ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/373467415

The Jeopardy of Learning from Over-Sampled Class-Imbalanced Medical Datasets

Conference Paper · July 2023

DOI: 10.1109/ISCC58397.2023.10218211

CITATIONS
10

READS
318

7 authors, including:

Ahmad Hassanat
Mutah University
150 PUBLICATIONS **3,372** CITATIONS
SEE PROFILE

Ghada Awad Altarawneh
Mutah University
34 PUBLICATIONS **613** CITATIONS
SEE PROFILE

Ibraheem M Alkhawaldeh
Mutah University
83 PUBLICATIONS **125** CITATIONS
SEE PROFILE

Yasmeen Jamal Alabdallat
Hashemite University
62 PUBLICATIONS **102** CITATIONS
SEE PROFILE

All content following this page was uploaded by Ahmad Hassanat on 29 August 2023.

The user has requested enhancement of the downloaded file.

The Jeopardy of Learning from Over-Sampled Class-Imbalanced Medical Datasets

1st Ahmad Hassanat
Faculty of Information Technology
Mutah University
Karak, Jordan
hasanat@mutah.edu.jo

2nd Ghada Altarawneh
Accounting department
Mutah University
Karak, Jordan
ghadaa@mutah.edu.jo

3rd Ibraheem M. Alkhawaldeh
Faculty of Medicine
Mutah University
Karak, Jordan
Ibraheemfamous096@gmail.com

4th Yasmeen Jamal Alabdallat
Faculty of Medicine
Hashemite University
Zarqa, Jordan
abdallat.01@gmail.com

5th Amir F. Atiya
Computer Engineering Department
Cairo University
Giza, Egypt
amir@alumni.caltech.edu

6th Ahmad Abujaber
Executive Director
Hamad Medical Corp.
Doha, Qatar
AABUJABER@hamad.qa

7th Ahmad S. Tarawneh
Department of DS and AI
University of Petra
Amman, Jordan
ahmad.trwh@gmail.com

***Abstract*—**

The usefulness of the oversampling approach to class-imbalanced structured medical datasets is discussed in this paper. In this regard, we basically look into the oversampling approach's prevailing assumption that synthesized instances do belong to the minority class. We used an off-the-shelf over-sampling validation system to test this assumption. According to the experimental results from the validation system, at least one of the three medical datasets used had newly generated samples that were not belonging to the minority class as a result of the oversampling methods validated. Additionally, the error rate varied based on the dataset and oversampling method tested. Therefore, we claim that synthesizing new instances without first confirming that they are aligned with the minority class is a risky approach, especially in medical fields where misdiagnosis can have serious repercussions. As alternatives to oversampling, ensemble, data partitioning, and method-level approaches are advised since they do not make false assumptions.

***Index Terms*—machine learning, class imbalance, Medical apps, Easy ensemble,**

I. Introduction

Class imbalance happens when a dataset is trained with disproportionately more examples from one class than the other. Most often, the dominating class is referred to as the majority class, while the class with a much smaller number of examples is referred to as the minority class. Generally, Classifiers trained on unequal training sets have a prediction bias, which is connected to poor performance in the minority class(es), which is the primary cause of the class imbalance problem. The bias might vary greatly depending on the dataset used, from a slight imbalance to a serious imbalance.

The minority class is frequently of crucial importance since it provides positive examples that are uncommon in nature or expensive to obtain, which has led to the problem growing and becoming a substantial challenge. When it comes to apps and datasets for the medical field, this is especially true. Because Medical datasets are frequently class imbalanced [1], for instance, there are considerably more samples in the non-patients/negative (majority) class set than in the patients/positive (minority) class set [2]. Hence the issue of class imbalance is common in medical classification apps [3], since a classifier can achieve high accuracy even if it properly assigns all of the samples to the (patients/negative) majority class having the propensity to assign (patients/positive) minority samples to the majority class. consequently, this poses a risk in medical apps since incorrectly classifying the patient class set has more severe repercussions than incorrectly classifying the non-patient class set.

Class imbalance solutions fall into three approaches: data-level approach, algorithm-level approach, and a hybrid of both, such as the ensemble learning approach. The data-level approach involves three different approaches to data manipulation, namely, features engineering, undersampling the majority class, and oversampling the minority class.

Oversampling -starting with the Synthetic Minority Oversampling Technique (SMOTE) [4], is the most often used approach to solve the class imbalance problem, as seen by the multitude of oversampling methods published in the last two decades. For example, On January 26, 2022, a Google Scholar search for the term ”SMOTE” yielded 94,900 results, while a search for ”oversampling” yielded 360,000 results.

According to the plethora of oversampling approaches published over the past 20 years, oversampling, beginning with the Synthetic Minority Oversampling Technique (SMOTE) [4], is the most often employed approach to address the class imbalance problem. For instance, a Google Scholar search for ”SMOTE” on May 16, 2023 returned 94,900 results, whereas a search for ”oversampling” returned 360,000 results. The reasons for this unusual increase in oversampling research include the applicability of the clearly defined class imbalance problem and the ease of oversampling remedies [5].

This does not, however, necessarily mean that the oversampling approach is advantageous. By generating new cases out of thin air based only on their similarity to one or more of the minority's examples, oversampling techniques increase the number of minority-class instances. This is problematic since using such techniques could increase the risk of overfitting the learning process [6]–[9].

Overfitted artificially generated datasets yield positive machine learning outcomes, however, this is not necessarily the case in real-world medical applications. A more serious issue with oversampling is that, regardless of how close the made-up instances are to those of the patients/positive) minority, they could exist in the real world and belong to a different class [5].

By calculating the probability distribution of the SMOTE-generated samples, Elreedy and coworkers [10] developed a novel theoretical analysis of the SMOTE method and came to the conclusion that the synthetic data produced by SMOTE might not exactly match the original minority class distribution, which could affect the classification performance.

Almost similarly, Tarawneh and coworkers [5] Argue that *"Oversampling in its current forms and methodologies is a misleading approach that should be avoided since it feeds the learning process with falsified instances that are pushed to be members of the minority class when they are most likely members of the majority."*

This conclusion was reached based on the findings of their recommended validation system, which essentially applied numerous oversampling methods to a number of class-imbalanced datasets, hid a number of majority examples, and then checked their similarity to the synthesized examples by each oversampling method tested.

The aim of this paper is to employ the same tester, proposed by [5] in order to determine whether the oversampling approach is beneficial for class-imbalanced medical datasets, specifically those used by various oversampling methods, as claimed by some researchers such as [11]–[14], or detrimental as claimed by some researchers such as [2], [5], [10].

II. Related work

Many publications have used oversampling to create artificial samples from minority samples to address the issue of class imbalance in the medical field. For example, searching PubMed for the term ("oversampling" OR "smote") returned 2157 results published between 2000 and 2022, while searching the Web of Science (WOS) for the same query (but filtering the results to the medical subjects only) returned 2185 results. This merely serves as a prelude to the emerging trend of oversampling research that dealt with or simply discussed oversampling in the medical literature. The sharp rise in the number of articles that discussed used or addressed oversampling or SMOTE is seen in Figure 1.

All types of data, including time series [15], medical images [16], and structured data [11]–[14], [17] were subjected to oversampling methods. In this paper, we will focus on the structured data, as the tester we are employing is intended for this type of machine learning data.

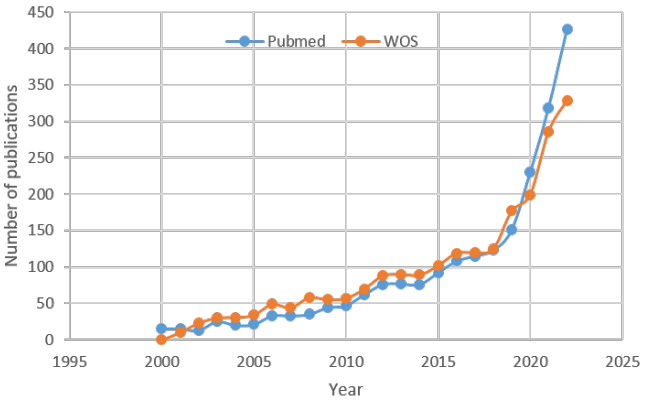

Fig. 1. Number of medical publications that discussed used or addressed oversampling or SMOTE between 2000 and 2022.

Examples of studies that used the oversampling approach on structured medical data include the work of Sreejith et al. [11], who presented a framework for creating a clinical decision support system that handles class imbalance, employing a SMOTE-improved approach to balance a dataset at the data level. The experimental findings on three clinical datasets indicated that utilizing oversampling improves the classification of Liver Patients, Thoracic Surgery, and Diabetes.

Naseriparsa et al. [12] proposed RSMOTE, a modified SMOTE-based medical diagnosis system, which generates new synthetic samples at the data level to create a balance between minority and majority classes. It does this by globally identifying the minority sample region and applying resampling close to a particular group of samples. The diagnosis system then evaluates the synthesized balanced dataset before making a decision. Their results reveal that the proposed medical diagnosis system performs better when compared to eight SMOTE-like methods such as SMOTE, Borderline SMOTE, and critical SMOTE. Their system was evaluated on four structured medical datasets: Heart, Diabetes, Hepatitis, and Breast Cancer Wisconsin (WDBC).

Sowjanya and Mrudula [13] proposed a two-phase oversampling approach for structured class-imbalanced medical data. The first phase involves modifying SMOTE to balance the classes using Distance-based SMOTE and Biphasic SMOTE, which were then combined with selected classifiers for prediction. The second phase was the application of machine learning methods to create a stacking ensemble framework that achieved significantly greater accuracy—96–97%—than individual algorithms. Three structured medical datasets—Framingham, Coronavirus 2019, and WDBC—were used to assess their system. Being that ensembling is one of the good ways for class imbalance [18], we believe it would be preferable if they were confined to the second phase without using any oversampling.

Fotouhi et al. [14] examined 18 oversampling and undersampling techniques that were tested on 15 cancer datasets from the SEER program, including those for kidney, soft tissue, bladder, rectum, colon, bone, larynx, breast, cervix, prostate, oropharynx, melanoma, thyroid, testis, and lip. The examined oversampling methods included SMOTE, ADASYN, ADOMS, AHC, Borderline-SMOTE, and others. The results showed that employing oversampling techniques with the right classifier led to a noticeable boost in classification accuracy.

Other studies that oversampled structured medical data include [17], [19]–[22] and many more.

Without offering any assurance, all oversampling techniques make the assumption that the synthetic instances belong to the minority class based on some resemblance. And the high degree of precision they achieve serves to support their virtue. However, it is clear that if we increase the number of instances by comparable ones, some will be included in the training set and some will be included in the testing set, allowing for overfitting [6]–[9]. This is unacceptable in machine learning since it is as if we are training and testing on the same data.

III. Methods

In order to determine whether such an oversampling practice is appropriate for use with medical data, we are investigating a large number of oversampling methods in this paper on some of the most frequently used class-imbalanced structured medical data from previous studies. The application of these methods is automatically done by the tester recommended by [5].

The oversampling methods investigated in this paper include but are not limited to SMOTE [4], SMOTE TomekLinks [23], Borderline SMOTE [24], ADASYN [25], AHC [26], Distance SMOTE [27], polynom fit SMOTE [28], ADOMS [29], Safe Level SMOTE [30], MSMOTE [31], DE oversampling [32], SMOBD [33], SUNDO [34], MSYN [35], SVM balance [36], TRIM SMOTE [37], ProWSyn [38], SL graph SMOTE [39], LVQ SMOTE [40], SOI CJ [41], ROSE [42], SMOTE OUT [43], SMOTE Cosine [43], Selected SMOTE [43], LN SMOTE [44], MWMOTE [45], PDFOS [46], RWO sampling [47], NEATER [48], DEAGO [49], Gazzah [50], SMOTE IPF [51], KernelADASYN [52], MOT2LD [53], etc. The names of the other methods are mentioned in Tables I, II, and III. For a complete review of these methods the readers may refer to [5]. It is worth mentioning that most of these methods were utilized for class imbalance medical data as can be seen from the literature.

The tester works by concealing a portion of the majority's examples from the oversampling method tested, assuming that they were not obtained from the real world. Assuring that the class imbalance problem still exists after concealing the examples.

The tester will then use the remaining dataset to create new instances using the oversampling method validated. The training set then receives the concealed subset back.

In order to determine whether these synthesized examples belong to the majority or minority class, the tester examines how similar the synthetic examples are to every other example in the training set.

The similarity measure used by the tester is the Hassanat distance (HD) [54], [55], because it had been established that HD performed better than a variety of machine learning similarity metrics [56]–[60]. However, in our experiments, we used Euclidean distance (ED) since it is much faster.

The number of synthetic examples that are more similar to any of the majority's examples is divided by the total number of synthetic examples to get the oversampling error.

A. Data

Three structured medical datasets, some of which are often employed by oversampling techniques to address the issue of class imbalance in the medical field, were used in our investigations:

- Diabetes: consists of 268 positive/minority examples and 500 negative/majority examples with 8 numeric features. Data source: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
- Framingham: heart study dataset, consists of 644 positive/minority examples and 3596 negative/majority examples with 15 numeric features. Data source: https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset
- Thoracic Surgery: post-operative life expectancy in lung cancer patients, consists of 70 positive/minority examples and 400 negative/majority examples with 16 numeric and categorical features. The class (True/False) shows whether a patient will live for at least a year following surgery. Data source: https://archive-beta.ics.uci.edu/dataset/277/thoracic+surgery+data

IV. Results and discussion

After running the tester on the three aforementioned medical data sets, which each had 25% of the majority of examples hidden, we were able to validate the oversampling errors of the oversampling methods for the Diabetes, Framingham, and Thoracic Surgery datasets, as shown in Tables I, I, and I, respectively.

A careful analysis of these findings reveals that all oversampling methods validated lead to mistakes in the synthesized samples. In other words, they produce instances that are supposed to represent the minority but are actually more like the majority or fall within the decision boundary of the majority class.

Depending on the validated method and the oversampled dataset, the error rate ranged from 0 to 100% percent. However, none of these techniques achieves zero error on all datasets, indicating that they cannot accurately oversample medical records.

The cause of these errors is that these oversampling methods mistakenly assume that the synthesized examples actually belong to the minority since they fill in the feature space gap based on similarities to one or more minority instances.

This misleads the training of these examples and increases the likelihood that the classifier will become overfitted on false data. Because of this, it is feasible that the entire machine learning system would fail severely when used for real-world medical applications where even one erroneous synthesized example could cause severe damage.

Therefore, we think it is questionable practice, especially in the case of medical data, to oversample structured medical datasets by synthesizing new instances based on their resemblance to the minority examples without ensuring that the additional samples genuinely fall within the minority examples.

V. Conclusion

The usefulness of the oversampling approach with class-imbalanced structured medical datasets is addressed in this work. In order to determine if the synthesized instances genuinely belong to the minority example or not (as assumed by the oversampling approach), we utilized an oversampling validation system that was recommended by [5].

The experimental results of the validation system demonstrate that each of the oversampling methods examined had generated new samples in at least one of the three examined medical datasets that do not belong to the minority class. Additionally, the error rate varies based on the dataset utilized and the oversampling method that has been validated.

As a result, we believe that it is a dangerous practice to oversample structured medical datasets by synthesizing new instances based on their resemblance to the minority examples without ensuring that the generated samples actually fall within the minority examples, especially in the case of medical data, where patients may lose their lives as a result of wrong diagnosis of medical applications [14].

As an alternative to oversampling, we strongly recommend using ensemble approaches such as Easy Ensemble [61], Random Data Partitioning [18], and method-level approaches [62], because these approaches do not have wrong assumptions.

As our paper investigated a limited number of structured medical datasets, Future research should aim to evaluate the performance of oversampling methods and alternative approaches on a wider range of medical datasets, including different medical domains and various data types, such as time series, medical images, and unstructured text data.

The class imbalance issue in medical datasets will benefit from the development of more practical and trustworthy solutions as a result. It's crucial to assess how oversampling and alternative methods may affect actual medical applications. This may be accomplished by working together with subject-matter experts and medical professionals to develop and assess these approaches. Their opinions might offer insightful advice and direct the selection or modification of suitable techniques. Medical experts and data scientists working together can produce more trustworthy and efficient solutions.

TABLE I
Results of Diabetes dataset used to validate oversampling methods utilizing a 25% concealed percentage.

Diabetes			
Method	**No. Errors**	**No. Added**	**Error**
ADASYN	6	107	0.056
ADOMS	11	107	0.103
AHC	267	267	1.000
AMSCO	4	70	0.057
AND_SMOTE	2	107	0.019
ANS	28	107	0.262
ASMOBD	2	107	0.019
Assembled_SMOTE	3	107	0.028
Borderline_SMOTE1	8	107	0.075
Borderline_SMOTE2	2	107	0.019
CBSO	14	107	0.131
CCR	26	49	0.531
CE_SMOTE	1	107	0.009
cluster_SMOTE	2	107	0.019
CURE_SMOTE	9	107	0.084
DE_oversampling	3	79	0.038
DEAGO	70	107	0.654
distance_SMOTE	7	107	0.065
DSMOTE	22	107	0.206
DSRBF	4	107	0.037
Edge_Det_SMOTE	1	107	0.009
G_SMOTE	1	107	0.009
GASMOTE	21	765	0.027
Gaussian_SMOTE	43	107	0.402
polynom_fit_SMOTE_star	106	268	0.396
polynom_fit_SMOTE_bus	131	267	0.491
polynom_fit_SMOTE_poly	56	107	0.523
polynom_fit_SMOTE_mesh	41	107	0.383
KernelADASYN	46	107	0.430
Lee	1	107	0.009
LN_SMOTE	8	107	0.075
LVQ_SMOTE	40	107	0.374
MDO	49	107	0.458
MSYN	1	160	0.006
MWMOTE	17	107	0.159
NDO_sampling	0	107	0.000
NEATER	7	214	0.033
NRAS	11	107	0.103
NT_SMOTE	9	107	0.084
OUPS	27	107	0.252
PDFOS	41	107	0.383
ProWSyn	29	107	0.271
Random_SMOTE	14	107	0.131
ROSE	25	107	0.234
RWO_sampling	0	107	0.000
Safe_Level_SMOTE	30	107	0.280
SDSMOTE	2	107	0.019
Selected_SMOTE	3	107	0.028
SL_graph_SMOTE	26	107	0.243
SMOBD	9	107	0.084
SMOTE_Cosine	16	107	0.150
SMOTE_D	6	122	0.049
SMOTE_FRST_2T	5	126	0.040
SMOTE_IPF	2	107	0.019
SMOTE_OUT	1	107	0.009
SMOTE_PSO	1	588	0.002
SMOTE_PSOBAT	0	10	0.000
SMOTE_TomekLinks	2	51	0.039
SMOTE	0	107	0.000
SN_SMOTE	2	107	0.019
SOI_CJ	11	107	0.103
SSO	20	105	0.190
Supervised_SMOTE	15	107	0.140
SVM_balance	3	107	0.028
SYMPROD	0	107	0.000
V_SYNTH	67	107	0.626
MSMOTE	12	107	0.112

TABLE II
RESULTS OF FRAMINGHAM DATASET USED TO VALIDATE OVERSAMPLING METHODS UTILIZING A 25% CONCEALED PERCENTAGE.

Framingham			
Method	**No. Errors**	**No. Added**	**Error**
ADASYN	278	1769	0.157
ADOMS	353	1769	0.200
AHC	556	556	1.000
AND_SMOTE	83	1769	0.047
ANS	671	1769	0.379
ASMOBD	111	1769	0.063
Assembled_SMOTE	237	1769	0.134
Borderline_SMOTE1	240	1769	0.136
Borderline_SMOTE2	29	1769	0.016
CBSO	522	1769	0.295
CCR	1403	1892	0.742
CE_SMOTE	175	1769	0.099
cluster_SMOTE	150	1769	0.085
CURE_SMOTE	264	1769	0.149
DE_oversampling	210	1769	0.119
DEAGO	1636	1769	0.925
distance_SMOTE	333	1769	0.188
DSMOTE	1472	1769	0.832
DSRBF	174	1769	0.098
Edge_Det_SMOTE	249	1769	0.141
G_SMOTE	195	1769	0.110
GASMOTE	217	1697	0.128
Gaussian_SMOTE	1231	1769	0.696
polynom_fit_SMOTE_star	910	1671	0.545
polynom_fit_SMOTE_bus	807	1668	0.484
polynom_fit_SMOTE_poly	1503	1769	0.850
polynom_fit_SMOTE_mesh	1056	1769	0.597
Gazzah	77	135	0.570
KernelADASYN	1212	1769	0.685
Lee	101	1769	0.057
LN_SMOTE	57	1769	0.032
LVQ_SMOTE	1491	1769	0.843
MDO	1151	1769	0.651
MWMOTE	689	1769	0.389
NDO_sampling	20	1769	0.011
NEATER	501	3538	0.142
NRAS	327	1769	0.185
NT_SMOTE	311	1769	0.176
OUPS	990	1769	0.560
PDFOS	1217	1769	0.688
ProWSyn	685	1769	0.387
Random_SMOTE	472	1769	0.267
ROSE	898	1769	0.508
RWO_sampling	0	1769	0.000
Safe_Level_SMOTE	907	1769	0.513
SDSMOTE	216	1769	0.122
Selected_SMOTE	264	1769	0.149
SL_graph_SMOTE	220	1769	0.124
SMOBD	416	1769	0.235
SMOTE_Cosine	685	1769	0.387
SMOTE_D	238	1743	0.137
SMOTE_FRST_2T	269	2065	0.130
SMOTE_IPF	219	1769	0.124
SMOTE_OUT	153	1769	0.086
SMOTE_PSO	22	6453	0.003
SMOTE_PSOBAT	0	176	0.000
SMOTE_TomekLinks	209	1691	0.124
SMOTE	195	1769	0.110
SN_SMOTE	98	1769	0.055
SOI_CJ	116	1769	0.066
SSO	341	1765	0.193
SUNDO	41	41	1.000
Supervised_SMOTE	225	1769	0.127
SVM_balance	205	1769	0.116
SYMPROD	518	1769	0.293
V_SYNTH	1499	1769	0.847
MSMOTE	561	1769	0.317

TABLE III
RESULTS OF THORACIC SURGERY DATASET USED TO VALIDATE OVERSAMPLING METHODS UTILIZING A 25% CONCEALED PERCENTAGE.

Thoracic Surgery			
Method	**No. Errors**	**No. Added**	**Error**
ADASYN	67	230	0.291
ADOMS	70	230	0.304
AHC	69	69	1.000
AMSCO	61	174	0.351
AND_SMOTE	37	230	0.161
ANS	93	230	0.404
ASMOBD	82	230	0.357
Assembled_SMOTE	58	230	0.252
Borderline_SMOTE1	66	230	0.287
Borderline_SMOTE2	19	230	0.083
CBSO	72	230	0.313
CCR	191	300	0.637
CE_SMOTE	54	230	0.235
cluster_SMOTE	52	230	0.226
CURE_SMOTE	62	230	0.270
DE_oversampling	45	189	0.238
DEAGO	225	230	0.978
distance_SMOTE	84	230	0.365
DSMOTE	128	230	0.557
DSRBF	59	230	0.257
Edge_Det_SMOTE	58	230	0.252
G_SMOTE	65	230	0.283
GASMOTE	61	235	0.260
Gaussian_SMOTE	152	230	0.661
polynom_fit_SMOTE_star	93	210	0.443
polynom_fit_SMOTE_bus	84	207	0.406
polynom_fit_SMOTE_poly	204	230	0.887
polynom_fit_SMOTE_mesh	103	230	0.448
Gazzah	12	32	0.375
KernelADASYN	161	230	0.700
Lee	28	230	0.122
LN_SMOTE	28	230	0.122
LVQ_SMOTE	111	230	0.483
MDO	86	230	0.374
MSYN	69	345	0.200
MWMOTE	79	230	0.343
NDO_sampling	20	230	0.087
NEATER	142	460	0.309
NT_SMOTE	102	230	0.443
OUPS	129	230	0.561
PDFOS	123	230	0.535
ProWSyn	56	230	0.243
Random_SMOTE	109	230	0.474
ROSE	91	230	0.396
RWO_sampling	19	230	0.083
Safe_Level_SMOTE	145	230	0.630
SDSMOTE	52	230	0.226
Selected_SMOTE	91	230	0.396
SL_graph_SMOTE	75	230	0.326
SMOBD	88	230	0.383
SMOTE_Cosine	63	230	0.274
SMOTE_D	73	234	0.312
SMOTE_FRST_2T	67	235	0.285
SMOTE_IPF	63	230	0.274
SMOTE_OUT	73	230	0.317
SMOTE_PSO	89	873	0.102
SMOTE_PSOBAT	161	460	0.350
SMOTE_TomekLinks	47	198	0.237
SMOTE	62	230	0.270
SN_SMOTE	47	230	0.204
SSO	87	230	0.378
SUNDO	19	19	1.000
Supervised_SMOTE	17	230	0.074
SVM_balance	68	230	0.296
V_SYNTH	170	230	0.739
MSMOTE	79	230	0.343
MSMOTE	561	1769	0.317

References

[1] S. Belarouci, M. A. Chikh, Medical imbalanced data classification, Advances in Science, Technology and Engineering Systems Journal 2 (3) (2017) 116–124.

[2] L. Liu, X. Wu, S. Li, Y. Li, S. Tan, Y. Bai, Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection, BMC Medical Informatics and Decision Making 22 (1) (2022) 1–16.

[3] J. Bi, C. Zhang, An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme, Knowledge-Based Systems 158 (2018) 81–93.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[5] A. S. Tarawneh, A. B. Hassanat, G. A. Altarawneh, A. Almuhaimeed, Stop oversampling for class imbalance learning: A review, IEEE Access 10 (2022) 47643–47660.

[6] K. K. Hauner, R. E. Zinbarg, W. Revelle, A latent variable model approach to estimating systematic bias in the oversampling method, Behavior Research Methods 46 (3) (2014) 786–797.

[7] P. Branco, L. Torgo, R. P. Ribeiro, A survey of predictive modeling on imbalanced domains, ACM Computing Surveys (CSUR) 49 (2) (2016) 1–50.

[8] P. Fergus, M. Selvaraj, C. Chalmers, Machine learning ensemble modelling to classify caesarean section and vaginal delivery types using cardiotocography traces, Computers in biology and medicine 93 (2018) 7–16.

[9] M. Al-Nashashibi, W. Hadi, N. El-Khalili, G. Issa, A. AlBanna, A new two-step ensemble learning model for improving stress prediction of automobile drivers, The International Arab Journal of Information Technology 18 (6) (2021) 819–829.

[10] D. Elreedy, A. F. Atiya, F. Kamalov, A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning, Machine Learning (2023) 1–21.

[11] S. Sreejith, H. K. Nehemiah, A. Kannan, Clinical data classification using an enhanced smote and chaotic evolutionary feature selection, Computers in Biology and Medicine 126 (2020) 103991.

[12] M. Naseriparsa, A. Al-Shammari, M. Sheng, Y. Zhang, R. Zhou, Rsmote: improving classification performance over imbalanced medical datasets, Health information science and systems 8 (2020) 1–13.

[13] A. M. Sowjanya, O. Mrudula, Effective treatment of imbalanced datasets in health care using modified smote coupled with stacked deep learning algorithms, Applied Nanoscience (2022) 1–12.

[14] S. Fotouhi, S. Asadi, M. W. Kattan, A comprehensive data level analysis for cancer diagnosis on imbalanced data, Journal of biomedical informatics 90 (2019) 103089.

[15] M. Sharma, J. Tiwari, U. R. Acharya, Automatic sleep-stage scoring in healthy and sleep disorder patients using optimal wavelet filter bank technique with eeg signals, International journal of environmental research and public health 18 (6) (2021) 3087.

[16] S. Roy, M. Tyagi, V. Bansal, V. Jain, Svd-clahe boosting and balanced loss function for covid-19 detection from an imbalanced chest x-ray dataset, Computers in Biology and Medicine 150 (2022) 106092.

[17] J. Shen, J. Wu, M. Xu, D. Gan, B. An, F. Liu, A hybrid method to predict postoperative survival of lung cancer using improved smote and adaptive svm, Computational and mathematical methods in medicine 2021 (2021).

[18] A. B. Hassanat, A. S. Tarawneh, S. S. Abed, G. A. Altarawneh, M. Alrashidi, M. Alghamdi, Rdpvr: Random data partitioning with voting rule for machine learning from class-imbalanced datasets, Electronics 11 (2) (2022) 228.

[19] J. Liu, Z. S. Wong, H.-Y. So, K. L. Tsui, Evaluating resampling methods and structured features to improve fall incident report identification by the severity level, Journal of the American Medical Informatics Association 28 (8) (2021) 1756–1764.

[20] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, S. Annamalai, Cervical cancer identification with synthetic minority oversampling technique and pca analysis using random forest classifier, Journal of medical systems 43 (2019) 1–19.

[21] E. M. Karabulut, T. Ibrikci, Effective automated prediction of vertebral column pathologies based on logistic model tree with smote preprocessing, Journal of medical systems 38 (2014) 1–9.

[22] V. P. K. Turlapati, M. R. Prusty, Outlier-smote: A refined oversampling technique for improved detection of covid-19, Intelligence-based medicine 3 (2020) 100023.

[23] G. E. Batista, R. C. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD explorations newsletter 6 (1) (2004) 20–29.

[24] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International conference on intelligent computing, Springer, 2005, pp. 878–887.

[25] H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), IEEE, 2008, pp. 1322–1328.

[26] J. Wang, M. Xu, H. Wang, J. Zhang, Classification of imbalanced data by using the smote algorithm and locally linear embedding, in: 2006 8th international Conference on Signal Processing, Vol. 3, IEEE, 2006, pp. 1–4.

[27] J. De La Calleja, O. Fuentes, A distance-based over-sampling method for learning from imbalanced data sets., in: FLAIRS Conference, 2007, pp. 634–635.

[28] S. Gazzah, N. E. B. Amara, New oversampling approaches based on polynomial fitting for imbalanced data sets, in: 2008 the eighth iapr international workshop on document analysis systems, IEEE, 2008, pp. 677–684.

[29] S. Tang, S.-P. Chen, The generation mechanism of synthetic minority class examples, in: 2008 International Conference on Information Technology and Applications in Biomedicine, IEEE, 2008, pp. 444–447.

[30] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, in: Pacific-Asia conference on knowledge discovery and data mining, Springer, 2009, pp. 475–482.

[31] S. Hu, Y. Liang, L. Ma, Y. He, Msmote: Improving classification performance when training data is imbalanced, in: 2009 second international workshop on computer science and engineering, Vol. 2, IEEE, 2009, pp. 13–17.

[32] L. Chen, Z. Cai, L. Chen, Q. Gu, A novel differential evolution-clustering hybrid resampling algorithm on imbalanced datasets, in: 2010 Third International Conference on Knowledge Discovery and Data Mining, IEEE, 2010, pp. 81–85.

[33] S. Wang, Z. Li, W. Chao, Q. Cao, Applying adaptive over-sampling technique based on data density and cost-sensitive svm to imbalanced learning, in: The 2012 International Joint Conference on Neural Networks (IJCNN), IEEE, 2012, pp. 1–8.

[34] S. Cateni, V. Colla, M. Vannucci, Novel resampling method for the classification of imbalanced datasets for industrial and other real-world problems, in: 2011 11th International Conference on Intelligent Systems Design and Applications, IEEE, 2011, pp. 402–407.

[35] X. Fan, K. Tang, T. Weise, Margin-based over-sampling method for learning from imbalanced datasets, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2011, pp. 309–320.

[36] M. A. H. Farquad, I. Bose, Preprocessing unbalanced data using support vector machine, Decision Support Systems 53 (1) (2012) 226–233.

[37] K. Puntumapon, K. Waiyamai, A pruning-based approach for searching precise and generalized region for synthetic minority over-sampling, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2012, pp. 371–382.

[38] S. Barua, M. M. Islam, K. Murase, Prowsyn: Proximity weighted synthetic oversampling technique for imbalanced data set learning, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2013, pp. 317–328.

[39] C. Bunkhumpornpat, S. Subpaiboonkit, Safe level graph for synthetic minority over-sampling techniques, in: 2013 13th International Symposium on Communications and Information Technologies (ISCIT), IEEE, 2013, pp. 570–575.

[40] M. Nakamura, Y. Kajiwara, A. Otsuka, H. Kimura, Lvq-smote–learning vector quantization based synthetic minority over–sampling technique for biomedical data, BioData mining 6 (1) (2013) 1–10.

[41] A. I. Sanchez, E. F. Morales, J. A. Gonzalez, Synthetic oversampling of instances using clustering, International Journal on Artificial Intelligence Tools 22 (02) (2013) 1350008.

[42] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data mining and knowledge discovery 28 (1) (2014) 92–122.

[43] F. Koto, Smote-out, smote-cosine, and selected-smote: An enhancement strategy to handle imbalance in data level, in: 2014 International Conference on Advanced Computer Science and Information System, IEEE, 2014, pp. 280–284.

[44] T. Maciejewski, J. Stefanowski, Local neighbourhood extension of smote for mining imbalanced data, in: 2011 IEEE symposium on computational intelligence and data mining (CIDM), IEEE, 2011, pp. 104–111.

[45] S. Barua, M. M. Islam, X. Yao, K. Murase, Mwmote–majority weighted minority oversampling technique for imbalanced data set learning, IEEE Transactions on knowledge and data engineering 26 (2) (2012) 405–425.

[46] M. Gao, X. Hong, S. Chen, C. J. Harris, E. Khalaf, Pdfos: Pdf estimation based over-sampling for imbalanced two-class problems, Neurocomputing 138 (2014) 248–259.

[47] H. Zhang, M. Li, Rwo-sampling: A random walk over-sampling approach to imbalanced data classification, Information Fusion 20 (2014) 99–116.

[48] B. A. Almogahed, I. A. Kakadiaris, Neater: filtering of over-sampled data using non-cooperative game theory, Soft Computing 19 (11) (2015) 3301–3322.

[49] C. Bellinger, N. Japkowicz, C. Drummond, Synthetic oversampling for advanced radioactive threat detection, in: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), IEEE, 2015, pp. 948–953.

[50] S. Gazzah, A. Hechkel, N. E. B. Amara, A hybrid sampling method for imbalanced data, in: 2015 IEEE 12th International Multi-Conference on Systems, Signals & Devices (SSD15), IEEE, 2015, pp. 1–6.

[51] J. A. Sáez, J. Luengo, J. Stefanowski, F. Herrera, Smote–ipf: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering, Information Sciences 291 (2015) 184–203.

[52] B. Tang, H. He, Kerneladasyn: Kernel based adaptive synthetic data generation for imbalanced learning, in: 2015 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2015, pp. 664–671.

[53] Z. Xie, L. Jiang, T. Ye, X. Li, A synthetic minority oversampling method based on local densities in low-dimensional space for imbalanced learning, in: International Conference on Database Systems for Advanced Applications, Springer, 2015, pp. 3–18.

[54] A. B. Hassanat, Dimensionality invariant similarity measure, arXiv preprint arXiv:1409.0923 (2014).

[55] A. Hassanat, E. Alkafaween, A. S. Tarawneh, S. Elmougy, Applications review of hassanat distance metric, in: 2022 International Conference on Emerging Trends in Computing and Engineering Applications (ETCEA), IEEE, 2022, pp. 1–6.

[56] H. A. Abu Alfeilat, A. B. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. Eyal Salman, V. S. Prasath, Effects of distance measure choice on k-nearest neighbor classifier performance: a review, Big data 7 (4) (2019) 221–248.

[57] R. Ehsani, F. Drabløs, Robust distance measures for knn classification of cancer data, Cancer informatics 19 (2020) 1176935120965542.

[58] C. R. Kancharla, J. Vankeirsbilck, D. Vanoost, J. Boydens, H. Hallez, Latent dimensions of auto-encoder as robust features for inter-conditional bearing fault diagnosis, Applied Sciences 12 (3) (2022) 965.

[59] R. Veerachamy, R. Ramar, Agricultural irrigation recommendation and alert (aira) system using optimization and machine learning in hadoop for sustainable agriculture, Environmental Science and Pollution Research (2021) 1–20.

[60] M. Farooq, S. Sarfraz, C. Chesneau, M. U. Hassan, M. A. Raza, R. A. K. Sherwani, F. Jamal, Computing expectiles using k-nearest neighbours approach, Symmetry 13 (4) (2021) 645.

[61] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39 (2) (2008) 539–550.

[62] A. B. Hassanat, V. S. Prasath, M. Al-kasassbeh, A. S. Tarawneh, A. J. Al-shamailh, Magnetic energy-based feature extraction for low-quality fingerprint images, Signal, Image and Video Processing 12 (8) (2018) 1471–1478.

View publication stats