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) Section V (Methodology) and VII (Experiments) were thoroughly revised, especially sections V.B, VII.A, VII.C, and VII.E. Specifically, we added more information to improve the readability of the manuscript, including formulation descriptions, result explanations, discussion, and implementation details.
2) Section VII (Related Work) was reorganized to improve reading comprehension and moved to the second section. Two suggested papers from the reviewer were also considered and referred to in this section.
3) The prior estimation (was estimated as constants) in Equation 9 of the proposed method has been changed to use the Empirical Bayes method for a better approximation.
4) To have a better comparison, we replaced the Random Oversampling technique with a recently published work (DeepSMOTE) which is a deep learning-based oversampling method.
5) The experiments on 41 datasets were re-conducted according to the changes in Items 3 and 4. Figures 4, 5, 6, and Tables III, IV, V, VI, VII, and VIII were updated to reflect new experimental results.
6) All mistakes and typos from reviewers' comments have been reviewed and addressed. Several sentences thoughout the paper have been rewritten to improve the readability.

I. RESPONSE TO REVIEWER 1

Reviewer's Comment. Comments to the Author This work explores how class imbalanced data affects deep learning and proposes a data balancing technique for mitigation by generating more synthetic data for the minority class. The paper is well-written and easy to read. Several comments are listed below.

Comment 1.1 The detail of entropy-based active learning should be improved. For example, how many samples are used for model training initially? On p.3, the classifier with θ^0 is trained using a random batch of k selected samples. So only k samples are randomly drawn from the training set to train the model? Besides, the authors further state that the informative samples are obtained by selecting k samples based on the top k highest entropy. These two k are identical?

Response 1.1: Thank you for your comments and sorry for any confusions. We would like to response to your question and revise Section III.C as follows.

Active learning requires repeated phases. In the initial phase, a classifier is only trained on a small initial batch of data. The classifier is then used to estimate the remaining data entropy scores. Then a batch of informative samples are collected from the top entropy samples. This batch is concatenated to the training data for the next phase and also added to informative set. This phase is repeated so that the classifier is trained on larger dataset after every phase. For example, if the initial set contains 6 samples and each next informative batch contains 20 samples, the training size at the 4th phase is $6 + 20 * 4 = 86$.

In this revision, the detail of entropy-based active learning implementation under the experimental setting section is enhanced to improve the readability (e.g., including number of initial training samples). Additionally, the "Entropy-based Active Learning paragraph" is rewritten for better readability. The two changed paragraphs are listed below.

(This modification can be viewed at page 4, line 23.)

"' ... We consider a dataset containing N pairs of samples X and corresponding labels y, and a deep neural network with parameter θ. AL implementation requires repeated phases, and a batch of informative data is selected for each phase. At the first phase $t^{(0)}$, a classifier is trained with parameter $\theta^{(0)}$ (Note that this classifier differs from the classifier for the final classification problem) on a random initial batch of labeled samples and use the model $\theta^{(0)}$ to estimate the entropy of the remaining data.

The entropy scores are then estimated for the remaining samples based on Equation 4. The first batch of informative samples is determined by selecting k highest entropy samples. This batch is then concatenated with the initial training data for the training classifier parameter ($\theta^{(1)}$) in the next phase ($t^{(1)}$) and also accumulated to the informative set. In the next phase, similarly, the updated model is

then used to estimate the entropy of the remaining data. The next informative batch is selected and also added to the informative set. Phases are repeated until the number of accumulated informative samples reaches a pre-set informative portion (IP). For example, $IP = 0.3$ will select 30% training samples as informative samples. ”’

“‘ ... 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 (Note that the classifier for finding the informative subset differs from the classifiers for the final classification evaluation after all balancing techniques are applied to the data). The detailed steps are introduced in Section III-C. The model contains two fully connected hidden layers with *relu* activation functions and 10 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 when the loss is not significantly improved after updating weights. The model is trained firstly on a random set of three samples each class (six samples two classes). This model is then used to estimate entropy scores for the remaining data. We then select next 20 highest entropy samples (k=20) for the next informative data batch. This batch is concatenated to the initial batch for updating the classifiers and accumulated to the informative set. The steps are repeated until the informative set reaches desire informative portion (IP). In these experiments, we set IP=0.3 corresponding to 30 percent of the training size selected for the informative set. ”’

Comment 1.2 The authors simplify Eq. (8) by assume that the prior probabilities in the two classes are identical to cancel out this term. I can understand this assumption can make it easy to calculate, but is it reasonable? The priors for the majority and minority classes are apparently different. I suggest the authors consider the empirical priors to conduct experiments.

Response 1.2: Thank you for the insightful comment and suggestion. In this revision, we adopted the Empirical Bayes method to estimate the priors for equations (8) as suggested. The changes are made in the following paragraph (Section V.B) of the manuscript:

(This change can be viewed at page 5, line 8)

“‘

Approximation of priors in Equation 8: Additionally, we estimate the prior probabilities of observing samples in class A ($p(A)$) and class B ($p(B)$) (in Equation 8) by the widely-used Empirical Bayes Method [39] to leverage the existing information from the original data. The estimates are denoted as $\widehat{p(A)}$ and

$\widehat{p(B)}$ respectively.

Equation 8 Approximation: Let X_A and X_B be the subsets of dataset X which contain samples of class A and class B, $X_A = \{x : y = A\}$ and $X_B = \{x : y = B\}$. N_A and N_B are the numbers of samples in X_A and X_B. d is the number of data dimensions. h presents the width parameter of the Gaussian kernel. The posterior ratio for each synthetic sample x' then can be estimated as follows:

$$f = \frac{p(x'|y' = B)\, p(B)}{p(x'|y' = A)\, p(A)} \tag{9}$$

$$\propto \frac{\frac{1}{N_B h^d} \sum_{i=1}^{N_B} (2\pi)^{-\frac{d}{2}}\, e^{\frac{1}{2}(\frac{x' - X_{B_i}}{h})^2}\, \widehat{p(B)}}{\frac{1}{N_A h^d} \sum_{j=1}^{N_A} (2\pi)^{-\frac{d}{2}}\, e^{\frac{1}{2}(\frac{x - X_{A_j}}{h})^2}\, \widehat{p(A)}} \tag{10}$$

$$\propto \frac{\frac{1}{N_B h^d} \sum_{i=1}^{N_B} e^{\frac{1}{2}(\frac{x' - X_{B_i}}{h})^2}\, \widehat{p(B)}}{\frac{1}{N_A h^d} \sum_{j=1}^{N_A} e^{\frac{1}{2}(\frac{x - X_{A_j}}{h})^2}\, \widehat{p(A)}} \tag{11}$$

”””

Comment 1.3 The proposed method involves many hyper-parameters. For example, h in Eq. (12). The values of these hyper-parameters and how to obtain them should be presented in the manuscript.

Response 1.3: Thank you for your suggestions. In this revision, we updated missing hyper-parameters and explain how we selected them. For example, the update for the bandwidth parameter h in Eq. 12 is as follows.

(This change be viewed at page 5, line 21.)

“‘ ... ***Selecting bandwidth parameter*** h ***for Gaussian kernel:*** The bandwidth is automatically selected for each dataset using the most common method, namely Scott's rule of thumb, proposed by Scott [40]. With an attempt to minimize the mean integrated squared error, the parameter is estimated as $h = N^{(-\frac{1}{d+4})}$ where N, d are the number of data points and the number of dimensions respectively. This study utilizes a scikitlearn python library for KDE, including bandwidth selection. The implementation detail can be found at [41].

”’

Comment 1.4 The comparison methods should include recently published methods and deep learning methods. Among the comparison methods, only GOD was a method published in 2022. SMOTE,

Boderline-SMOTE, EE, and ADASYN where published before 2010. SVMCS was in 2012.

Response 1.4: Thank you for your recommendation. In the revised version, we have included a comparison with DeepSMOTE, a data balancing technique based on deep learning that was published in 2022. Further details can be found in Section VII (Experiment).

Comment 1.5 In Section VI. E (Data Visualization), how to calculate No. of intersection samples?

Response 1.5: In order to provide further clarification on the calculation of the number of intersection samples, we have added additional details to the paragraph (Section VII.E), which now reads as follows.

"‘ ... A hard-to-differentiate ratio is defined as the ratio of the number of samples in the intersection between 2 classes to the total of minority samples ($HDR = \frac{No.\ Intersection\ samples}{No.\ Minority\ samples} 100\%$) where the number of intersection samples is estimated by counting samples in the overlapped bins between the two classes in the 1D histograms "’

Comment 1.6 . In Fig. 6, how to obtain the densities of the results? If they are obtained by using KDE, are they unfair to the other methods that don't use KDE?

Response 1.6: Thank you for your feedback. The densities presented in Figure 6 were estimated using KDE on the dimensionally reduced data, but this was done only for visualization purposes. All the quantified comparisons, such as HDR, were actually computed based on the 1D histogram, as explained in Comment 1.5. To make this clearer and to avoid any confusion, we have revised the relevant paragraph (Section VII.E) accordingly.

(The modifications can be viewed at page 10, line 17.)

"‘ To explore more on how the techniques perform, we visualize the generated data by projecting them onto lower dimension space (i.e., one and two dimensions) using the Principle Component Analysis technique (PCA) [50]. Data's 2-Dimension (2D) plots and 1-Dimension histograms are presented with a hard-to-differentiate ratio (HDR) for each technique. 1D histograms are computed by dividing one-dimensional-reduced data into 20 bins (intervals) and counting the number of samples within the interval of each bin. A hard-to-differentiate ratio is defined as the ratio of the number of samples in the intersection between 2 classes to the total of minority samples ($HDR = \frac{No.\ Intersection\ samples}{No.\ Minority\ samples} 100\%$) where the number of intersection samples is estimated by counting samples in the overlapped bins between the two classes

in the 1D histograms. This ratio is expected to be as small as 0% if the two classes are well separated; in contrast, 100% indicates that the two classes cannot be distinguished in the projected 1D space. Besides HDR, we show the absolute numbers of Minority, Majority, and Intersection samples for each technique in the bottom tables. From the plots, we observe how the data are distributed in 2D space and quantify samples that are hard to be differentiated in the 1D space histograms. ”’

Comment 1.7 In Fig. 7 & 8, the results show that α and IP do not significantly affect the performance, but they are counter-intuitive. The authors should provide more discussion to explain the results rather than only presenting the results.

Response 1.7: We appreciate your valuable feedback. In this revised version, we have included additional discussion and explanation of the results accordingly in Section VII.G as follows.

(This modification can be viewed at page 13, line 9, and page 13, line 1.)

“‘ ... The figure obtained from the experiment indicates that the r factor, with a radius distribution standard deviation ranging from 0.6R to R, has minimal impact on the classification performance. While there are slight variations within the α range of 0.6 to 1, the performance improves between 0.2 to 0.6 (such as for ecoli1, abalone9-18, and yeast4). This is because the performance mainly depends on the classifier’s decision boundary, and the synthetic data are placed far away from the decision boundary towards the minority class area; thus, the radius does not have much effect on the accuracy results. However, in the case of multi-classed data, the performance might be affected by a significant value of R. ”””

“‘ ... As we can see from the figure, while datasets with outstanding performance (new-thyroid1, ecoli1) have little impact, there are fluctuations in other datasets’ F1-score and AUC score (abalonce9-18, glass0, yeast4). This is because, for the easy-separated dataset such as new-thyroid1 and ecoli1, the IP change does not affect the classification performance as the data classes are easily separated. While in more challenging datasets, IP changes might affect the balance at the informative region; thus, this leads to performance variations. The resulting figure also suggests tuning IP for each dataset between a range of (0.2, 0.6) could achieve higher performance. ”””

Comment 1.8 . In Section VI. F (Processing Time), the sentence ”To explore more ... the To better evaluate the technique” should be re-phrased.

Response 1.8: Thank you for your suggestion. The sentence in your comment has been reworded in the revised version, and we have also rephrased several other sentences to enhance the readability of the manuscript. The suggested sentence was modified as follows.

“‘ Data processing times for oversampling-based approaches on 41 datasets are compared to provide a more comprehensive comparison.... ”””

II. RESPONSE TO REVIEWER 2

Reviewer's Comment. This work explores how class-imbalanced data affects deep learning and proposes a data-balancing technique to mitigate by generating more synthetic data for minority classes. The author has done a lot of work, and the experimental results are very good. The content of the article is substantial. I hope the following comments could be useful for the further improvement of the paper:

Comment 2.1 The language of the manuscript is not academic enough. There are too many "we" in the text. I suggest the writer use the passive voice

Response 2.1:

Thank you for your feedback. In this version, we have thoroughly revised the manuscript according to the comments. Multiple sentences and paragraphs throughout the manuscript have been rephrased to refine the language and enhance its comprehensibility. A few examples are provided below.

Example 1:

"‘ ...We elaborate on how our strategy is developed in the rest of this section. "’’’

is rewritten as

"‘ ...The remainder of this section provides further information about how our approach was developed. "’’’

Example 2:

"‘ ... Because we want to generate neighbors for each minority sample that maximizes Function f in Equation 11, we examine points lying on the sphere centered at the minority sample with a small radius r. As a result, we can find a vector $\vec{v}$ so that it can be added to the sample to generate a new sample. "’’’

is rewritten as

"‘ ...To generate neighbors for each minority sample that maximizes Function f in Equation 11, points on the r-radius sphere centered at a minority sample are considered synthetic instances. As a result, a vector $\vec{v}$ can be added to a minority sample for generating a new instance. "’’’

Comment 2.2 The contribution: 3) We applied our technique to 41 real datasets with a diversity of imbalance ratio and the number of features. I don't think this is a contribution. This is a statement of work.

Response 2.2: Thank you for your feedback. We have revised accordingly by moving the contribution 3 to the statement of work in Section VII as follows.

(The change can be viewed at page 7, line 43.)

‘‘‘ In this section, we explore the techniques via binary classification problems on an artificial dataset (i.e., Moon) and 41 real-world datasets (i.e., KEEL, UCI, Credit Card Fraud) with a diversity of imbalance ratios and different numbers of features. Samples in Moon have two features, while other datasets contain various numbers of features and imbalance ratios. ... ’’’’’

Comment 2.3 AUC is an acronym. It is better to use an acronym after mentioning the full name in the “Introduction”.

Response 2.3:

Thanks for your suggestion. We have included the complete name of the AUC abbreviation in this version (Section I).

(The change can be viewed at page 2, line 36.)

‘‘‘ ... Section III introduces related concepts that will be used in this work, i.e., Imbalance Ratio, Macro F1-score, Area Under the Curve (AUC), and Entropy-based active learning... ’’’’’

Comment 2.4 The letters used in the formulas are not explained clearly, especially formula (9).

Response 2.4: Thanks for your feedback. To enhance the comprehensiveness of the revision, we have included descriptions for every letter in formula (9) (Section V.B). The detail is listed as follows.

(This modification can be viewed in column 2 on page 5.)

‘‘‘ ***Approximation of priors in Equation 8:*** Additionally, we estimate the prior probabilities of observing samples in class A ($p(A)$) and class B ($p(B)$) (in Equation 8) by the widely-used Empirical Bayes Method [39] to leverage the existing information from the original data. The estimates are denoted as $\widehat{p(A)}$ and $\widehat{p(B)}$ respectively.

Equation 8 Approximation: Let X_A and X_B be the subsets of dataset X which contain samples of class A and class B, $X_A = \{x : y = A\}$ and $X_B = \{x : y = B\}$. N_A and N_B are the numbers of samples in X_A and X_B. d is the number of data dimensions. h presents the width parameter of the Gaussian kernel. The posterior ratio for each synthetic sample x' then can be estimated as follows:

$$f = \frac{p(x'|y'=B)\,p(B)}{p(x'|y'=A)\,p(A)} \tag{9}$$

$$\propto \frac{\frac{1}{N_B h^d}\ \sum_{i=1}^{N_B} (2\pi)^{-\frac{d}{2}}\ e^{\frac{1}{2}(\frac{x'-X_{B_i}}{h})^2}\ \widehat{p(B)}}{\frac{1}{N_A h^d}\ \sum_{j=1}^{N_A} (2\pi)^{-\frac{d}{2}}\ e^{\frac{1}{2}(\frac{x-X_{A_j}}{h})^2}\ \widehat{p(A)}} \tag{10}$$

$$\propto \frac{\frac{1}{N_B h^d}\ \sum_{i=1}^{N_B} e^{\frac{1}{2}(\frac{x'-X_{B_i}}{h})^2}\ \widehat{p(B)}}{\frac{1}{N_A h^d}\ \sum_{j=1}^{N_A} e^{\frac{1}{2}(\frac{x-X_{A_j}}{h})^2}\ \widehat{p(A)}} \tag{11}$$

Selecting bandwidth parameter h ***for Gaussian kernel:*** The bandwidth is automatically selected for each dataset using the most common method, namely Scott's rule of thumb, proposed by Scott [40]. With an attempt to minimize the mean integrated squared error, the parameter is estimated as $h = N^{(-\frac{1}{d+4})}$ where N, d are the number of data points and the number of dimensions respectively. This study utilizes a scikitlearn python library for KDE, including bandwidth selection. The implementation detail can be found at [41]. "'

Comment 2.5 More new published methods are suggested to be used as the comparison algorithms

Response 2.5: Thank you for your recommendation. In the revised version, we have included a comparison with DeepSMOTE, a data balancing technique based on deep learning that was published in 2022. (More detail can be viewed in Section VII at Page 7.)

Comment 2.6 "Related work" should be put to the second section of the paper. In addition, the following references are suggested to be present and cited:

1

Wei Feng*, Gabriel Dauphin, Wenjiang Huang, Yinghui Quan, Wenxin g Bao, Mingquan Wu, Qiang Li, "Dynamic synthetic minority over-sampling technique based rotation forest for the classification of imbal a nced hyperspectral data," IEEE Journal of Selected Topics in AppliedEarth Observations and Remote Sensing, vol. 12, no. 7, pp. 2159–2169, 2019.

2

Wei Feng*, Wenjiang Huang and Wenxing Bao, "Imbalanced Hyperspectral Image Classification With

an Adaptive Ensemble Method Based on SMOTE and Rotatio n F orest With Differentiated Sampling Rates," IEEE Geoscience and Remote Sensing Le tters, vol. 16, no. 12, pare suggestedggestedc. 2019.

Response 2.6: Thank you for your suggestions. In this revision, the two suggested papers have been taken into account and are listed as [14] and [15]. Besides, Section "Related Work" (was Section VII in previous version) has been moved to Section II to improve readability of the manuscript.