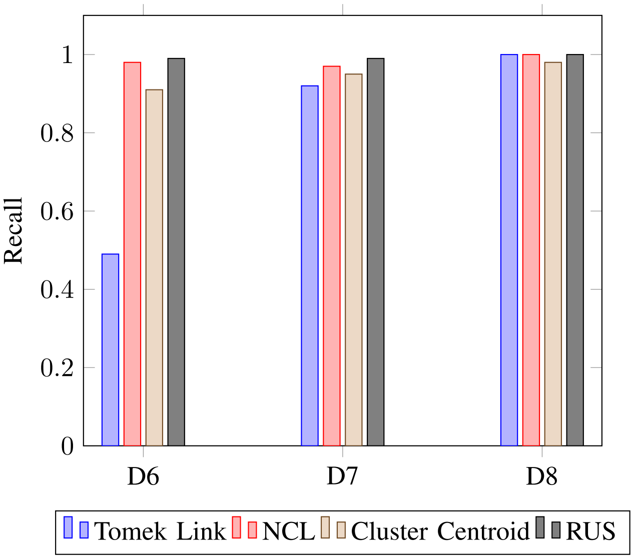

Fig. 4. Recall results for all medium imbalance rate datasets.

B. F1-Measure

From Table III, Fig. 3, 6 to 16 we can concluded the following points

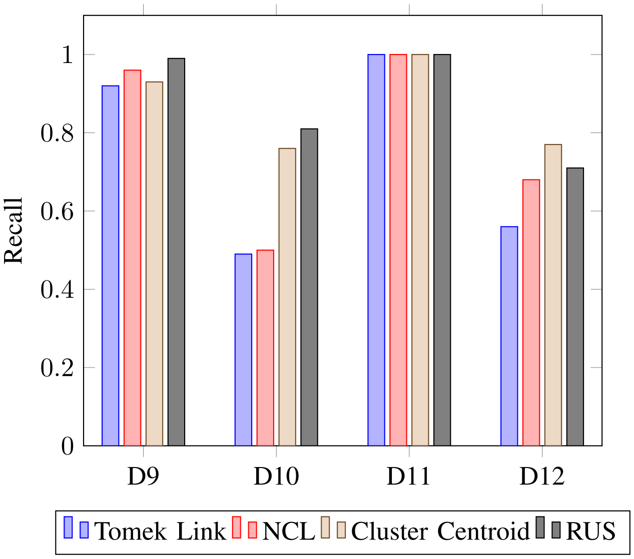

Fig. 5. Recall results for all high imbalance rate datasets.

- RUS provides less performance than NCL for all low imbalance rate datasets
- NCL provides better or equal performance than Tomek link for all medium imbalance rate group datasets
- Clusters Centroid provides better or equal performance than Tomek link for all medium imbalance rate group datasets
- RUS provides better or equal performance than Clusters Centroid for all medium imbalance rate group datasets
- Clusters Centroid provides the worst performance for all high imbalance rate group datasets except for last one with imbalance rate = 129.44

C. Gmean

From Table IV we can concluded the following points:

- Tomek link provides the worst performance for all low, medium, and high imbalance rate datasets
- NCL has better performance than Cluster Centroid for all datasets in the low imbalance rate group except the first one with imbalance rate = 1.82

TABLE III. F1-MEASURE RESULTS FOR ALL EXAMINED DATASETS USING THE SELECTED UNDERSAMPLING TECHNIQUES

Dataset	Tomek Link	NCL	Cluster Centroids	RUS
glass1	.74	.81	.65	.71
ecoli-0_vs_1	.64	.64	.62	.42
vehicle0	.8	.89	.91	.81
ecoli3	.65	.72	.71	.69
page-blocks0	.9	.9	.77	.86
glass4	.49	.74	.52	.83
car-good	.93	.96	.77	.92
kr-vs-k-one_vs_fifteen	1	1	.79	.95
kr-vs-k-zero_vs_eight	.95	.96	.6	.88
Winquality	.49	.5	.37	.4
kr-vs-k-zero_vs_fifteen	1	1	.89	.95
abalone19	.54	.44	.46	.41

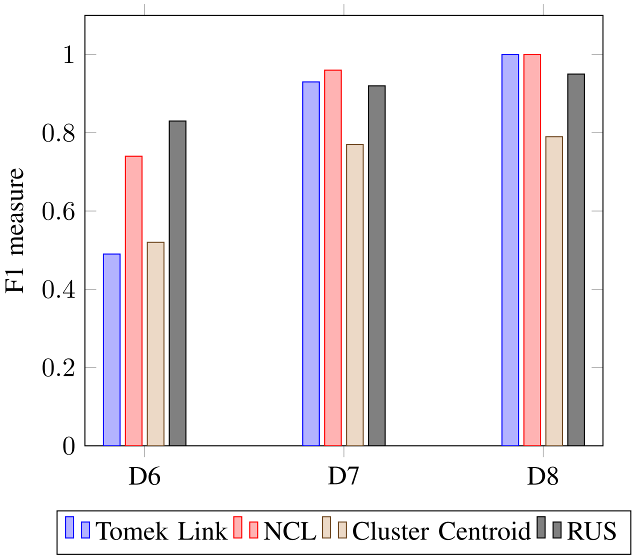

Fig. 6. F1 Measure results for all medium imbalance rate datasets.

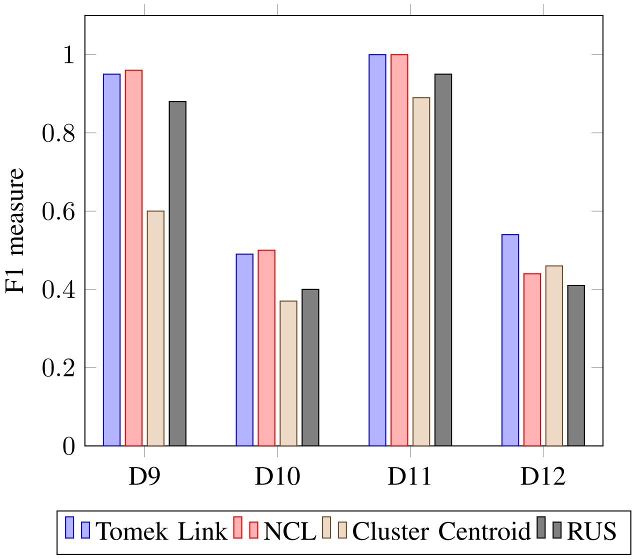

Fig. 7. F1 Measure results for all high imbalance rate datasets.

- NCL and RUS provide the same gmean value =1 for all datasets in the medium imbalance rate group and these sampling techniques outperformed Cluster Centroid

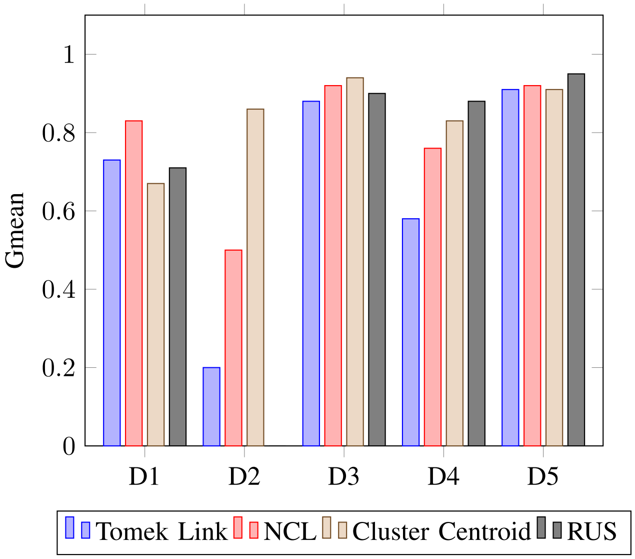

Fig. 8. Gmean results for all low imbalance rate datasets.

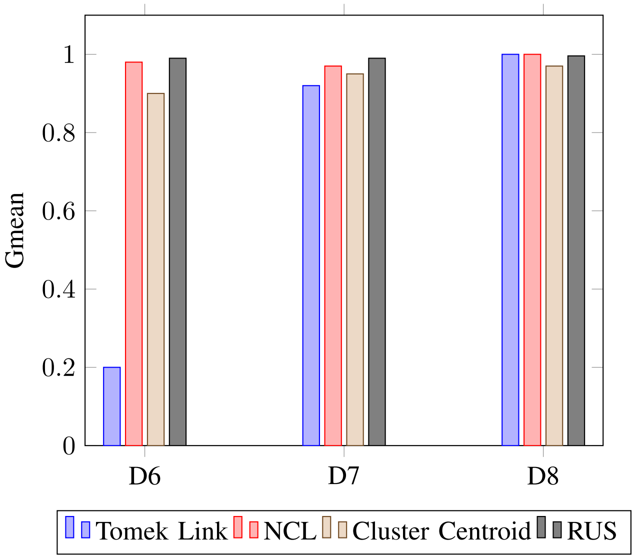

Fig. 9. Gmean results for all medium imbalance rate datasets.

TABLE IV. GMEAN RESULTS FOR ALL EXAMINED DATASETS USING THE SELECTED UNDERSAMPLING TECHNIQUES

Dataset	Tomek Link	NCL	Cluster Centroids	RUS
glass1	.73	.83	.67	.71
ecoli-0_vs_1	.2	.5	.86	0
vehicle0	.88	.92	.94	.9
ecoli3	.58	.76	.83	.88
page-blocks0	.91	.92	.91	.95
glass4	.2	.98	.9	.99
car-good	.92	.97	.95	.99
kr-vs-k-one_vs_fifteen	1	1	.97	.996
kr-vs-k-zero_vs_eight	.91	.96	.92	.99
Winquality	0	0	.73	.78
kr-vs-k-zero_vs_fifteen	1	1	.996	.998
abalone19	.35	.68	.77	.71

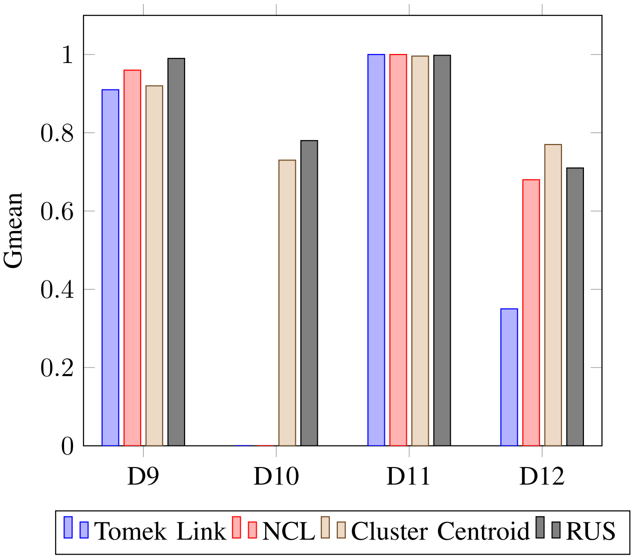

Fig. 10. Gmean results for all high imbalance rate datasets.

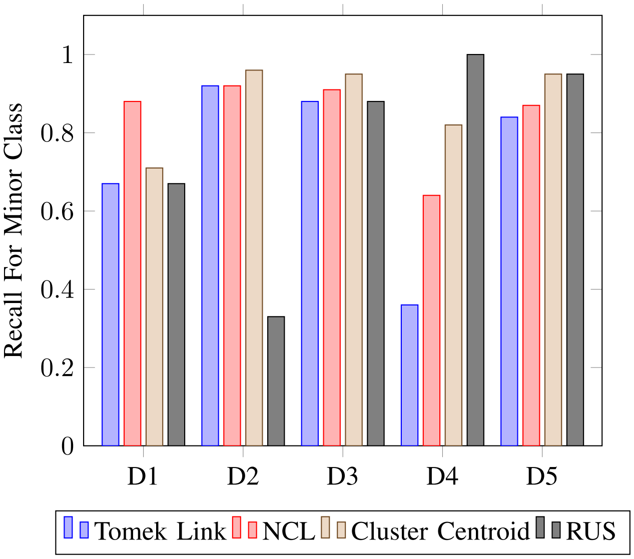

Fig. 11. Recall for minor class results for all low imbalance rate datasets.

D. Recall for Minor Class

From Table V we can concluded the following points:

- NCL outperformed Cluster Centroid for all datasets in the Low imbalance rate group except the first one with an imbalance rate = 1.82.
- NCL provides better or equal performance than Tomek link for all Low imbalance rate datasets.
- Clusters Centroid provides better performance than Tomek link for all Low imbalance rate datasets.
- Clusters Centroid and RUS provide recall value = 1 for the minor class for all datasets in the medium imbalance rate group.
- Tomek link provides the worst performance for all medium and high imbalance rate datasets.
- Clusters Centroid and RUS provide the same recall value for the minor class for all datasets in the high imbalance rate group.

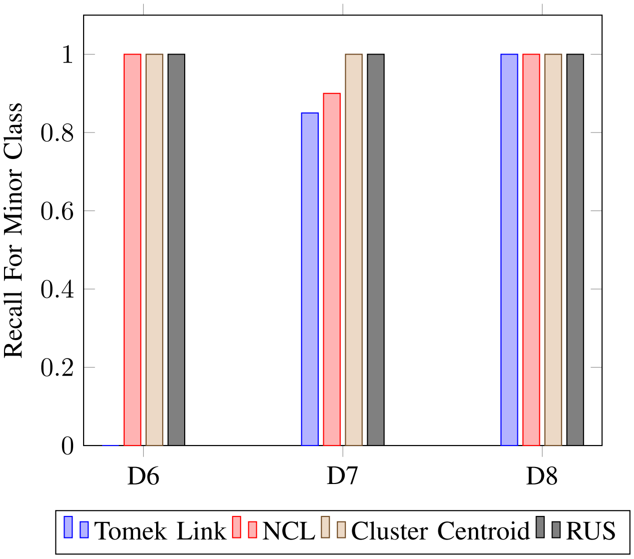

Fig. 12. Recall for minor class results for all medium imbalance rate datasets.

E. F1 Measure for Minor Class

From Table VI we can conclude the following points:

- NCL provides better or equal performance than Tomek link and Cluster Centroid for all low imbalance rate

TABLE V. Recall for Minor Class

Dataset	Tomek Link	NCL	Cluster Centroids	RUS
glass1	.67	.88	.71	.67
ecoli-0_vs_1	.92	.92	.96	.33
vehicle0	.88	.91	.95	.88
ecoli3	.36	.64	.82	1
page-blocks0	.84	.87	.95	.95
glass4	0	1	1	1
car-good	.85	.9	1	1
kr-vs-k-one_vs_fifteen	1	1	1	1
kr-vs-k-zero_vs_eight	.83	.92	1	1
Winquality	0	0	1	1
kr-vs-k-zero_vs_fifteen	1	1	1	1
abalone19	.12	.62	.75	.75

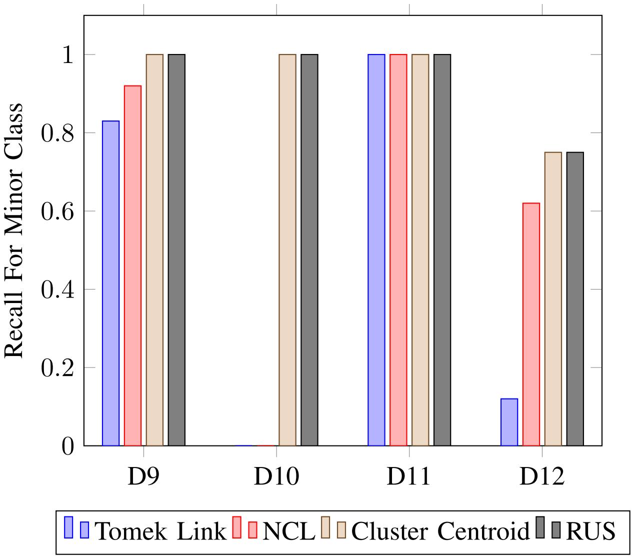

Fig. 13. Recall for minor class results for all high imbalance rate datasets.

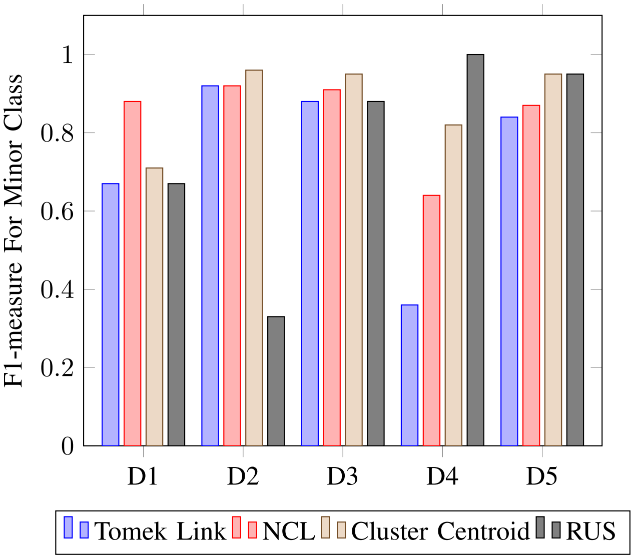

Fig. 14. F1-measure for minor class results for all low imbalance rate datasets.

datasets except vehicle0 dataset with imbalance rate = 3.25

- RUS provides less performance than NCL for all medium imbalance rate datasets except the first one with imbalance rate = 15.47
- Cluster Centroid provides less performance than NCL and RUS for all medium imbalance rate datasets
- RUS provides better performance than Clusters Centroid for all high imbalance rate datasets except for the abalone19 dataset with imbalance rate = 129.44

V. Conclusion and Future Work

Sampling techniques are one of the most effective ways of handling imbalanced data set problems in machine learning.

This paper is concerned on comparing the effects of four common under-shambling techniques including Tomek Link, NCL, RUS, and Clusters Centroid in handling imbalance datasets problems for various Imbalance ratios ranges from low, medium, and high IR. Twelve imbalanced data sets with various IR have been used for comparison. DT has been used for classification.

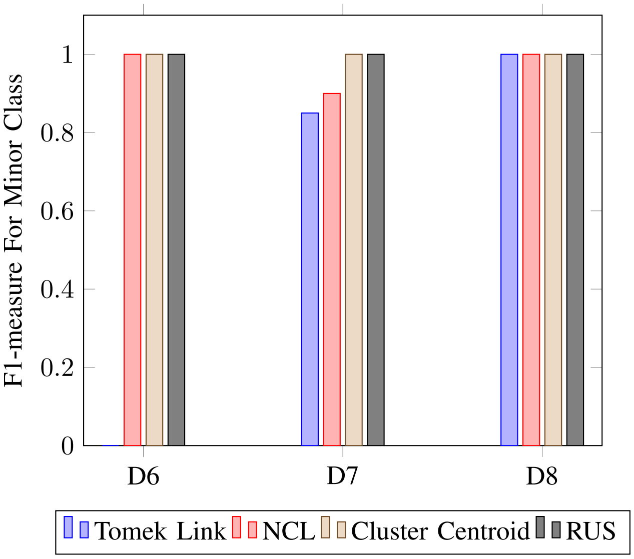

Fig. 15. F1-measure for minor class results for all medium imbalance rate datasets.

Results from all low IR datasets clearly show that NCL

TABLE VI. F1-Measure for Minor Class

Dataset	Tomek Link	NCL	Cluster Centroids	RUS
glass1	.67	.78	.61	.64
ecoli-0_vs_1	.96	.96	.96	.74
vehicle0	.87	.83	.87	.87
ecoli3	.38	.52	.51	.51
page-blocks0	.82	.82	.6	.76
glass4	0	0.5	.14	.67
car-good	.87	.92	.28	.77
kr-vs-k-one_vs_fifteen	1	1	.61	.91
kr-vs-k-zero_vs_eight	.91	.92	.28	.77
Winquality	0	0	.04	.05
kr-vs-k-zero_vs_fifteen	1	1	.78	0.9
abalone19	.08	.03	.04	.03

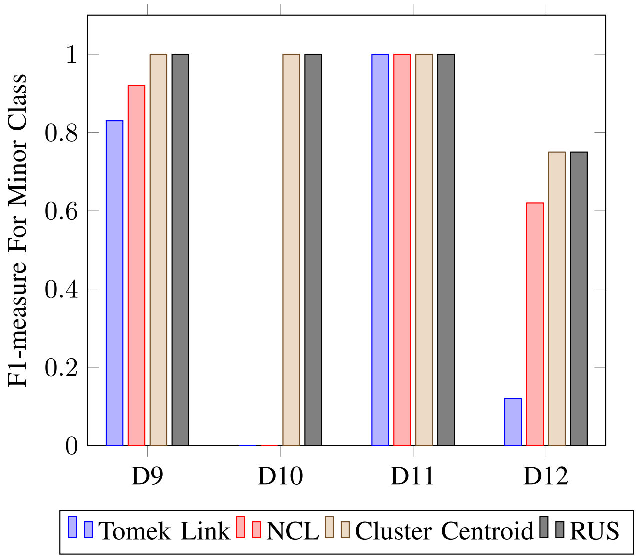

Fig. 16. F1-measure for minor class results for all high imbalance rate datasets.

outperformed both Tomek Link and RUS, while Clusters Centroid outperformed NCL based on recall. Based on minor class recall and gmean, NCL outperformed Clusters Centroid. RUS provides less performance than NCL for all low IR datasets.

For all medium IR datasets, Tomek Link provides the worst performance, while NCL and RUS outperformed other techniques in terms of recall, minor class recall, and gmean values. Based on the minor class F1 measure NCL outperformed RUS which outperformed Clusters Centroid based on the average F1 measure.

For all high IR datasets, Tomek Link provides the worst performance, while NCL and RUS outperformed other techniques based on recall, minor class recall, and gmean values. Based on the average F1 measure and minor class F1 measure, RUS provides better performance than Clusters Centroid.

Finally, the results presented in this paper were derived from the databases used here and according to the rates that were set to classify these databases for only four common under sampling techniques.

In the future, we need to compare the effect of these techniques by applying them to more databases in each IR category. Also, we can study more under-sampling techniques, and comparing them with other oversampling techniques.

References

[1] N. S. Shikha Gupta, "Machine learning driven threat identification to enhance fanet security using genetic algorithm," *The International Arab Journal of Information Technology (IAJIT)*, vol. 21, no. 04, pp. 711 – 722, 1970.

[2] H. A. Owida, H. S. Migdadi, O. S. M. Hemied, N. F. F. Alshdaifat, S. F. A. Abuowaida, and R. S. Alkhawaldeh, "Deep learning algorithms to improve covid-19 classification based on ct images," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 5, pp. 2876–2885, 2022.

[3] H. Owida, O. S. M. HEMIED, R. S. ALKHAWALDEH, N. F. F. ALSHDAIFAT, and S. F. A. ABUOWAIDA, "Improved deep learning approaches for covid-19 recognition in ct images," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 13, pp. 4925–4931, 2022.

[4] A. Y. Alhusenat, H. A. Owida, H. A. Rababah, J. I. Al-Nabulsi, and S. Abuowaida, "A secured multi-stages authentication protocol for iot devices." *Mathematical Modelling of Engineering Problems*, vol. 10, no. 4, 2023.

[5] S. ABUOWAIDA, E. ELSOUD, A. AL-MOMANI, M. ARABIAT, H. A. OWIDA, N. ALSHDAIFAT, and H. Y. CHAN, "Proposed enhanced feature extraction for multi-food detection method," *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 24, 2023.

[6] H. Abu Owida, "Recent biomimetic approaches for articular cartilage tissue engineering and their clinical applications: narrative review of the literature," *Advances in Orthopedics*, vol. 2022, no. 1, p. 8670174, 2022.

[7] H. A. Owida, B. A.-h. Moh'd, and M. Al Takrouri, "Designing an integrated low-cost electrospinning device for nanofibrous scaffold fabrication," *HardwareX*, vol. 11, p. e00250, 2022.

[8] A. Al-Momani, M. N. Al-Refai, S. Abuowaida, M. Arabiat, N. Alshdaifat, and M. N. A. Rahman, "The effect of technological context on smart home adoption in jordan," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, p. 1186 – 1195, 2024.

[9] E. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," *Computers in Biology and Medicine*, vol. 140, p. 105051, 2022.

[10] Z. Salah and E. Abu Elsoud, "Enhancing network security: A machine learning-based approach for detecting and mitigating krack and kr00k attacks in ieee 802.11," *Future Internet*, vol. 15, no. 8, p. 269, 2023.

[11] H. Alazzam, A. Al-Adwan, O. Abualghanam, E. Alhenawi, and A. Alsmady, "An improved binary owl feature selection in the context of android malware detection," *Computers*, vol. 11, no. 12, p. 173, 2022.

[12] R. Al-Sayyed, E. Alhenawi, H. Alazzam, A. Wrikat, and D. Suleiman, "Mobile money fraud detection using data analysis and visualization techniques," *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17 093–17 108, 2024.

[13] S. Shukri, R. Al-Sayyed, H. Al-Bdour, E. Alhenawi, T. Almarabeh, and H. Mohammad, "Internet of things: Underwater routing based on user's health status for smart diving," *International Journal of Data and Network Science*, vol. 7, no. 4, pp. 1715–1728, 2023.

[14] M. Haj Qasem, M. Aljaidi, G. Samara, R. Alazaidah, A. Alsarhan, and M. Alshammari, "An intelligent decision support system based on multi agent systems for business classification problem," *Sustainability*, vol. 15, no. 14, p. 10977, 2023.

[15] ——, "An intelligent decision support system based on multi agent systems for business classification problem," *Sustainability*, vol. 15, no. 14, p. 10977, 2023.

[16] T. Sabbah, M. Ayyash, and M. Ashraf, "Hybrid support vector machine based feature selection method for text classification." *The International Arab Journal of Information Technology (IAJIT).*, vol. 15, no. 3A, pp. 599–609, 2018.

[17] E. Alhenawi, R. A. Khurma, P. A. Castillo, M. G. Arenas, and A. M. Al-Hinawi, "Effects of term weighting approach with and without stop words removing on arabic text classification," in *2023 9th International Conference on Optimization and Applications (ICOA)*, 2023, pp. 1–6.

[18] H. Alazzam, O. AbuAlghanam, A. Alsmady, and E. Alhenawi, "Arabic documents clustering using bond energy algorithm and genetic algorithm," in *2022 13th International Conference on Information and Communication Systems (ICICS)*. IEEE, 2022, pp. 4–8.

[19] O. K. A. Alidmat, K. Y. Umi, E. Alhenawi, H. J. Badarneh, R. Alazaidah, and L. Al-Rbabah, "Simulation of exit selection behavior evacuation based on an improved cellular automata model during fire disaster," in *2023 24th International Arab Conference on Information Technology (ACIT)*. IEEE, 2023, pp. 1–8.

[20] O. Alidmat, H. A. Owida, U. K. Yusof, A. Almaghthawi, A. Altalidi, R. S. Alkhawaldeh, S. Abuowaida, N. Alshdaifat, and J. AlShaqsi, "Simulation of crowd evacuation in asymmetrical exit layout based on improved dynamic parameters model," *IEEE Access*, 2024.

[21] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in *2020 11th international conference on information and communication systems (ICICS)*. IEEE, 2020, pp. 243–248.

[22] R. Alazaidah, A. Al-Shaikh, M. Al-Mousa, H. Khafajah, G. Samara, M. Alzyoud, N. Al-Shanableh, and S. Almatarneh, "Website phishing detection using machine learning techniques," *Journal of Statistics Applications & Probability*, vol. 13, no. 1, pp. 119–129, 2024.

[23] R. Alazaidah, F. K. Ahmad, M. F. M. Mohsin, and W. A. AlZoubi, "Multi-label ranking method based on positive class correlations," *Jordanian Journal of Computers and Information Technology*, 2020.

[24] O. AbuAlghanam, H. Alazzam, E. Alhenawi, M. Qatawneh, and O. Adwan, "Fusion-based anomaly detection system using modified isolation forest for internet of things," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–15, 2022.

[25] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.

[26] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, pp. 1–30, 2018.

[27] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[28] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Int'l Conf. on Artificial Intelligence*, vol. 56. Citeseer, 2000, pp. 111–117.

[29] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *2008 Fourth international conference on natural computation*, vol. 4. IEEE, 2008, pp. 192–201.

[30] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent data analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[31] S.-J. Yen and Y.-S. Lee, "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation*. Springer, 2006, pp. 731–740.

[32] ——, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5718–5727, 2009.

[33] Y.-P. Zhang, L.-N. Zhang, and Y.-C. Wang, "Cluster-based majority under-sampling approaches for class imbalance learning," in *2010 2nd IEEE International Conference on Information and Financial Engineering*. IEEE, 2010, pp. 400–404.

[34] M. M. Rahman and D. Davis, "Cluster based under-sampling for unbalanced cardiovascular data," in *Proceedings of the world congress on engineering*, vol. 3, 2013, pp. 3–5.

[35] H. A. Owida, N. Alshdaifat, A. Almaghthawi, S. Abuowaida, A. Aburomman, A. Al-Momani, M. Arabiat, and H. Y. Chan, "Improved deep learning architecture for skin cancer classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 1, p. 501 – 508, 2024. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85200149026&doi=10.11591%2fijeecs.v36.i1.pp501-508&partnerID=40&md5=0b7a43e8a8aac2d4ae6e3688bd3f6f93

[36] T. Hasanin and T. Khoshgoftaar, "The effects of random undersampling with simulated class imbalance for big data," in *2018 IEEE international conference on information reuse and integration (IRI)*. IEEE, 2018, pp. 70–79.

[37] T. Hasanin, T. M. Khoshgoftaar, J. Leevy, and N. Seliya, "Investigating random undersampling and feature selection on bioinformatics big data," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2019, pp. 346–356.

[38] M. Saripuddin, A. Suliman, S. Syarmila Sameon, and B. N. Jorgensen, "Random undersampling on imbalance time series data for anomaly detection," in *2021 The 4th International Conference on Machine Learning and Machine Intelligence*, 2021, pp. 151–156.

[39] T. Elhassan and M. Aljurf, "Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method," *Global J Technol Optim S*, vol. 1, 2016.

[40] R. Zuech, J. Hancock, and T. M. Khoshgoftaar, "Detecting web attacks using random undersampling and ensemble learners," *Journal of Big Data*, vol. 8, no. 1, pp. 1–20, 2021.

[41] J. L. Leevy, J. Hancock, T. M. Khoshgoftaar, and N. Seliya, "Iot reconnaissance attack classification with random undersampling and ensemble feature selection," in *2021 IEEE 7th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2021, pp. 41–49.

[42] M. H. Popel, K. M. Hasib, S. A. Habib, and F. M. Shah, "A hybrid under-sampling method (husboost) to classify imbalanced data," in *2018 21st international conference of computer and information technology (ICCIT)*. IEEE, 2018, pp. 1–7.

[43] K. Agustianto and P. Destarianto, "Imbalance data handling using neighborhood cleaning rule (ncl) sampling method for precision student modeling," in *2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*. IEEE, 2019, pp. 86–89.

[44] Y. Wu, J. Yao, S. Chang, and B. Liu, "Limcr: Less-informative majorities cleaning rule based on naïve bayes for imbalance learning in software defect prediction," *Applied Sciences*, vol. 10, no. 23, p. 8324, 2020.

[45] Y. Zhang, H. Zhang, X. Zhang, and D. Qi, "Deep learning intrusion detection model based on optimized imbalanced network data," in *2018 IEEE 18th International Conference on Communication Technology (ICCT)*. IEEE, 2018, pp. 1128–1132.

[46] P. Gulati, "Hybrid resampling technique to tackle the imbalanced classification problem," *Applied Sciences*, 2020.

[47] S. Choirunnisa and J. Lianto, "Hybrid method of undersampling and oversampling for handling imbalanced data," in *2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2018, pp. 276–280.

[48] Q. Ning, X. Zhao, and Z. Ma, "A novel method for identification of glutarylation sites combining borderline-smote with tomek links technique in imbalanced data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021.

[49] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and smote approaches for machine fault classification with an imbalanced dataset," *Sensors*, vol. 22, no. 9, p. 3246, 2022.

[50] A. Bansal and A. Jain, "Analysis of focussed under-sampling techniques with machine learning classifiers," in *2021 IEEE/ACIS 19th International Conference on Software Engineering Research, Management and Applications (SERA)*. IEEE, 2021, pp. 91–96.

[51] O. Tarawneh, Q. Saber, A. Almaghthawi, H. A. Owida, A. Issa, N. Alshdaifat, G. Jaradat, S. Abuowaida, and M. Arabiat, "The effect of pre-processing on a convolutional neural network model for dorsal hand vein recognition." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 3, 2024.

[52] H. A. Owida, M. R. Hassan, A. M. Ali, F. Alnaimat, A. Al Sharah, S. Abuowaida, and N. Alshdaifat, "The performance of artificial intelligence in prostate magnetic resonance imaging screening," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 2, pp. 2234–2241, 2024.

[53] A. Al Sharah, H. A. Owida, F. Alnaimat, and S. Abuowaida, "Application of machine learning in chemical engineering: outlook and perspectives," *Int J Artif Intell*, vol. 13, no. 1, pp. 619–630, 2024.

[54] N. Alshdaifat, M. A. Osman, and A. Z. Talib, "An improved multi-object instance segmentation based on deep learning," *Kuwait Journal of Science*, vol. 49, no. 2, 2022.

[55] J. Wang and Y. Wang, "Fd technology for hss based on deep convolutional generative adversarial networks." *The International Arab Journal of Information Technology (IAJIT)*, vol. 21, no. 2, pp. 299–312, 2024.

[56] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.

[57] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," *arXiv preprint arXiv:1608.06048*, 2016.

[58] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Conference on artificial intelligence in medicine in Europe*. Springer, 2001, pp. 63–66.

[59] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, "Handling imbalanced datasets: A review," *GESTS international transactions on computer science and engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[60] H. Rawashdeh, S. Awawdeh, F. Shannag, E. Henawi, H. Faris, N. Obeid, and J. Hyett, "Intelligent system based on data mining techniques for prediction of preterm birth for women with cervical cerclage," *Computational biology and chemistry*, vol. 85, p. 107233, 2020.

[61] H. Abu Owida, G. AlMahadin, J. I. Al-Nabulsi, N. Turab, S. Abuowaida, and N. Alshdaifat, "Automated classification of brain tumor-based magnetic resonance imaging using deep learning approach." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 14, no. 3, 2024.

[62] A. K. Shukla, P. Singh, and M. Vardhan, "A hybrid gene selection method for microarray recognition," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 4, pp. 975–991, 2018.