CLE-SMOTE: Addressing Imbalanced Chest X-Ray Classification with Contrastive Learning-Enhanced SMOTE

Author Name1[*1,2] ABC@SAMPLE.EDU
[1] *Address 1*
[2] *Address 2*

Author Name2[*1] XYZ@SAMPLE.EDU
Author Name3[2] ALPHABETA@EXAMPLE.EDU
Author Name4[†3] UVW@FOO.AC.UK
[3] *Address 3*

Author Name5[*4] FGH@BAR.COM
[4] *Address 4*

Editors: Under Review for MIDL 2024

Abstract

Class imbalance is a prevalent issue in many healthcare tasks, where diseases of interest are exceedingly rare in datasets. This issue is especially relevant in chest X-ray classification, where specific subconditions appear significantly less frequently than others. Synthetic Minority Oversampling Technique (SMOTE) is a widely used oversampling method that addresses class imbalance by generating synthetic minority class examples. While effective, SMOTE occasionally introduces harmful examples into the dataset, hindering model performance. In this work, we introduce Contrastive Learning-Enhanced SMOTE (CLE-SMOTE), a method to identify and reduce the influence of these noisy SMOTE-generated examples. We demonstrate CLE-SMOTE's effectiveness in learning from datasets with severe interclass and intraclass imbalance through a chest X-ray classification task. CLE-SMOTE delivers promising results, substantially outperforming all baselines.

Keywords: class imbalance, chest x-rays, supervised learning, unsupervised learning

1. Introduction

The phenomenon of class imbalance arises when the number of samples in one or more target classes is significantly lower than in others (Johnson and Khoshgoftaar, 2019). Class imbalance is especially notorious in healthcare applications, where rare diseases or conditions may be vastly underrepresented in datasets (Li et al., 2010). This underrepresentation can skew conventional machine learning algorithms, leading to models that largely ignore minority classes and inaccurately predict their occurrences. This has severe consequences in healthcare, as minority classes often represent diseases or conditions of critical interest. To improve healthcare outcomes and ensure equitable treatment, it is imperative to develop robust techniques designed to address the complexities of class imbalance in medical data.

Computer-aided diagnostic methods for chest X-ray (CXR) analysis provide a prime example of how class imbalance disrupts medical image analysis. CXRs play a critical role

[*] Contributed equally
[†] Contributed equally

in medical imaging, as they are crucial for the identification, diagnosis, and management of cardiothoracic and pulmonary pathologies. CXRs are the most frequently performed radiological examination worldwide, with an estimated 837 million CXRs performed annually (on the Effects of Atomic Radiation, 2011). While the widespread demand, accessibility, and utility of CXRs merit celebration, they present a formidable review workload for both radiologists and healthcare professionals (Nakajima et al., 2008; Kawooya, 2012).

Developing computer-aided diagnostic methods for CXRs has been the focus of a multitude research efforts over the past decade, with researchers developing robust algorithms that can detect specific diseases such as pneumonia, fracture, and tuberculosis (Majkowska et al., 2020). More recently, the concept of a CXR abnormality detector has emerged, which classifies a given CXR as either normal or abnormal (Nabulsi et al., 2021).

While a plethora of CXR diagnostic algorithms have been developed, the efficacy of the models in this domain is often hindered by the scarcity of labeled data and the inherent class imbalance within available datasets (Majkowska et al., 2020). Previous research, especially in the context of medical imaging, has largely concentrated on the concept of interclass imbalance. However, an equally critical and often overlooked aspect of class imbalance is intraclass imbalance. In the context of CXR normal/abnormal classification, interclass imbalance refers to the fact that there are typically more normal CXRs than abnormal CXRs in a given dataset. Intraclass imbalance refers to the phenomenon where, within the abnormal class, the distribution of subconditions is uneven, as certain subconditions are significantly underrepresented compared to others. Both manifestations of class imbalance lead to increased misclassification of examples from minority classes and/or subclasses, as a model may not learn the subtle but critical indicators of a rare disease if samples from affected patients are overshadowed by the overwhelming number of examples from other patients. Further, the repercussions of false negatives are significant, potentially resulting in delayed treatment and exacerbated health outcomes for patients.

Previous research has explored various methods to ameliorate the negative impact of interclass imbalance. Common techniques to involve weighting the loss of samples based on their class frequency (Cui et al., 2019), artificially increasing the number of examples in the underrepresented class(es) by oversampling (Shelke et al., 2017), and decreasing the number of examples in the majority class(es) by undersampling (He and Garcia, 2009). Currently, one of the most effective methods to address interclass imbalance is Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic minority class examples through interpolation between randomly chosen minority class neighbors (Chawla et al., 2002). While SMOTE generally performs well in practice, it can occasionally introduce noisy or unhelpful synthetic examples, thereby negatively impacting the learning process (Batista et al., 2004). Further, there is no evidence that SMOTE alleviates the effects of intraclass imbalance.

In this work, we present Contrastive Learning-Enhanced SMOTE (CLE-SMOTE), a method that supplements SMOTE by limiting the influence of noisy, synthetically-generated examples on the network's training, allowing the network to learn robust class representations despite severe levels of both interclass and intraclass imbalance. We evaluate our method on chest X-ray data, simulating various levels of interclass and intraclass imbalance. Our methods demonstrate state-of-the-art results, outperforming all baselines.

2. Materials & Methods

2.1. Overview

In the development of CLE-SMOTE, our objective was to establish an architecture that enables the network to benefit from synthetically generated examples while minimizing the influence of out-of-distribution or noisy examples. CLE-SMOTE consists of three stages: a data augmentation stage using SMOTE, a pretraining stage, and a supervised finetuning stage. An overview schematic of CLE-SMOTE is shown in Figures 1a and 1b.

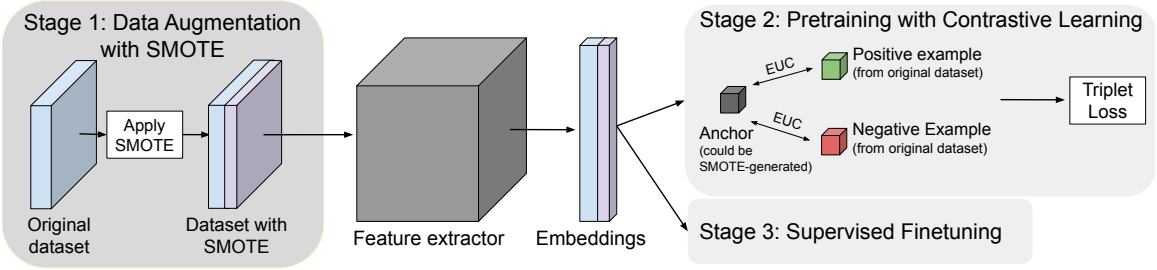

Figure 1a: Overview of the CLE-SMOTE model architecture and training procedure, consisting of three stages. Stage 1 applies Synthetic Minority Oversampling Technique (SMOTE) to augment the original dataset, creating a balanced dataset with synthetic examples. Stage 2 uses contrastive learning for pretraining, where embeddings from a feature extractor are compared using Euclidean distance (EUC). Anchors (which may include SMOTE-generated samples) are paired with positive examples (from the original dataset) and negative examples (from the original dataset), optimizing the embedding space through triplet loss. Stage 3 involves supervised finetuning on the final task.

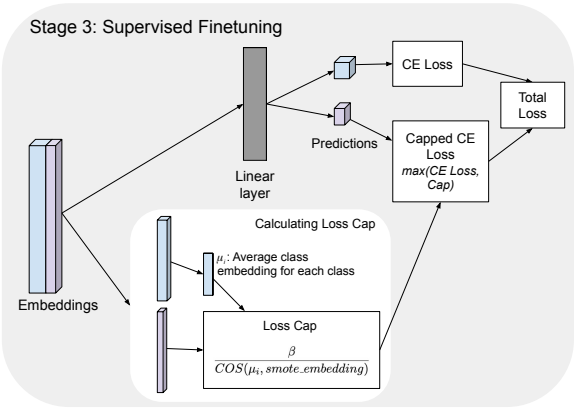

Figure 1b: The supervised fine-tuning stage of CLE-SMOTE, illustrating the method for calculating the loss cap applied to synthetic minority oversampling technique (SMOTE) examples. Input embeddings are processed through a linear layer to generate predictions, which are used to compute the cross-entropy (CE) loss. The capped CE loss is determined as the maximum of the CE loss and the calculated loss cap. The loss cap is computed using the cosine distance (COS) between the average class embedding (μ_i) and the embedding of the SMOTE example, scaled by a factor β. The total loss combines both the capped CE loss and the standard CE loss.

2.2. Stage 1: Data Augmentation with SMOTE

The initial Data Augmentation stage of CLE-SMOTE is conducted prior to any training. In this stage, the dataset is balanced through the application of SMOTE (Chawla et al., 2002), an oversampling algorithm that generates synthetic examples of a dataset's minority class(es). SMOTE creates each synthetic example by interpolating between a random minority class example and one of its k-nearest neighbors; this parameter for nearest neighbors is set to k=5, determined empirically.

2.3. Stage 2: Pretraining using Supervised Contrastive Learning

With a balanced, albeit noisy dataset created, we proceed to the pretraining stage. The goal of pretraining is twofold. First, we aim to enable the network to learn distinct feature representations for each class despite the presence of class imbalance, leveraging contrastive learning over standard supervised methods (Marrakchi et al., 2021). Second, by learning distinct feature spaces for the minority class(es), our network is better equipped to identify out-of-distribution, noisy SMOTE-generated examples and cap their influence on parameter updates during training. To learn these feature spaces, we employ supervised contrastive learning and train on the entire SMOTE-augmented dataset (Khosla et al., 2021).

Specifically, for each example in a batch from our SMOTE-augmented training set (referred to as the anchor), we randomly choose a positive example (an example from the same class as the anchor) and a negative example (an example from a different class as the anchor). We pass each of the three CXR images through our network, skipping the final classification layer to obtain the pre-head embeddings. We compute the Euclidean distance between the anchor and both the positive and negative examples, and use the following loss:

$$L = max(0, EUC(A, P) - EUC(A, N)) \tag{1}$$

where EUC is the euclidean distance, A is the anchor's embedding, P is the embedding of the positive example, and N is the embedding of the negative example. We confine the selection of positive and negative examples to solely originate from non-SMOTE examples.

2.4. Stage 3: Supervised Finetuning

With the network pretrained using supervised contrastive learning, our goal is to tune the model to the classification task of interest. We add a single linear (classification) layer on top of the pretrained embedding network and finetune the entire network using a customized cross-entropy loss (Equation 2). To cap the influence of potentially harmful SMOTE-generated examples, the pre-head embeddings (the output of the second-to-last layer of the network) for each non-SMOTE example are calculated at the start of each epoch. We compute the average embedding for each class and cap the loss for each SMOTE example using the following formula:

$$L_s = min(ce\_loss, \frac{\beta}{COS(\mu_i, smote\_embedding)}) \tag{2}$$

where L_s is the loss of the SMOTE example, ce_loss is the regular cross-entropy loss for the example, COS is the cosine distance, μ_i is the average embedding for the minority class

of the SMOTE example (computed at the start of each epoch), and β is the noise tolerance hyperparameter that regulates the extent to which noisier synthetic examples are capped.

We perform this procedure at the beginning of each epoch because the continuous updates to the network's parameters cause the learned feature spaces for each class to shift throughout the training process.

A higher cosine distance signifies a more substantial deviation from the feature space of its class, indicating that the example is dissimilar from the rest of its class. Our loss function is designed to limit the influence of these out-of-distribution examples on parameter updates. For non-SMOTE examples, vanilla cross-entropy loss (no capping) is used.

The primary feature of the SMOTE loss function is *beta*, the noise tolerance hyperparameter, which controls the extent to which we cap the SMOTE-generated examples. A higher *beta* signifies a higher noise tolerance and less aggressive capping. An illustration depicting how the loss cap is calculated is shown in Figure 1b.

3. Dataset

To evaluate the effectiveness of CLE-SMOTE and benchmark it against other methods on its ability to effectively learn from an imbalanced CXR dataset, we use two public datasets: CheXpert (Irvin et al., 2019) and ChestX-ray14 from the National Institutes of Health (NIH) Clinical Center (Wang et al., 2017). CheXpert serves as the training set for both CLE-SMOTE and the benchmark methods. To evaluate the quality of the trained models, we use both the test split of CheXpert and the test split of ChestX-ray14.

3.1. CheXpert

CheXpert consists of 224,316 chest radiographs of 65,240 patients who underwent a radiographic examination from Stanford Health Care between October 2002 and July 2017. 14 observations were selected based on clinical relevance, comprising the following classes: No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices. Each report was labeled for the presence of the 14 classes as positive, negative, or uncertain. An automated labeling system assigned labels to each observation: blank for unmentioned, 0 for negative, -1 for uncertain, and 1 for positive.

To precisely control interclass and intraclass imbalance ratios during training, we filtered the original training set. We first excluded all lateral CXR images to ensure the exclusivity of frontal CXRs and maintain consistency across all patient data. Due to the low frequency of certain subconditions and the high number of subclasses in the labels, we combined subconditions with similar manifestations. A correlation matrix analysis was conducted among the labels for the 14 observations to identify highly correlated classes. This analysis, consistent with the observations of relevant works (Nabulsi et al., 2021), revealed high co-occurrence rates between Atelectasis, Lung Opacity, Pneumonia, and Consolidation, as well as between Cardiomegaly and Enlarged Cardiomediastinum. Based on these findings, we decided to consolidate these similar classes into broader categories. Specifically, Atelectasis, Lung Opacity, Pneumonia, and Consolidation were merged into a single class termed "Focal/Multifocal Lung Opacity", while Cardiomegaly and Enlarged Cardiomediastinum were amalgamated into "Enlarged Cardiac Silhouette". Consequently, the

individual classes pertaining to these amalgamated categories were eliminated from the dataset, thereby simplifying the classification task.

To further refine the dataset, two types of exceptional cases were identified and filtered out: 1) cases labeled as having a finding but without a specific condition identified (because it could not be determined what subcondition was present), and 2) cases labeled as having no finding but in which a specific condition was identified. Furthermore, CXRs with multiple subconditions present were filtered out, ensuring that only single-condition cases and healthy cases remained for subsequent analysis. After these preliminary steps, Edema, Pleural Effusion, and Pleural Other were found to lack sufficient remaining sample quality and were therefore removed from the dataset. Our final training dataset consisted of six classes—Normal, Lung Lesion, Pneumothorax, Fracture, Focal/multifocal Lung Opacity, and Enlarged Cardiac Silhouette—and 41,052 train samples. A STARD diagram depicting our inclusion/exclusion criteria is shown in Figure 2.

Because we wanted to evaluate the robustness of CLE-SMOTE when trained on an imbalanced dataset, we used the test split of CheXpert as is, with no filtering.

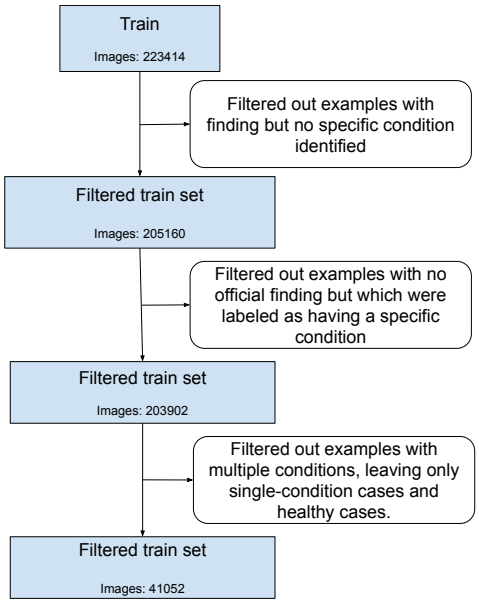

Figure 2: STARD Diagram depicting our inclusion/exclusion criteria for the train split of CheXpert.

ChestX-ray14

For a robust evaluation of the different methods' ability to generalize to other datasets, we used a portion of the test split of the ChestX-ray14 dataset to evaluate the different trained methods. We used the expert labels made available by Nabulsi et al. (2021). As with the test split for CheXpert, we did not filter the dataset.

4. Results & Evaluation

4.1. Architecture & Hyperparameters

For all experiments in this work, we use the same architecture in order to ensure fair comparisons across the different methods. Specifically, we use a ResNet50 (He et al., 2015) with default settings. For each method, we ran a hyperparameter search to determine the optimal learning rate.

4.2. Baselines

To evaluate the performance of CLE-SMOTE, we designed a series of imbalanced classification experiments. To ensure a comprehensive evaluation, we benchmarked CLE-SMOTE against several widely used methods for addressing class imbalance, including the following baselines:

- **Oversampling Minority Class:** Oversampling the minority class is a widely used method when training on imbalanced datasets (Shelke et al., 2017). We randomly oversample all minority classes until all classes have approximately the same number of examples (i.e., until there is no more class imbalance).
- **Weighted Loss:** We apply weighted loss, a common method used where every example's contribution to the loss is weighted based on its class frequency (Cui et al., 2019).
- **Vanilla SMOTE:** We apply SMOTE with no adjustments, as described in the original paper (Chawla et al., 2002). This means that synthetic samples are generated for the minority classes without any filtering, and noisy (out-of-distribution) synthetic examples are not excluded during the process.
- **No Method:** We apply no special methods to balance the class imbalance.

4.3. Experiments

To benchmark CLE-SMOTE against the baselines described in the previous section, we designed and executed 17 classification experiments on subsets of the CheXpert dataset. These experiments were designed to evaluate the efficacy of each data preprocessing technique in addressing five different types of class imbalance.

- **Interclass Imbalance (One Experiment):** This type of class imbalance refers to imbalance *across* classes. We configured the training set with a 20:1 ratio between normal and abnormal classes. Within the abnormal class, we maintained a balanced distribution among the subconditions.
- **Intraclass Imbalance, Single Majority Subclass (Four Experiments):** This type of class imbalance refers to imbalance *within* a single class. In this framework, one subcondition (of the five) within the abnormal class was designated as the majority subclass, while the other four subconditions served as minority subclasses. A 20:1 ratio was maintained between the majority subclass and each minority subclass. To

Experiment	Normal	Abnormal					
		SC 1	SC 2	SC 3	SC 4	SC 5	Total
Interclass	19624	196	196	196	196	196	980
Intraclass (single majority)	2194	87	1759	87	87	87	2147
Intraclass (multiple majority)	3858	62	62	1245	1245	1245	3859
Inter & Intraclass (single majority)	19624	40	817	40	40	40	977
Inter & Intraclass (multiple majority)	19624	15	15	316	316	316	978

Table 1: Example class counts from each of our experiments. SC stands for subcondition, and the abnormal class count is the sum of all subcondition counts.
.

Experiment	Normal	Abnormal					
		SC 1	SC 2	SC 3	SC 4	SC 5	Total
Interclass	100	1	1	1	1	1	5
Intraclass (single majority)	24	20	1	1	1	1	24
Intraclass (multiple majority)	62	20	1	1	20	20	62
Inter & Intraclass (single majority)	480	20	1	1	1	1	24
Inter & Intraclass (multiple majority)	1240	20	1	1	20	20	62

Table 2: Normalized ratios of normal and abnormal class counts across different experimental conditions. For experiments with interclass imbalance (Interclass, Inter & Intraclass with single-class majority, and Inter & Intraclass with multi-class majority), the overall ratio of normal to abnormal cases is 20:1. For experiments without interclass imbalance (Intraclass with single-class majority and Intraclass with multi-class majority), the normal-to-abnormal ratio is balanced at 1:1. Additionally, for experiments containing intraclass imbalance (Intraclass with single-class majority, Intraclass with multi-class majority, Inter & Intraclass with single-class majority, and Inter & Intraclass with multi-class majority), the ratio of majority to minority subclass is 20:1.
.

balance the class counts between the normal and abnormal classes, the training set followed a ratio of 24:20:1 for normal:majority subclass:minority subclass.[1]

- **Intraclass Imbalance, Multiple Majority Subclasses (Four Experiments):** This framework is identical to the intraclass imbalance with a single majority subclass framework, albeit with a distinction: rather than selecting a single subcondition within the abnormal class as the majority subclass, three subconditions were designated as the majority subclasses, with the remaining two serving as minority subclasses. A 20:1 ratio is maintained between the majority and minority subclasses. To balance the normal and abnormal class counts, the training set follows a 62:20:1 ratio for normal:majority subclass:minority subclass.[1]

- **Interclass & Intraclass Imbalance, Single Majority Subclass (Four Experiments):** This framework introduces imbalance both *between* classes and *within* a single class. A 24:1 ratio is maintained between the normal and abnormal classes. Within the abnormal class, one subcondition is designated as the majority subclass, while the other four serve as minority subclasses, with a 20:1 ratio maintained between the majority and minority subclasses. Combined, this results in a training set ratio of 480:20:1 for normal:majority subclass:minority subclass.[1]

- **Interclass & Intraclass Imbalance, Multiple Majority Subclasseses (Four Experiments):** This framework is identical to the interclass & intraclass imbalance with a single majority subclass framework, with an exception: instead of selecting a single subcondition within the abnormal class as the majority subclass, three subconditions were selected as the majority subclasses, and the remaining two were designated as minority subclasses. Similarly, we maintained a ratio of 24:1 between the normal and abnormal classes, resulting in a ratio of 1,240:20:1 for normal:majority subclass:minority subclass.[1]

To evaluate the different methods, we computed the area under the reciever operating characteristic curve (AUROC). For each experiment, we used the average AUROC over five trials. Figure 3 shows the model's training over time. For each type of class imbalance (excluding interclass), we calculated the mean AUROC for each baseline (oversampling, weighted loss, vanilla SMOTE, no method) and CLE-SMOTE over all four permutations. For interclass imbalance, we only had one permutation and thus did not need to calculate an average.

4.4. Results

Across all five types of class imbalances, we find that CLE-SMOTE strongly outperforms the four competing baseline approaches. CLE-SMOTE was able to achieve .85 AUC for each imbalance experiment, converging between 10 and 30 epochs. Much of CLE-SMOTE's outperformance is acquired within the first 10 epochs, after which the increase in its performance is roughly matched by the baselines. Of note, we observe that while the "no-method",

1. Due to the presence of five subconditions (subclasses) within the abnormal class, we ran all permutations of this experiment. However, lung lesion was excluded from being a majority class due to insufficient data. Thus, we ran four permutations of the experiment.

Method	Binary AUROC (Variance)	Multiclass AUROC (Variance)
No Method	0.824 (0.00145)	0.673 (0.000414)
Oversampling Minority Class	0.815 (0.00411)	0.727 (0.00125)
Vanilla SMOTE	0.829 (0.00230)	0.691 (0.00415)
CLE-SMOTE	0.876 (0.000220)	0.753 (0.00108)
No Class Imbalance	0.903 (0.00150)	0.8332 (0.0124)

Table 3: AUROC results from baseline and CLE-SMOTE experiments. Multiclass experiments were performed with 3 classes.

.

"oversampled", and "weighted" baselines all converge on a similar AUC for each experiment (between .6 and .65), SMOTE consistently outperforms them, converging to a .80 AUC. Thus, CLE-SMOTE is able to outperform simple SMOTE by .05 AUC, indicating that the contrastive learning and triplet loss additions to the SMOTE framework allow the model to better learn representations of minority classes.

5. Conclusion

Class imbalance presents a barrier to the use of machine learning in healthcare tasks, including chest X-ray classification. While there exist many methods to alleviate interclass imbalance, including SMOTE, the issue of intraclass imbalance remains largely unaddressed. In this work, we propose CLE-SMOTE, a method that caps the loss of noisy SMOTE-generated examples based on their distance from the original dataset, thus reducing their influence. Our experiments on the CheXpert and ChestX-ray14 datasets demonstrate the effectiveness of CLE-SMOTE against baselines.

We hope that by releasing our trained network, as well as all training and statistical code, we can aid future work in the field. CLE-SMOTE presents a step towards developing robust training models that can effectively learn from datasets with both interclass and intraclass imbalance. Future research could include testing CLE-SMOTE on different classification tasks and datasets, or even applying CLE-SMOTE to generative tasks. Additionally, further research could be done into utilizing image embeddings to weight the influence of different examples.

Acknowledgments

We thank a bunch of people.

References

Gustavo Batista, Ronaldo Prati, and Maria-Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations*, 6: 20–29, 06 2004. doi: 10.1145/1007730.1007735.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June

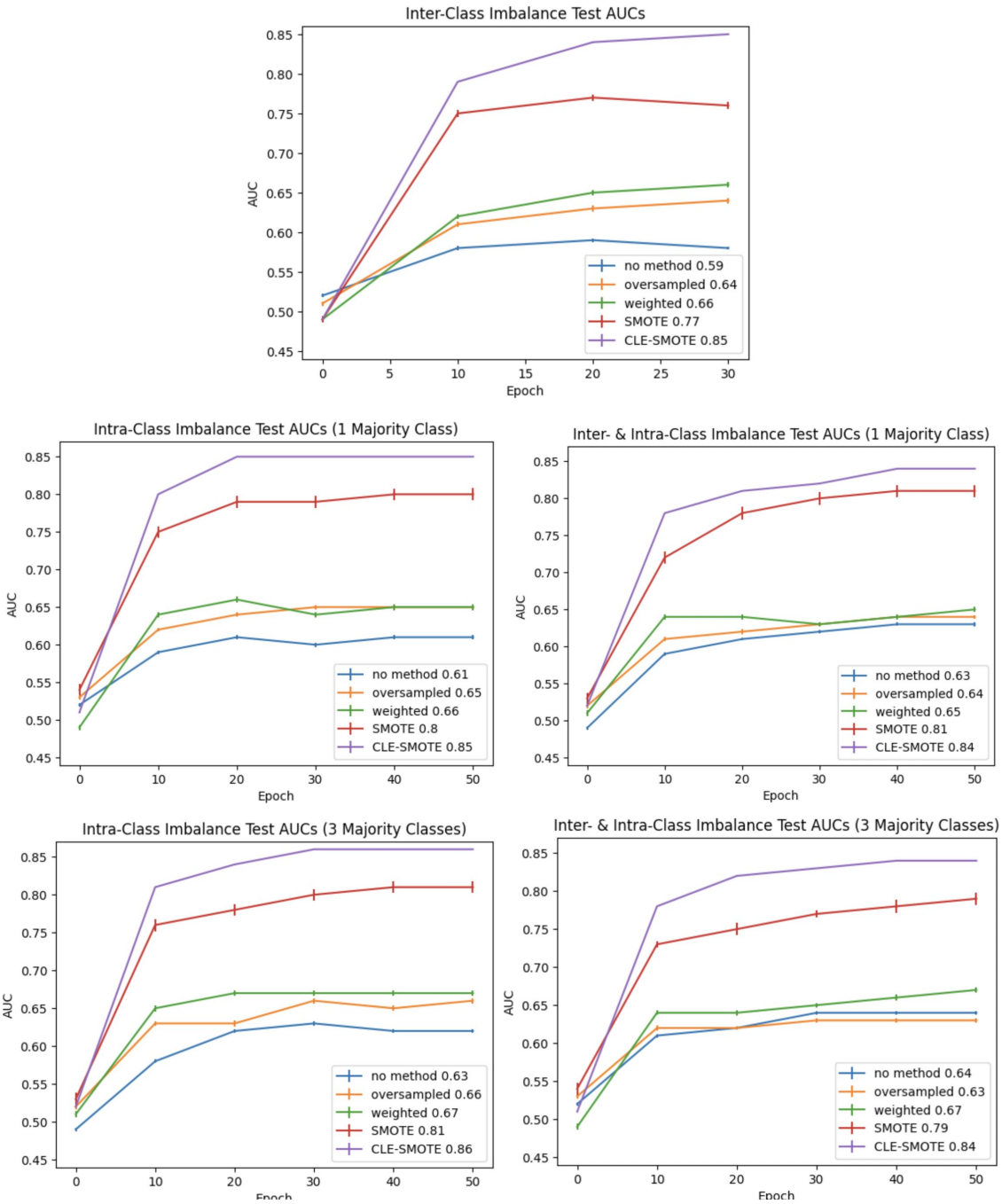

Figure 3: AUROCs and Variances over the first 50 epochs of training. The top graph shows the results of the interclass imbalance experiment. The second row shows the results of the intraclass and interclass & intraclass imbalance for the cases with a single majority class, while the third row shows results for the cases with multiple majority classes. CLE-SMOTE quickly outperforms all baselines in every type of class imbalance.

2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019. URL https://api.semanticscholar.org/CorpusID:58014111.

Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597, Jul. 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://ojs.aaai.org/index.php/AAAI/article/view/3834.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, Mar 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0192-5. URL https://doi.org/10.1186/s40537-019-0192-5.

Michael Kawooya. Training for rural radiology and imaging in sub-saharan africa: Addressing the mismatch between services and population. *Journal of clinical imaging science*, 2:37, 06 2012. doi: 10.4103/2156-7514.97747.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Der-Chiang Li, Chiao-Wen Liu, and Susan C. Hu. A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine*, 40(5):509–518, 2010. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2010.03.005. URL https://www.sciencedirect.com/science/article/pii/S0010482510000405.

Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020.

Yassine Marrakchi, Osama Makansi, and Thomas Brox. Fighting class imbalance with contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. URL https://api.semanticscholar.org/CorpusID:236513238.

Zaid Nabulsi, Andrew Sellergren, Shahar Jamshy, Charles Lau, Edward Santos, Atilla P. Kiraly, Wenxing Ye, Jie Yang, Rory Pilgrim, Sahar Kazemzadeh, and et al. Deep learning for distinguishing normal versus abnormal chest radiographs and generalization to two unseen diseases tuberculosis and covid-19. *Scientific Reports*, 11(1), Sep 2021. doi: 10.1038/s41598-021-93967-2.

Yasuo Nakajima, Kei Yamada, Keiko Imamura, and Kazuko Kobayashi. Radiologist supply and workload: international comparison–working group of japanese college of radiology. *Radiation medicine*, 26:455–65, 11 2008. doi: 10.1007/s11604-008-0259-2.

United Nations Scientific Committee on the Effects of Atomic Radiation. *Sources and Effects of Ionizing Radiation, United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2008 Report, Volume II*. United Nations, 2011. URL https://www.un-ilibrary.org/content/books/9789210544825.

Mayuri S Shelke, Prashant R Deshmukh, and Vijaya K Shandilya. A review on imbalanced data handling using undersampling and oversampling technique. *Int. J. Recent Trends Eng. Res*, 3(4):444–449, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. doi: 10.1109/cvpr.2017.369. URL http://dx.doi.org/10.1109/CVPR.2017.369.

Appendix A. Ablation Study

Before arriving at our current version of CLE-SMOTE, we conducted an ablation study on the CIFAR-10 dataset (Krizhevsky, 2009) to investigate alternative methods for capping the loss of SMOTE-generated examples. We tried the below formulations:

- **Constant Cap**: We employ a constant cap for all SMOTE-generated examples on the loss. Empirically, we determined the optimal cap to be 5 and 1 for binary and multiclass formulations, respectively.
- **Class Avg Cosine Distance Cap**: We utilize the same capping function as CLE-SMOTE (Equation 1), but we do not pretrain the network.
- **Batch Avg Cosine Distance Cap**: The same as Class Avg Cosine Distance Cap, but in lieu of the average class embedding of the entire dataset, we compute the average class embedding of the current batch.

Method (with SMOTE)	Binary AUROC (Variance)	Multiclass AUROC (Variance)
Vanilla SMOTE	0.829 (0.00230)	0.691 (0.00415)
Constant Cap	0.858 (0.00138)	0.727 (0.000580)
Class Avg Cosine Distance Cap	0.856 (0.00102)	0.744 (0.000493)
Batch Avg Cosine Distance Cap	0.868 (0.000354)	0.712 (0.00145)
Euclidean Distance Cap	0.822 (0.000923)	0.664 (0.000679)
CLE-SMOTE v1	0.835 (0.0126)	0.523 (0.00549)
CLE-SMOTE	0.876 (0.000220)	0.753 (0.00108)

Table 4: CLE-SMOTE's training as compared to baselines. AUROC results from our ablation study. Multiclass experiments were performed with 3 classes.

- **Euclidean Distance Cap**: The same as Class Avg Cosine Distance Cap, but in lieu of cosine distance, we use euclidean distance..
- **CLE-SMOTE v1**: In the initial version of CLE-SMOTE, we excluded SMOTE-generated examples from being used as anchors during the pretraining stage.

Results for our ablation study are shown in Figure 4 and Table 4.

We compared our final version of CLE-SMOTE to baselines on the CIFAR-10 dataset. We also use another baseline, No Class Imbalance, where we apply no methods and train the network on a balanced dataset. This baseline is not comparable to others that are trained on imbalanced datasets because of the inherent differences in the training set, but is useful as it gives a theoretical upper bound on performance of methods trained on imbalanced datasets. Results are shown in Figure 5 and Table 3. CLE-SMOTE outperforms all baselines and approaches the performance of an equivalent network trained on a balanced dataset.

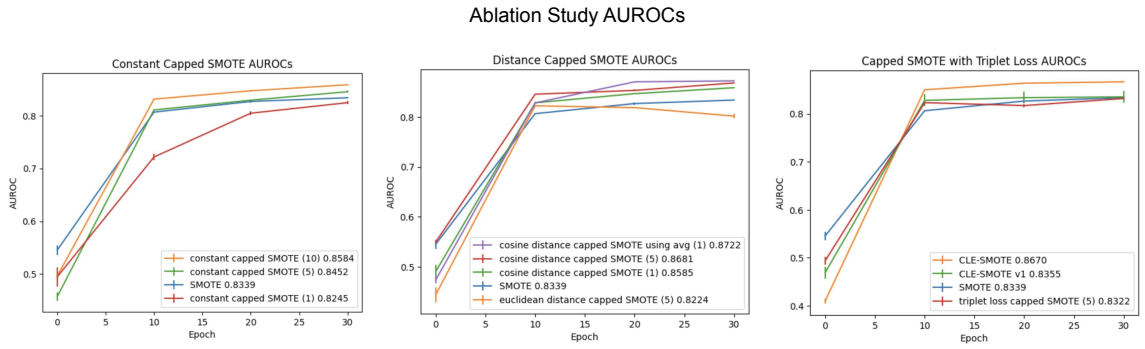

Figure 4: AUROCs and variances during the first 30 epochs of training for binary classification during the ablation study. The left graph shows the results of using a constant cap with a varying β. The center graph shows the results of both cosine and euclidean distance caps using a batch average. The right graph compares CLE-SMOTE and CLE-SMOTE v1 with a capping mechanism that employs triplet loss.

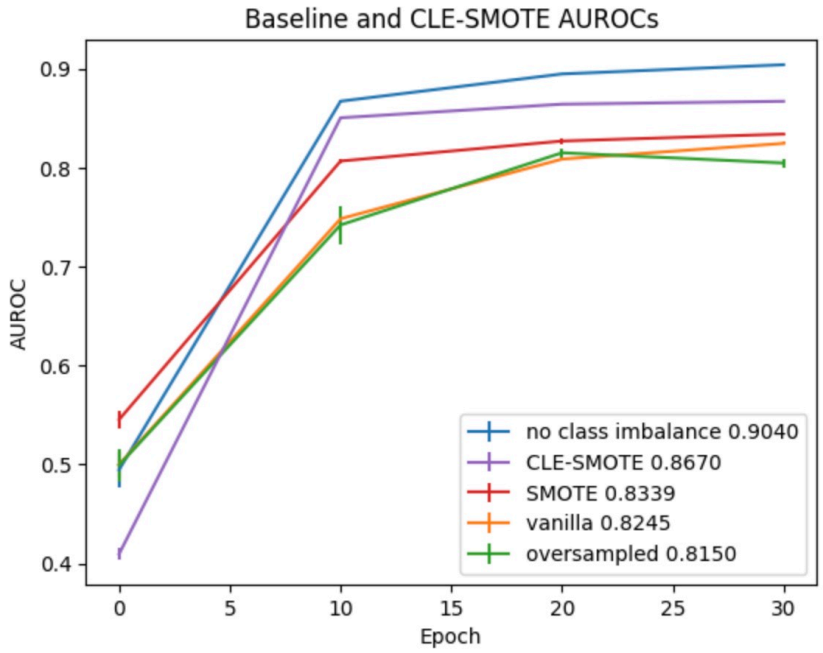

Figure 5: AUROCs and variances during the first 30 epochs of training for binary classification.