SIMPOR: SUMMARY OF CHANGES

We are immensely pleased to be offered the helpful suggestions, and have thoroughly revised the manuscript according to the reviewers' comments. In this revision, we have taken the opportunity to address reviewers' concerns that were kindly drawn to our attention. We thoroughly considered each comment and made changes to clarify in the manuscript accordingly. The following items summarize the main changes in the latest revision conducted to the paper.

1) A number of sections in the manuscript including Abstract, Introduction, Related Work, Methodology and Experiment have been revised in terms of language to improve the reading comprehension.
2) Section Related Work was re-organized to improve reading comprehension. Specifically, we categorized the related work into three categories, i.e., sampling-based approach, cost sensitive learning approach and ensemble learning approach.
3) The Experiment section was rewritten to improve reading comprehension and added more results.
4) To avoid generating more noise due to borderline and noisy samples, we added an extra step to the proposed technique. This step enforces a policy to reject the minority candidates which are selected to generate synthetic data. The main idea is to reject candidates surrounded mostly by other class samples, which are likely to be mislabelled.
5) Additional to 5 current experimental techniques (including oversampling and under-sampling techniques), we compared with two more approaches, i.e., SVM Cost Sensitive and Easy Ensemble for a better comparison.
6) The experiments on 41 Datasets have been re-conducted, and the results were updated accordingly. Several additional results have been added to the Experiment Section based on the reviewers' suggestions, e.g., Precision, Recall results, processing time, etc.
7) In this revision, we conducted a statistical test (Wilcoxon Singed-Rank Hypothesis Test) based on the results from 41 datasets for each pair of techniques to provide a more convincing comparison.
8) All mistakes and typos from reviewers' comments have been reviewed and addressed.

I. RESPONSE TO REVIEWER 1

Reviewer's Comment. This study proposed SIMPOR, an oversampling method for dealing with imbalanced data. The topic of the article is an interesting one. The article is presented clearly, and easy to follow.

Response: The authors would like to express sincere thanks to the reviewer for the thorough and constructive comments. In what follows, we present detailed comments in response to the individual points raised by the reviewer and elaborate on how the manuscript has been revised.

Comment 1.a There are currently many approaches to handle the imbalanced data such as sampling method (under, over, and hybrid), cost-sensitivity learning, and ensemble methods. Therefore, the author needs to review the above approaches.

Response: Thank you for the insightful comment. In the previous version, we did review recent methods. However, we looked at the methods from different aspects (i.e., model-centric approach and data-centric approach). However, after reviewing the work, we re-categorize the recent approaches following the reviewer's suggestion (i.e., sampling-based approach, cost-sensitive learning and Ensemble learning approaches). By looking at the related work from this angle, audiences might be easier to follow the background of the work. The revised details are described in Chapter VII (Related Work). The chapter's header is briefly listed as follows.

"In the last few decades, many solutions have been proposed to alleviate the negative impacts of data imbalance in machine learning. However, most of them are not efficiently extended for deep learning. This section reviews algorithms to tackle class imbalanced data that can be extended for deep learning. These techniques can be categorized into three main categories, i.e., sampling, cost-sensitive and ensemble learning approaches.

...

A. Sampling-based approach.

...

B. Cost-sensitive learning approach.

...

C. Ensemple learning approach.

...

"

Comment 1 In table 2, why are these parameters chosen?

Response: Thank you for the insightful comment. In this revision, to clarify, we added details on how we selected the parameters for Table II as follows.

" The classifiers are constructed by neural networks with the input and output sizes corresponding to the number of datasets' features and unique labels. We use the same classifier structure (number of hidden layers, number of neurons in each layer, learning rate, optimizer) for all compared datasets. The detail of neural network implementation is described in Table II. For baseline technique settings, we follow the experimental parameter sets in [4] as we share very similar datasets and comparison techniques. "

Comment 2 The processing time should be compared among the experimental methods.

Response: Thank you for the suggestion. In this revision, in addition to the time complexity analysis, we also added the actual processing time (in seconds) to compare experimental methods as suggested. The detail is added to Section V (Algorithm Time Complexity). The processing time table is added as follows.

"

A. Processing Time.

To explore more on how the techniques perform, we visualize the To better evaluate the technique, we record the data processing time of resampling-based methods over 41 datasets. We don't compare to the other approaches, i.e., cost-sensitive learning and ensemble learning, because they only need negligible data processing time as they focus on classifiers other than improving the data. The processing time was recorded from our machine, which uses an Intel i7 32-thread processor and two NVIDIA 3090 Ti GPUs. Table X shows the recorded processing time over 41 datasets. Overall, our technique takes longer than other techniques as we have to compute the kernel estimation for each data point as mentioned in Section V. From the table, GDO is the second slower technique, and ROS is the fastest one among compared ones. In other words, the proposed technique is slower, but it provides better F1 and AUC scores than others.

"

Comment 3 The classes have been balanced after using the oversampling method, so the author needs to provide the accuracy to compare the experimental methods.

Response: Thank you for the suggestion, and sorry for the confusion. Our testing data are taken from the original datasets before any imbalance handler is applied. Thus, they are possibly class imbalanced. For that reason, we consider F1-score a suitable metric to evaluate the performance. To avoid confusion, we added detail on how the testing results are reported in Section VI (Experiments). The changes are briefly described as follows.

TABLE X: Processing time (in seconds) over 41 datasets.

	SIMPOR	GDO	SMOTE	BL-SMOTE	ADASYN	ROS
glass1	0.1147	0.0576	0.0020	0.0033	0.0032	0.0007
wisconsin	2.0805	0.1769	0.0024	0.0044	0.0046	0.0007
pima	0.2032	0.2066	0.0025	0.0049	0.0050	0.0006
glass0	0.2157	0.0553	0.0023	0.0035	0.0036	0.0009
yeast1	0.2457	0.4749	0.0035	0.0108	0.0104	0.0008
haberman	0.0517	0.1560	0.0022	0.0033	0.0036	0.0008
vehicle1	0.4365	0.1237	0.0025	0.0059	0.0059	0.0007
vehicle2	6.2913	0.1512	0.0029	0.0053	0.0061	0.0010
vehicle3	0.2821	0.1237	0.0024	0.0060	0.0061	0.0007
creditcard	2.1200	0.3783	0.0087	0.0184	0.0182	0.0017
glass-0-1-2-3_vs_4-5-6	0.3376	0.0459	0.0023	0.0035	0.0035	0.0008
vehicle0	7.3645	0.1198	0.0024	0.0054	0.0058	0.0007
ecoli1	0.0418	0.0337	0.0010	0.0018	0.0017	0.0004
new-thyroid1	0.5352	0.0304	0.0015	0.0024	0.0024	0.0006
new-thyroid2	0.3881	0.0359	0.0025	0.0033	0.0031	0.0009
ecoli2	0.2516	0.0266	0.0011	0.0017	0.0016	0.0004
glass6	0.3196	0.0268	0.0014	0.0025	0.0023	0.0006
yeast3	0.1374	0.2422	0.0023	0.0060	0.0059	0.0009
ecoli3	0.0658	0.0378	0.0015	0.0025	0.0024	0.0006
page-blocks0	7.9654	2.0918	0.0045	0.0143	0.0138	0.0015
yeast-2_vs_4	2.4310	0.0624	0.0017	0.0028	0.0028	0.0007
yeast-0-5-6-7-9_vs_4	0.0868	0.0632	0.0016	0.0029	0.0027	0.0007
vowel0	4.7675	0.1312	0.0018	0.0039	0.0037	0.0008
glass-0-1-6_vs_2	0.0482	0.0207	0.0013	0.0023	0.0022	0.0006
glass2	0.0501	0.0227	0.0013	0.0024	0.0024	0.0006
yeast-1_vs_7	0.4697	0.0420	0.0017	0.0026	0.0026	0.0007
glass4	0.1141	0.0197	0.0012	0.0024	0.0023	0.0006
ecoli4	0.1087	0.0310	0.0015	0.0024	0.0024	0.0006
page-blocks-1-3_vs_4	1.8742	0.0445	0.0015	0.0027	0.0026	0.0007
abalone9-18	2.9722	0.0716	0.0015	0.0028	0.0026	0.0006
yeast-1-4-5-8_vs_7	0.0881	0.0673	0.0017	0.0031	0.0028	0.0006
glass5	0.2815	0.0241	0.0017	0.0033	0.0036	0.0008
yeast-2_vs_8	0.1239	0.0441	0.0016	0.0027	0.0028	0.0007
car_eval_4	0.4381	0.1746	0.0026	0.0066	0.0049	0.0012
wine_quality	0.1622	0.8587	0.0030	0.0144	0.0137	0.0015
yeast_me2	0.1060	0.1379	0.0018	0.0042	0.0039	0.0008
yeast4	0.1083	0.1386	0.0018	0.0041	0.0039	0.0008
yeast-1-2-8-9_vs_7	0.0924	0.0757	0.0017	0.0031	0.0030	0.0008
yeast5	0.1188	0.1312	0.0019	0.0037	0.0040	0.0008
yeast6	0.0613	0.1419	0.0018	0.0037	0.0036	0.0008
abalone19	0.0890	0.3161	0.0022	0.0053	0.0054	0.0012

“... Considering each imbalanced dataset as a classification problem, we use the classification testing performance for the technique comparison. Each dataset is randomly split into two parts, 80% for training and 20% for testing. The classifiers are trained on training sets after applying the techniques. The results are reported on the raw testing sets (There isn’t any technique applied on the testing sets; thus, they are also possibly class imbalanced). We use F1-score and AUC for the evaluation metrics as they are suitable and widely used to evaluate imbalanced data. Reported testing results for each dataset are the averages of 5 experimental trials. ...”

II. RESPONSE TO REVIEWER 2

Comment 1.a What attributes does the MOON dataset have?

Response: Thank you for the insightful comment, and sorry for the confusion. In this revision, the experiment settings section was revised to improve reading comprehension. For the MOON dataset, the data generation is explained as follows.

"... We implement techniques on an artificial 2-dimension dataset for demonstration purposes. We first generate the balanced synthetic MOON dataset using python library *sklearn.datasets.make_moons*. The generated MOON contains 3000 samples labeled in two classes, and each instance has two numerical features with values ranging from 0 to 1. We then make the dataset artificially imbalanced with an Imbalance Ratio of 7:1 by randomly removing 1285 samples from one class. ..."

Comment 1.b There were 41 actual real-world datasets from KEEL, UCI, and Credit Card Fraud. In these cases, data classes belonging to each data category should be specifically indicated. So that readers can figure out the imbalance dataset precisely. It would be preferable to cite them appropriately; Recheck reference 19-21. Comment 1.c Table I displays the Imbalance Ratio (IR). It is unclear whether the author created the IR or if it is based on original data; please explain briefly.

Response: Thank you for the recommendation. In this revision, we added the references that directly point to the dataset description sites followed by the data contributors' suggested references in Section VI-C. Besides the imbalance ratio (IR), we added the number of samples per class in Table I to give more precise information. We briefly mentioned how we achieve imbalance ratios in the same section. The corresponding changes are listed as follows.

"... In this section, we compare the proposed technique on 41 real two-class datasets with a variable number of features and Imbalance Ratios, i.e., KEEL datasets [18], [19], UCI datasets fetched from Sklearn tool [20], [21] and Credit Card Fraud [22] dataset. Since the original Credit Card Fraud contains a large number of banking normal and fraud transaction samples (284,807) which significantly reduces our experimental efficiency, we reduced the dataset size by randomly removing normal class transactions to reach an imbalance ratio of 3.0. Other datasets are kept as their original versions after removing bad samples (containing Null values). The datasets are described in Table I.

... "

TABLE I: Dataset Description.

dataset	#samples	#features	IR
glass1	214	9	1.8 (138:76)
wisconsin	683	9	1.9 (444:239)
pima	768	8	1.9 (500:268)
glass0	214	9	2.1 (144:70)
yeast1	1484	8	2.5 (1055:429)
haberman	306	3	2.8 (225:81)
vehicle1	846	18	2.9 (629:217)
vehicle2	846	18	2.9 (628:218)
vehicle3	846	18	3.0 (634:212)
creditcard	1968	30	3.0 (1476:492)
glass-0-1-2-3_vs_4-5-6	214	9	3.2 (163:51)
vehicle0	846	18	3.3 (647:199)
ecoli1	336	7	3.4 (259:77)
new-thyroid1	215	5	5.1 (180:35)
new-thyroid2	215	5	5.1 (180:35)
ecoli2	336	7	5.5 (284:52)
glass6	214	9	6.4 (185:29)
yeast3	1484	8	8.1 (1321:63)
ecoli3	336	7	8.6 (301:35)
page-blocks0	5472	10	8.8 (4913:559)
yeast-2_vs_4	514	8	9.0 (463:51)
yeast-0-5-6-7-9_vs_4	528	8	9.4 (477:51)
vowel0	988	13	10.0 (898:90)
glass-0-1-6_vs_2	192	9	10.3 (175:17)
glass2	214	9	11.6 (197:17)
yeast-1_vs_7	459	7	14.3 (429:30)
glass4	214	9	15.5 (201:13)
ecoli4	336	7	15.8 (316:20)
page-blocks-1-3_vs_4	472	10	15.9 (444:28)
abalone9-18	731	8	16.4 (689:42)
yeast-1-4-5-8_vs_7	693	8	22.1 (663:30)
glass5	214	9	22.8 (205:9)
yeast-2_vs_8	482	8	23.1 (462:20)
car_eval_4	1728	21	25.6 (1663:65)
wine_quality	4898	11	25.8 (4715:183)
yeast_me2	1484	8	28.0 (1433:51)
yeast4	1484	8	28.1 (1433:51)
yeast-1-2-8-9_vs_7	947	8	30.6 (917:30)
yeast5	1484	8	32.7 (1440:44)
yeast6	1484	8	41.4 (1449:35)
abalone19	4174	8	129.4 (689:42)

Comment 2.a In A. Experimental Setup, What dataset is being used, and what is the total size of the data? Where is the testing accuracy? If the conditions are the same for all the models, what is the model's time complexity? It would be preferable to provide Precision and Recall statistics, so that readers may accurately gauge how well the model performs with the proposed approach.

Response: Thank you for your comment. The Experimental Setup includes our general settings for the proposed technique and evaluation settings on the 41 experimental datasets. To improve reading comprehension and avoid confusion, we rewrote the section and clarified every section with bold headlines. Besides, we provide the time complexity analysis in Section V. The Precision and Recall statistics are also added in Section VI-C1, Table VI and VII. The respective Experimental Setup, Time Complexity, and Statistic Results are briefly listed as follows.

Experimental Setup:

"

B. Experimental Setup

This section describes the general settings and implementation details for the experimental techniques.

1) SIMPOR settings: In order to find the informative subset, we leverage entropy-based active learning. We first utilize a neural network model playing a role as a classifier to find high-entropy samples. The detailed steps are introduced in Subsection II-C. The model contains two fully connected hidden layers with *relu* activation functions and 100 neurons in each layer. The output layer applies the soft-max activation function. The model is trained in a maximum of 300 epochs with an early stop option until the loss does not change after updating weights. The model is trained firstly on a random set of data; this model is then used to predict the remaining data and compute the sample entropy. We select the top 30 percent of high-entropy samples (IP=0.3) for the informative subsets. Note that the classifier for finding informative subsets differs from the classifiers for the evaluation after all balancing techniques are applied to the data.

To solve the optimization problem in Equation 16 for finding optima (this differs from the classification optimization for the evaluation) introduced in Section IV-B, we use a gradient ascent method with the gradient rate of $1e-5$ and the maximum iteration of 300.

2) Evaluation Classification settings: Considering each imbalanced dataset as a classification problem, we use the classification testing performance for the technique comparison. Each dataset is randomly split into two parts, 80% for training and 20% for testing. The classifiers are trained on training sets after applying the techniques. The results are reported on the raw testing sets (There isn't any technique applied on the testing sets; thus, they are also possibly class imbalanced). We use F1-score and AUC for the evaluation metrics as they are suitable and widely used to evaluate imbalanced data. Reported testing results for each dataset are the averages of 5 experimental trials.

The classifiers are constructed by neural networks with the input and output sizes corresponding to the number of datasets' features and unique labels. We use the same classifier structure (number of hidden layers, number of neurons in each layer, learning rate, optimizer) for all compared datasets. The detail of neural network implementation is described in Table II. For baseline technique settings, we follow the experimental parameter sets in [4] as we share very similar datasets and comparison techniques.

...”

Time Complexity:

“

The costly part of SIMPOR is that each synthetic sample requires computing a kernel density estimation of the entire dataset. Elaborately, let n be the number of samples of the dataset. In the worst case, the numbers of samples of minority and majority class are $N_B = 1$ and $N_A = n - 1$, respectively. We need to generate $n - 2$ synthetic samples to balance the dataset completely. Since each generated sample must loop through the entire dataset of size n to estimate the density, the algorithm complexity is $O(n^2)$.

Although generating synthetic data is only a one-time process, and this does not affect the classification efficiency in the testing phase, we still try to alleviate its weakness by providing parallelized implementations to reduce the time complexity to $O(n)$. Specifically, each exponential component in Equation 12 is computed parallelly, utilizing GPU or CPU threads. Ellaborately, Equation 12 can be rewritten as N_B components of $e^{\frac{1}{2}(\frac{x-X_{B_i}}{h})^2}$ and N_A components of $e^{\frac{1}{2}(\frac{x-X_{A_i}}{h})^2}$. Fortunately, they are all independent and can be processed parallelly. Thus, with a sufficient hardware resource, the consumption time for the kernel density estimation of each synthetic data point is then reduced by $N_A + N_B = n$ times, which significantly simplifies the complexity to $O(n)$.

...”

Presision and Recall statistics:

TABLE VI: Precision results over 41 datasets.

	SIMPOR	GDO	SMOTE	BL-SMOTE	ADASYN	ROS	SVMCS	EE
glass1	**0.755**	0.695	0.714	0.720	0.701	0.699	0.729	0.714
wisconsin	**0.965**	**0.965**	0.962	0.959	0.961	0.964	0.963	0.960
pima	**0.762**	0.727	0.713	0.722	0.731	0.727	0.747	0.755
glass0	**0.816**	0.774	0.790	0.780	0.764	0.766	0.796	0.813
yeast1	**0.734**	0.676	0.688	0.685	0.676	0.688	0.722	0.725
haberman	0.628	0.607	0.614	0.600	0.614	0.581	0.643	**0.646**
vehicle1	0.760	0.767	0.778	**0.785**	0.774	0.756	0.774	0.784
vehicle2	0.967	0.916	0.955	0.969	0.968	0.948	**0.970**	0.964
vehicle3	**0.768**	0.733	0.736	0.739	0.719	0.738	0.745	0.751
creditcard	0.959	0.945	0.952	0.950	0.951	0.956	**0.971**	0.964
glass-0-1-2-3_vs_4-5-6	0.911	0.894	**0.924**	**0.924**	0.921	0.913	0.909	0.920
vehicle0	**0.966**	0.926	0.951	0.952	0.965	0.955	0.965	0.962
ecoli1	0.845	0.819	0.815	0.795	0.799	0.803	**0.848**	0.840
new-thyroid1	0.960	0.940	0.960	0.960	**0.963**	**0.963**	0.960	0.960
new-thyroid2	0.963	**0.976**	0.961	0.963	0.963	0.963	0.961	0.961
ecoli2	**0.924**	0.839	0.903	0.851	0.851	0.867	0.915	0.915
glass6	0.976	0.926	0.929	**0.984**	0.959	0.949	0.982	**0.984**
yeast3	**0.885**	0.778	0.819	0.814	0.821	0.815	0.871	0.883
ecoli3	0.856	0.748	0.773	0.739	0.679	0.763	0.847	**0.859**
page-blocks0	0.927	0.846	0.865	0.844	0.824	0.865	**0.929**	0.921
yeast-2_vs_4	**0.897**	0.821	0.873	0.864	0.865	0.894	0.739	0.803
yeast-0-5-6-7-9_vs_4	**0.839**	0.715	0.731	0.775	0.767	0.726	0.822	0.833
vowel0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
glass-0-1-6_vs_2	0.713	0.636	0.661	0.671	0.730	**0.750**	0.465	0.535
glass2	0.692	0.643	0.718	0.701	**0.724**	0.696	0.642	0.625
yeast-1_vs_7	**0.944**	0.631	0.589	0.623	0.577	0.601	0.932	0.904
glass4	**0.945**	0.784	0.875	0.898	0.875	0.898	0.942	0.918
ecoli4	**0.902**	0.795	0.848	0.862	0.838	0.848	0.877	0.877
page-blocks-1-3_vs_4	**0.997**	0.908	0.937	0.948	0.959	0.927	0.986	0.986
abalone9-18	0.837	0.658	0.650	0.672	0.685	0.676	**0.942**	0.933
yeast-1-4-5-8_vs_7	0.479	0.571	0.548	0.561	0.566	**0.581**	0.479	0.479
glass5	**0.995**	0.852	0.961	0.869	0.842	0.971	0.687	0.687
yeast-2_vs_8	**0.883**	0.674	0.656	0.688	0.629	0.736	**0.883**	**0.883**
car_eval_4	**1.000**	0.954	0.990	0.981	0.990	0.990	**1.000**	**1.000**
wine_quality	**0.809**	0.658	0.673	0.650	0.674	0.684	0.767	0.719
yeast_me2	0.788	0.609	0.662	0.661	0.657	0.620	**0.922**	0.887
yeast4	0.859	0.614	0.604	0.637	0.613	0.632	0.815	**0.888**
yeast-1-2-8-9_vs_7	**0.989**	0.580	0.564	0.602	0.577	0.592	0.988	0.988
yeast5	0.835	0.748	0.833	0.840	0.835	0.816	**0.859**	0.836
yeast6	0.827	0.645	0.616	0.740	0.658	0.691	**0.831**	0.810
abalone19	0.497	0.510	0.513	0.506	0.506	**0.517**	0.497	0.497

TABLE VII: Recall results over 41 datasets.

	SIMPOR	GDO	SMOTE	BL-SMOTE	ADASYN	ROS	SVMCS	EE
glass1	**0.738**	0.693	0.705	0.719	0.702	0.698	0.718	0.697
wisconsin	0.969	**0.978**	0.969	0.967	0.969	0.970	0.966	0.964
pima	**0.746**	0.740	0.720	0.734	0.744	0.735	0.734	0.742
glass0	0.818	0.799	0.807	0.809	0.789	0.792	0.807	**0.821**
yeast1	0.702	0.702	**0.713**	**0.713**	0.702	0.706	0.683	0.685
haberman	0.559	0.622	0.621	0.623	**0.638**	0.590	0.565	0.578
vehicle1	0.759	0.802	0.803	0.798	**0.806**	0.792	0.738	0.755
vehicle2	0.978	0.951	0.926	0.967	**0.979**	0.968	0.972	0.975
vehicle3	0.751	**0.785**	0.749	0.768	0.742	0.762	0.732	0.750
creditcard	**0.947**	0.940	0.942	0.938	0.936	0.935	0.933	0.936
glass-0-1-2-3_vs_4-5-6	0.924	**0.951**	0.929	0.927	0.935	0.925	0.909	0.919
vehicle0	0.954	0.972	0.957	0.973	0.970	**0.980**	0.979	0.959
ecoli1	0.828	0.837	**0.857**	0.849	0.841	0.837	0.851	0.838
new-thyroid1	0.953	**0.972**	0.953	0.953	0.963	0.963	0.953	0.953
new-thyroid2	0.920	**0.981**	0.910	0.920	0.920	0.920	0.910	0.910
ecoli2	0.893	0.891	**0.902**	0.883	0.882	0.897	0.885	0.885
glass6	**0.888**	0.854	0.819	0.843	0.841	0.821	0.823	0.840
yeast3	0.871	**0.897**	0.895	0.881	0.869	0.896	0.869	0.871
ecoli3	0.841	**0.886**	**0.886**	0.867	0.797	0.875	0.822	0.824
page-blocks0	0.888	0.948	0.932	0.951	0.945	**0.952**	0.856	0.894
yeast-2_vs_4	0.867	**0.868**	0.865	0.856	0.858	0.865	**0.868**	0.805
yeast-0-5-6-7-9_vs_4	0.770	0.827	0.706	**0.832**	0.763	0.759	0.728	0.740
vowel0	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
glass-0-1-6_vs_2	0.625	**0.807**	0.725	0.727	0.760	0.709	0.500	0.545
glass2	0.602	0.797	0.853	0.833	**0.875**	0.803	0.607	0.607
yeast-1_vs_7	0.678	**0.744**	0.654	0.656	0.647	0.626	0.662	0.628
glass4	0.873	**0.921**	0.848	0.830	0.848	0.830	0.786	0.832
ecoli4	0.904	**0.936**	0.900	0.901	0.899	0.900	0.902	0.902
page-blocks-1-3_vs_4	0.952	0.993	0.996	0.996	0.997	0.995	**0.999**	**0.999**
abalone9-18	0.631	0.819	**0.846**	0.814	0.835	0.842	0.693	0.716
yeast-1-4-5-8_vs_7	0.500	0.689	0.600	0.615	0.688	**0.697**	0.500	0.500
glass5	0.875	**0.988**	0.865	0.848	0.813	0.898	0.650	0.648
yeast-2_vs_8	0.797	0.778	**0.802**	0.767	0.783	0.790	0.797	0.797
car_eval_4	**1.000**	0.998	**1.000**	0.999	**1.000**	**1.000**	**1.000**	0.994
wine_quality	0.689	**0.704**	0.657	0.672	0.647	0.672	0.602	0.608
yeast_me2	0.627	**0.811**	0.763	0.714	0.719	0.706	0.610	0.609
yeast4	0.663	**0.795**	0.705	0.735	0.734	0.732	0.667	0.643
yeast-1-2-8-9_vs_7	0.620	**0.710**	0.652	0.629	0.667	0.630	0.589	0.580
yeast5	0.862	**0.972**	0.959	0.951	0.963	0.965	0.819	0.836
yeast6	0.735	**0.894**	0.777	0.812	0.854	0.805	0.706	0.730
abalone19	0.500	0.525	0.574	0.546	**0.581**	0.552	0.500	0.500

Comment 2.b Tables IV and V show results that are only slightly different from those of the other existing models while also outperforming the approaching model. If this is the case, what are the scientific merits of this approach when compared to other approaches?

Response:

Thank you for your comment. We agree with the reviewer that, based on the latest results, even though the proposed technique has shown significant improvement over the benchmark models in many datasets, there are scenarios where the differences between the proposed model and the benchmark ones are unclear. However, the results suggest that the technique outperforms others in terms of ”winning times” (i.e., the number of datasets that a technique achieved the highest result). To further evaluate the effectiveness of the proposed technique, in this revision, a Wilcoxon signed-rank hypothesis test has been performed. The test shows that, statistically, the proposed technique significantly improves the overall results. Therefore, with this study, we hope to provide the community with a better technique to tackle the imbalance issue of unseen datasets. The detail of the Wilcoxon test (added in Section VI-D) is described as follows.

“... To further evaluate the effectiveness of the technique, we also performed a Wilcoxon Signed Rank Test [23] on the 41 dataset results (F1 score and AUC). Wilcoxon hypothesis test is relevant to our study as it is a non-parametric statistical test and does not require a specific distribution assumption for the results. On the other hand, 41 data points (corresponding to 41 datasets results) are sufficient to support this test. Our null hypothesis is that the difference between the proposed technique results and those of the other technique is insignificant. Wilcoxon signed-rank test outputs are computed over the 41 dataset results and return a p-value for each technique pair. We then compare the p-value with the significant value $\alpha = 0.05$. Suppose the p-value is smaller than α. In that case, the evidence is sufficient to reject the hypothesis, which means the proposed technique does make a significant difference from the others, and vice versa. Table VIII shows the Wilcoxon p-value results.

TABLE VIII: Wilcoxon Signed Rank Hypothesis Test results.

	p-value	
SIMPOR vs.	F1-score	AUC
GDO	4.96E-03	4.99E-05
SMOTE	2.32E-02	1.27E-04
BL_SMOTE	2.90E-02	3.06E-06
ADASYN	3.30E-02	1.35E-05
ROS	1.39E-02	5.58E-05
SVMCS	1.44E-03	5.80E-05
EE	2.82E-03	3.33E-04

As we can see from Table VIII, the p-values are all smaller than the critical value of 0.05. Thus, the null hypothesis can be rejected as the supporting evidence is sufficient. In other words, the statistical result

shows that the proposed technique makes a significant improvement compared to others.
..."

Comment 2.c In some data attributes, the F1 and AUC are the same, what is the author's contribution/comments in such scenarios?

Response: In some datasets, F1-score and AUC are the same. This indicates that the proposed technique is comparable to other techniques for such datasets. It is worth noting that some oversampling techniques might accidentally create more noise (create synthetic samples in the majority class area instead of the minority region) and reduce the classification performance. The proposed technique, in fact, tends to reduce this mistake by generating synthetic samples toward the area of the minority.

Comment 3 Classification results: What are winning times? How the winning time is calculated, please explain; Table III, 1. Classification results, modify the inverted comma.

Response: Thank you for your comments. The Inverted comma mistakes have been fixed. The "winning times" calculation is already described in Section VI-C1. To improve the reading comprehension, however, we changed the writing and elaborated more on this part in this revision. The detailed explanation can be listed as follows.

"... We also provide the summary of the F1 and AUC scores by "winning times" scores. We count the number of datasets for which a technique achieves the highest scores among the compared techniques and name this number "winning times". For convention, if more than two techniques share the same highest score, the winning times will be increased for each technique. Figure 5 shows a summary of winning times.
..."

Comment 3.a Additionally, the authors mentioned that the proposed technique improves the training performance and alleviates the class imbalance problem; a. This claim needs to substantiate properly as there is no comparison or mention of training performance results or mention of them, and the issue of class imbalance is not adequately addressed.

Response: Thank you for your comment. As imbalanced data suffers a slow convergence rate [1], [2], the convergence rate might be improved when data is balanced. However, this improvement can be applied to other oversampling, and under-sampling techniques since these techniques tend to balance the data.

Thus, the claim might be trivial. On the other hand, training performance is not our focus to improve in this work. Therefore, we carefully removed this claim.

Comment 3.b How will overcome the synthesis of noisy and borderline samples as well as the tendency to interpolate the neighboring minority class? Need to include to justify the scientific merits of the proposed approach

Response: Thank you for your comment. In this revision, we have improved the proposed method by adding an extra step to avoid noisy and borderline samples. This step enforces a policy for rejecting minority sample seeds selected for re-sampling. More specifically, if a minority sample is surrounded by most of another class sample, this is considered a noisy sample and is rejected to be used for oversampling. The added detail is described in the Methodology Section as follows.

"... ***Avoiding synthesis of noise:*** To reduce the chance of misplacing synthetic samples on another class region because of noisy borderline and mislabeled minority samples, we set a policy for rejecting minority candidates which are selected for oversampling. The idea is to reject candidates surrounded mainly by other class samples. More specifically, we count the labels of the candidate's k-nearest neighbors and reject this candidate if there exists a class that its' number of samples is greater than the number of the minority samples). For example, the candidate is rejected when a class-A sample is selected for generating synthetic data, and its 5-nearest neighbors contain four class-B samples and one class-A sample. This is to avoid selecting mislabeled samples and noisy borderline samples for oversampling. ..."

Comment 4 Minor suggestion: a. Please check the word and line spacing such as in Fig 7, caption. b. Figure labelling are blurred, please change it. c. Check the inverted comas in Table II. d. We use a gradient ascent rate of 0.00001 but in Table II it is written as 0.1. Please make it consistent.

Response: Thank you for your comments. The mistakes have been reviewed and addressed.