Handling Imbalanced Medical Dataset with Continuous Class Features using Improved Contrastive Learning

Jungwoo Bae
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
baejungwoo@skku.edu

Jitae Shin†
Department of Electrical and Computer Engineering
Sungkyunkwan University
Suwon, Republic of Korea
jtshin@skku.edu

***Abstract*—Medical datasets often face data imbalance due to challenges in collecting abnormal data, hindering reliable classification model development. Additionally, while severity-based categorical classes derived from continuous features aid patient understanding, unclear class definitions can lead models to overlook their continuous nature. Typically, these datasets are trained using contrastive learning methods, which focus solely on class-level distinctions without considering the underlying continuous information. In this paper, we propose an Improved Contrastive Learning method (ICL) to effectively learn from medical datasets that represent continuous nature as classes. Our approach incorporates curriculum learning in a two-phase learning framework. In Phase 1, original contrastive learning is applied. In Phase 2, we improve the learning process by sampling proxy lists based on class distribution parameters to address data imbalance and by updating class distance ratios to capture the continuous features between each classes. Our method outperforms existing approaches on the APTOS dataset. Furthermore, low-dimensional manifold visualizations of the learned representations reveal that disease features are progressively distributed according to class severity.**

***Index Terms*—Contrastive learning, Imbalanced datasets, Medical Image Processing**

I. Introduction

In recent years, there has been a significant surge in research aiming to utilize artificial intelligence (AI) to replace human roles across various domains. The medical field, in particular, has accelerated active exploration of AI applications [1], which hold promise for enhancing healthcare services such as remote diagnosis and self-examination, thereby improving medical welfare. However, AI research in medicine is intrinsically linked to human life, making the development of robust models especially critical. Despite this necessity, the data imbalance problem [2] prevalent in medical datasets poses challenges to training robust models. Specifically, datasets often contain abundant data on healthy individuals but lack sufficient data on patients with certain diseases. This imbalance complicates the training of models capable of accurate disease diagnosis. From this perspective, research to address these limitations is being prompted.

A common task in medical diagnostics involves predicting the severity of a disease. Datasets for such tasks frequently suffer from data imbalance, exacerbated by the difficulty of collecting data as disease severity increases [3]. This results in a long-tailed data distribution [4] where higher severity classes are underrepresented. Moreover, traditional classification models tend to assign each class as an independent category, disregarding the inherent continuity of disease severity. Representing disease severity as discrete classes can impair classification performance because it fails to capture the gradual progression inherent in medical conditions. While numerical representation of severity is possible, patients often find it challenging to perceive exact severity levels, necessitating the categorization into discrete classes to express patient conditions incrementally. Consequently, addressing these challenges requires novel approaches that incorporate the continuity of disease progression.

Given datasets that provide only class labels without explicit ranges, models typically employ multi-class classification methods, treating each class independently. This approach neglects the continuous features of the classes, such as the sequential representation of class labels reflecting specific ranges of severity. As a result, this can potentially lead to suboptimal learning outcomes. The exclusion of these continuous characteristics of class labels from the training process can hinder the model's ability to generalize and accurately reflect real-world disease progression.

To overcome these challenges, this paper proposes a deep learning model that leverages contrastive learning and proxy list methods to incorporate the continuous characteristics of classes in datasets with inherent continuity (e.g., disease severity) while simultaneously addressing the data imbalance problem. By integrating these advanced techniques into the learning process and implementing strategies to mitigate data imbalance, the proposed model aims to enhance classification performance. We validate the effectiveness of our approach through a series of experiments on real-world medical datasets, demonstrating its potential to significantly improve robustness, accuracy, and generalization in critical medical diagnosis tasks.

II. Background

A. Contrastive Learning

Contrastive learning [5] has emerged as a powerful method in classification tasks, especially in computer vision. Traditional deep learning often relies on backpropagation of cross-entropy loss from one-hot labels, which can be insufficient for capturing essential features in high-dimensional image data. In contrast, contrastive learning focuses on mapping images into low-dimensional representations through a feature extractor, enabling more effective analysis of sample relationships and improving classification performance by emphasizing salient features.

Supervised contrastive learning [6] refines this concept by operating within a batch to push apart representations of different classes and pull together augmented instances of the same class. This encourages cohesive class representations and robust feature invariance that enhance generalization. Subsequently, a classifier leverages these low-dimensional representations, improving performance by relying on rich, discriminative features discovered during the contrastive learning phase. This aligns with the manifold hypothesis [7], suggesting that meaningful features reside on a low-dimensional manifold within high-dimensional input space.

B. Handling data imbalance problems

However, many real-world scenarios involve data imbalance, where certain classes are underrepresented. To address this, resampling methods like SMOTE [8] and ADASYN [9] create synthetic minority samples, while under-sampling reduces majority classes. Cost-sensitive learning [10] and focal loss [11] assign higher costs to minority misclassifications, mitigating bias. Ensemble methods [12] combine multiple models to enhance minority class performance. Recent advances also incorporate data augmentation with GANs [13] and diffusion models [14], generating synthetic samples of minority classes.

These solutions collectively improve robustness and fairness, ensuring that AI models maintain reliable performance even when confronted with imbalanced datasets common in industrial and medical domains.

III. Methodology

In this section, we describe the model architecture employed to leverage contrastive learning on a long-tailed imbalance dataset that have continuous characteristics among classes referred to as Improved Contrastive Learning(ICL). By utilizing curriculum learning to divide the training process into two distinct phases, we first learn robust representations and then address issues arising from data imbalance. By addressing data imbalance problems it enables the model to better capture class-specific features by conducting contrastive learning that considers class distributions within the representations. Consequently, during inference, we enhance classification robustness by incorporating the probability of the representation's class distribution with the classifier's output. The overall architecture is illustrated in Fig. 1.

A. Curriculum Learning

Curriculum learning [15] is a strategy that structures the training process into sequential phases based on specific objectives. We adopt this method to split the model training into two phases. In the first phase, we employ traditional supervised contrastive learning to learn initial representations. In the second phase, we continue training to mitigate the class imbalance problem by integrating class distance and a proxy list into the representation learning.

During Phase 1, we apply appropriate augmentations to the dataset and train the model using supervised contrastive learning. This involves pulling together representations of the same class and pushing apart those of different classes within a batch, described by the contrastive learning formula shown in Eq. (1).

$$L_{\text{contra}} = -\sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \tag{1}$$

where $\mathbf{z}_i$ and $\mathbf{z}_p$ are normalized representations, $P(i)$ is the set of positives for i, $A(i)$ is the set of all positives and negatives, and τ is a temperature parameter.

As a result, Phase 1 enhances the generalization of representations through augmentation and increases cohesion among representations of the same class. However, this method has limitations in capturing class-specific features, especially in the presence of data imbalance. Minor classes may not be adequately trained due to fewer instances compared to other classes within a batch, leading to less cohesive and distinguishable representations for these classes. To address these issues, we introduce a proxy list and a class distance update method in Phase 2, aiming to reflect the continuous characteristics of classes in the representations.

B. Weighted Contrastive Learning with Class Distance Ratio

To learn the continuous characteristics of classes in Phase 2, we propose a class distance update method. Two main challenges arise when attempting to reflect continuous class features: first, the actual feature differences between classes are unknown because classes represent categorical divisions of continuous task characteristics, which cannot be identified from the dataset. Second, there are inherent variances among instances within the same class leading to uncertainty.

To overcome these challenges, we propose a contrastive learning approach that employs a class distance ratio as a weighting factor. The class distance ratio represents the proportional distance from one class to another. During contrastive learning in Phase 2, we assign weights based on this ratio when pushing representations of other classes, encouraging the representations to be linearly distributed according to class unknown real continuous label values.

The class distance ratio is defined as a vector of length C-1, where C is the number of classes. When comparing instances, we sum the ratios within the index gap of the classes involved and use this sum as the weight. This definition allows us to

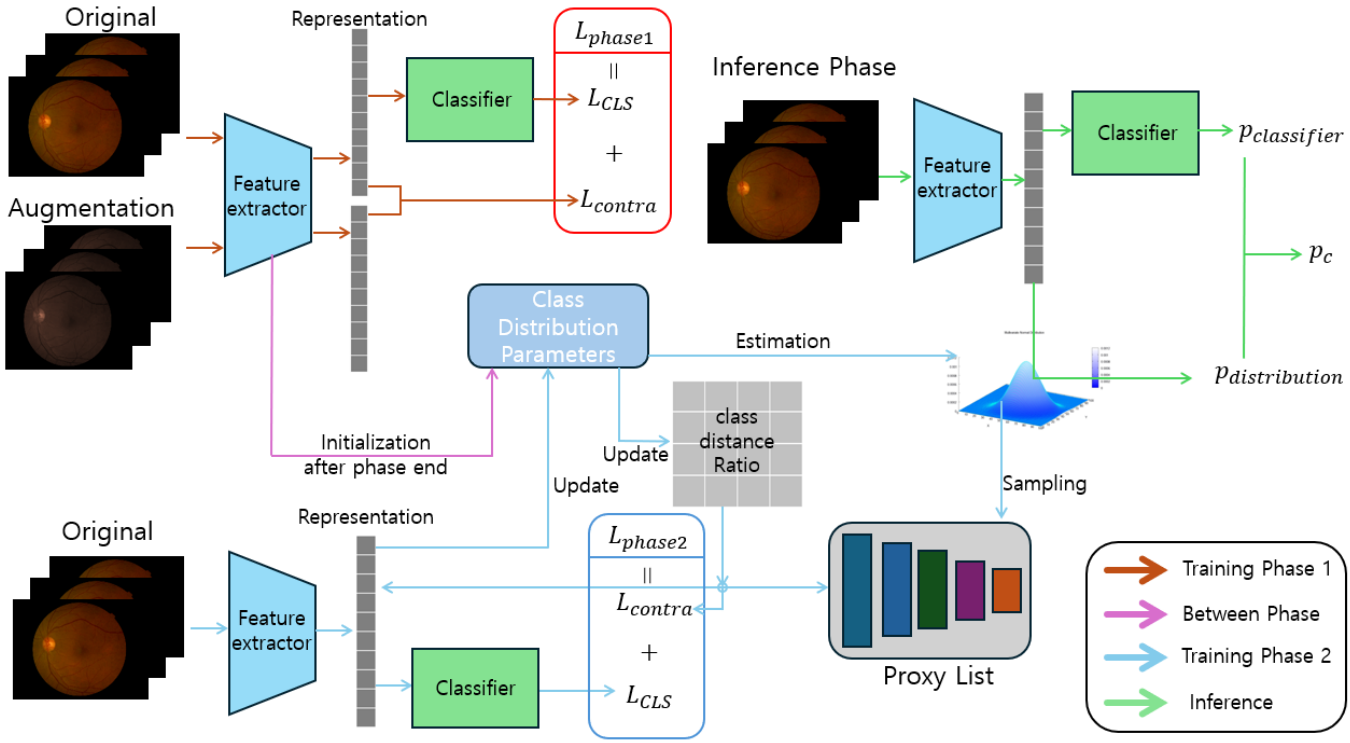

Fig. 1. Overall Architecture

assign sequential weights between instances with significant inter-class differences, enhancing training stability.

To initialize the class distance ratio vector, we utilize the feature extractor learned in Phase 1 to extract representations from the training dataset and define the distribution parameters for each class. These parameters consist of the mean and variance, with dimensions $\mu \in \mathbb{R}^h$ and $\sigma \in \mathbb{R}^h$ or $\mathbf{\Sigma} \in \mathbb{R}^{h \times h}$, where h is the hidden dimension size of the representation vector.

The class distribution parameters mean and variance are updated each epoch during Phase 2 to reflect changes in the representation distribution. The update equations are shown in Eq. (2) and (3).

$$\mu_c^{(t)} = \alpha \mu_c^{(t-1)} + (1-\alpha)\mu_c^{(batch)} \tag{2}$$

$$\sigma_c^{(t)} = \alpha \sigma_c^{(t-1)} + (1-\alpha)\sigma_c^{(batch)} \tag{3}$$

where α is the Exponential Moving Average (EMA) coefficient, and $\mu_c^{(batch)}$, $\sigma_c^{(batch)}$ are the mean and variance computed from the current batch for class c.

After updating the class distribution parameters, we update the class distance ratio using the updated means. This process enables contrastive learning to apply the features from the continuous characteristics of classes.

C. Addressing Class Imbalance with Proxy List

As previously mentioned, contrastive learning does not perform robustly on imbalanced datasets. To mitigate this, we incorporate additional samples from each class into contrastive learning during Phase 2, promoting more stable training. Unlike Phase 1, where various augmented images were used, Phase 2 utilizes representations generated referred to as proxies [16] from the class distributions.

These proxies are generated by estimating a Gaussian distribution using the class distribution parameters and sampling a specific number of representations. The number of proxies generated per class is determined based on the number of instances per class in the training dataset. The calculation for the number of proxies N_c^p for class c is shown in Eq. (4).

$$N_{c_p} = \left\lfloor \frac{b \times C}{N_c \sum_{i=0}^{C} \frac{1}{N_i}} \right\rfloor \tag{4}$$

where b is the batch size, C is the total number of classes, and N_c is the number of instances in class c.

By generating proxies in this manner, we consider the existing imbalance in the training set. During training, when an actual instance is processed, we perform contrastive learning with other instances in the batch and the proxies, weighted by the class distance ratio. This is the end of Phase 2.

D. Inference using both Representation and Classifier

Traditional classification methods typically compute the probability for each class through a classifier using the learned representations. However, this approach can degrade performance due to compression of representation features in the classifier's fully connected layer. To mitigate this issue, we enhance inference by leveraging the class distributions learned

during training to supplement feature information, akin to a residual connection.

During inference, we obtain class probabilities using the class distributions based on the extracted representations. We then perform soft voting [17] by combining these probabilities with those obtained from the classifier's output:

$$p_c(z) = \lambda p_{classifier,c}(\mathbf{z}) + (1-\lambda) p_{distribution,c}(\mathbf{z}) \quad (5)$$

where $p_c(\mathbf{z})$ is the classification probability for class c, $p_{classifier,c}(\mathbf{z})$ is the probability from the classifier, $p_{distribution,c}$ is the probability from the class distribution, λ is a weighting hyper-parameter and $\mathbf{z}$ is the representation vector. This method results in a more robust classification model by effectively combining both information of features in a different view.

IV. Results

A. Datasets and Evaluation Metrics

We evaluate our method on public medical image datasets that have data imbalance problems. APTOS 2019 blindness detection (2019) datasets [18] included 3662 retinal fundus images with five categories that defined the severity of diabetic retinopathy. We split 80%, 10%, 10% for train, valid, and test datasets. We use 5 metrics to evaluate our methods. Accuracy (ACC) and Area Under the ROC Curve (AUC) for classification, and we also use Macro-F1, balanced accuracy (B-ACC) and Matthew's correlation coefficient (MCC) to check model works well in imbalanced datasets. Results are presented in Table 1.

TABLE I
The Performance of different data imbalance handling methods

Method	$ACC(\uparrow)$	$AUC(\uparrow)$	$Macro-F1(\uparrow)$	$B-ACC(\uparrow)$	$MCC(\uparrow)$
Original	80.93	74.47	55.74	54.45	71.14
Under-sampling	79.29	78.78	63.14	62.89	69.08
Over-sampling	82.56	80.23	66.21	65.14	73.64
Contrastive learning	74.66	65.69	34.34	39.06	61.79
ICL(Ours)	**85.01**	**81.23**	**69.87**	**66.66**	**77.33**

B. Implementation Details

We use ResNet-18 as the backbone with the pre-trained model on the ImageNet dataset. We use random resized crop, color jitter, gaussian blur, flip for augmentation at phase 1. We trained ICL for 100 epochs with a batch size 128 and learning rate 1×10^{-5}. We set 20 epochs for phase 1 and the other for phase 2. We used Nvidia Tesla V100 for all the experiments.

C. Discussions

Fig. 2 presents a 3D Principle component analysis (PCA) visualization of the representation inferred from test datasets. The representation appears to be linear across classes also in test datasets. This visualization demonstrates that the feature extractor learns the features of classes present in a low-dimensional manifold. Fig. 3 also quantitatively shows that representations of the test dataset show greater distances with larger class differences.

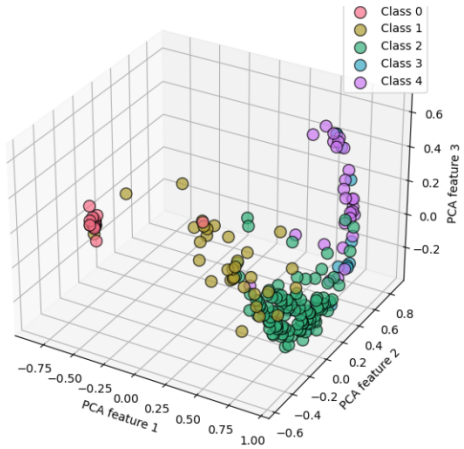

Fig. 2. 3D PCA visualization of representations

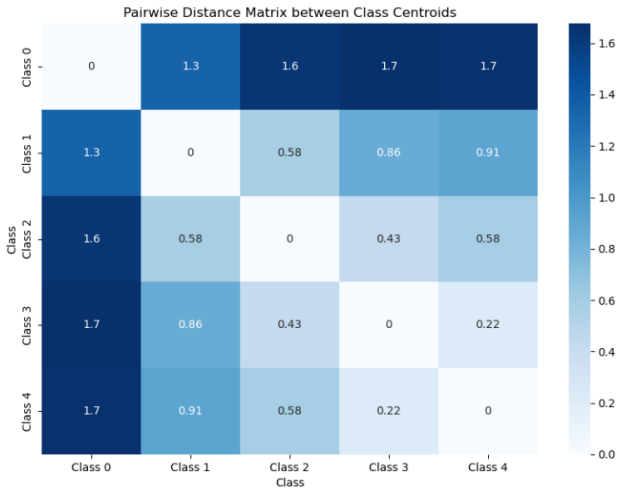

Fig. 3. Pairwise Distance Matrix between class representation centroids

V. Conclusion

In conclusion, we present new approach in contrastive learning to handle data imbalance problems that have continuous feature between classes. We evaluate our methods outperform previous methods, and show representation vector from feature extractor learns rich feature of class information. Future research would explore way to enhance the stability and robustness of model with advanced techniques for defining class distribution.

Acknowledgment

This work was supported by the BK21 FOUR Project.

References

[1] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, and A. K. Nandi, "Medical image segmentation using deep learning: A survey," *IET image processing*, vol. 16, no. 5, pp. 1243–1267, 2022.

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

[3] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[4] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 795–10 816, 2023.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[6] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[9] H. He, Y. Bai, E. A. Garcia, and 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.

[10] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, vol. 17, no. 1. Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[11] T. Lin, "Focal loss for dense object detection," *arXiv preprint arXiv:1708.02002*, 2017.

[12] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou, "A novel ensemble method for classifying imbalanced data," *Pattern Recognition*, vol. 48, no. 5, pp. 1623–1637, 2015.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[15] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[16] Y. Zhang, J. Chen, K. Wang, and F. Xie, "Ecl: Class-enhancement contrastive learning for long-tailed skin lesion classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 244–254.

[17] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, 2021.

[18] Karthik, Maggie, and S. Dane, "Aptos 2019 blindness detection," https://kaggle.com/competitions/aptos2019-blindness-detection, 2019, kaggle.