

# Proposal of self and semi-supervised learning for imbalanced classification of coronary heart disease tabular data

## Propuesta de aprendizaje auto-semi supervisado para la clasificación desequilibrada de datos tabulares de enfermedades coronarias

Danny Xie-Li<sup>1</sup>, Manfred González-Hernández<sup>2</sup>

---

Xie-Li, D; González-Hernández, M. Proposal of self and semi-supervised learning for imbalanced classification of coronary heart disease tabular data. *Tecnología en Marcha*. Vol. 37, special issue. August, 2024. IEEE International Conference on Bioinspired Processing. Pag. 38-43.


 <https://doi.org/10.18845/tm.v37i7.7295>


1 Instituto Tecnológico de Costa Rica. Costa Rica.

 [dxie@ic-itcr.ac.cr](mailto:dxie@ic-itcr.ac.cr)

 <https://orcid.org/0000-0003-1878-9460>

2 Universidad de Costa Rica. Costa Rica.

 [manfred.gonzalezhernandez@ucr.ac.cr](mailto:manfred.gonzalezhernandez@ucr.ac.cr)

 <https://orcid.org/0000-0002-5408-7901>

## Keywords

Self-supervised learning; semi-supervised learning; data augmentation; contrastive learning; imbalanced; medical datasets.

## Abstract

Triple Mixup is an augmentation policy in the hidden latent space we introduced in the Contrastive Mixup Self-Semi Supervised learning framework, to address the imbalanced data problem, for Cardiovascular Heart Diseases tabular dataset. Medical tabular datasets are known to present challenges as high imbalanced class, limited annotated quality samples due to the domain nature. Recent literature in Self and Semi supervised learning, has shown tremendous progress in learning useful representations, and leveraging unlabeled dataset and labeled dataset to train a learning model. Most existing methods are not feasible for tabular data due to the data augmentation scheme. In addition, the high imbalanced problem can show lower performance on machine learning algorithms. For this work, we propose the triple data augmentation method in hidden space to attack the unbalanced challenge in self-supervised and semi-supervised learning, from the possible applications of Contrastive Mixup, thus we will study the influence of it.

## Palabras clave

Aprendizaje Autosupervisado, Aprendizaje Semisupervisado, Aumentación de datos, Aprendizaje por Contraste, Desbalance de datos, datos médicos.

## Resumen

Triple Mixup es una política de aumento en el espacio latente oculto que introdujimos en el marco de aprendizaje autosupervisado de Mixup contrastivo, para abordar el problema de datos desequilibrados, para el conjunto de datos tabulares de enfermedades cardíacas cardiovasculares. Se sabe que los conjuntos de datos tabulares médicos presentan desafíos como muestras de calidad anotada limitada y de clase altamente desequilibrada debido a la naturaleza del dominio. La literatura reciente sobre el aprendizaje autosupervisado y semi supervisado ha mostrado un enorme progreso en el aprendizaje de representaciones útiles y en el aprovechamiento de conjuntos de datos no etiquetados y conjuntos de datos etiquetados para entrenar un modelo de aprendizaje. La mayoría de los métodos existentes no son factibles para datos tabulares debido al esquema de aumento de datos. Además, el problema de alto desequilibrio puede mostrar un rendimiento más bajo en los algoritmos de aprendizaje automático. Para este trabajo, proponemos el método de aumentación de datos triple en el espacio oculto para atacar el desafío desequilibrado en el aprendizaje autosupervisado y semi supervisado, desde las posibles aplicaciones de Contrastive Mixup, por ende estudiaremos la influencia de este .

## Introducción

Diversity on the training data is a key piece in the process of generalization, when training a supervised machine learning model. However, this can become a limitation due to the number and quality of the ground truth or related to how well diverse is the data. This problem has been attacked by data augmentation techniques in computer vision approaches [1]– [5] and, recent work has extended to tabular data approaches [6], [7]. However, tabular data often contain heterogeneous features that represent a mixture of continuous, categorical and ordinal values [8], and there is not an inherent positional information.

Unfortunately, obtaining labeled data is often infeasible in the healthcare domain, as annotation requires domain expert and manual labor. In addition, concerned with a particularly low representation of classes such as rare diseases. However, is often a wealth of unlabeled data available, but annotated data is only available for a small group. In order to take advantage of the unlabeled data, semi-supervised learning leverages the use of labeled and unlabeled data on training. Existing semi-supervised learning (SSL) algorithms from image and text domains are not effective for tabular data, because they heavily rely on spatial or semantic structure [9].

Some Tabular Health datasets are highly imbalanced, are often one or some minority classes, and most of the cases, the "minority" are more important than the "major" classes. As a result, traditional machine learning methods tend to produce results overwhelmed by the majority of classes, decreasing the model prediction accuracy. However, SSL has a limited assumption that the number of samples in different classes are balanced, and show lower performance for imbalanced class distribution datasets [10]. To augment data, the authors [11] explore the use of MixUp in the hidden space as augmentation for tabular data, and to consider class imbalanced [12] triple mixup to augment the data to generate more minority examples, but is only considered in the early stage for continuous variable.

Self-supervised learning frameworks have proven to be effective to learn representations from unlabeled data [13]. It is capable of adopting pseudo-labels based on the attributes learned for several downstream prediction tasks [11]. In addition, many self-supervised methods are based on contrastive representation learning. The authors in [14] for visual representation, extend the use of augmentation to generate "similar" samples, and normalized representations based on contrastive cross-entropy loss, to minimize the distance in the latent space of "positive" pairs and maximizes the distance of "negative" pairs. Recent approaches extend the use of self-supervised on tabular domain, authors in [9], proposed to recover the mask vector, in addition to the original sample with a novel corrupted sample generation for feature representation [11].

As indicated by [15], Cardiovascular Heart Diseases (CVDs) remain as one of the death causes with the highest mortality rates in the world. An estimated 17.9 million lives taken per year as mentioned by the World Health Organization. Although, several approaches focused on machine learning algorithms for CVDs have been proposed in the early years. This is the case for Chronic Kidney Disease where [16] suggests the implementation of Random Forest [17], a Supervised Machine Learning (SML) model to predict the occurrence of the disease. [18] proposed SML models to deal with the Coronary Heart Disease using the standard data set [19]. They opted to duplicate rows and get equal quantities for the number of rows they had per each class to manage the unbalanced data in the data set. Random Forest, Decision Tree and K-nearest neighbors were used in [18] to predict the existence of the deceased. In our proposed method we deal with the unbalanced data set [19] -collected by The Framingham Township Heart Institute offering a 10-year data set on coronary heart disease-via Triplet Mix up and a semi-supervised machine learning model.

## Preliminaries

To present our proposed method, we extend the work of [11] and [12] to formally introduce self-supervised, contrastive loss and semi-supervised loss. Given a dataset with  $N$  examples, there is a small subset defined as  $D_L = (x_i, y_i)_{i=1}^{N_L}$  for each example with the corresponding label, and  $D_U = (x_i, y_i)_{i=1}^{N_U}$  defined as the unlabeled dataset. Consider the  $x$  as the input features (consisting of numerical and categorical features) where  $y_i \in \{0,1\}$ ,  $1$  indicates that the person is prone to suffer CHD and  $0$  indicates the person is less likely to suffer CHD. In downstream tasks, we use supervised learning to find an optimized function  $f(X)$  that  $f(X) \rightarrow Y$  to minimize given the loss function  $l$ .

### Self-Supervised Learning: Contrastive Loss

Inspired by the recent contrastive learning framework [20] using metric learning methods, to learn representations. Given a batch of  $N$  samples is augmented using an augmentation function  $Aug(.)$  to create a multi-viewed batch with  $2N$  pairs,  $\{\tilde{x}_i, \tilde{y}_i\}_{i=1, \dots, 2N}$  where  $\tilde{x}_{2k}$  and  $\tilde{x}_{2k-1}$  are two random augmentations of the same sample  $x_k$  for  $k=1, \dots, N$ . The samples are fed to an encoder  $e: x \rightarrow z$ , which takes a sample  $x \in X$  to obtain a latent representation  $z = e(x)$ . For the pretext task defined, the model is trained jointly to minimize the self supervised contrastive loss function  $l$ .

$$\min_{e, h} \mathbb{E}_{(x, \tilde{y}) \sim P(X, \tilde{Y})} [l(\tilde{y}, h(e(x)))] \quad (1)$$

Where  $h$  maps  $z$  to an embedding space  $h: z \rightarrow v$ . Within a multi viewed batch  $i \in I = \{1, \dots, 2N\}$  the self supervised contrastive loss is defined as

$$l = \sum_{i \in I} -\log \left[ \frac{\exp(\text{sim}(v_i, v_j(i))/\tau)}{\sum_{n \in I \setminus \{i\}} \exp(\text{sim}(v_i, v_n)/\tau)} \right] \quad (2)$$

where  $\text{sim}(\dots) \in \mathbb{R}^+$  is a similarity function (e.g., dot product or cosine similarity),  $\tau \in \mathbb{R}^+$  is a scalar of temperature parameter,  $i$  is the anchor,  $A(i)$  is the positive(s) and  $I(i)$  are the negatives. The positive and negative samples refer to samples that are semantically similar and dissimilar, respectively. Intuitively, the objective of this function is to bring the positives and the anchor closer in the embedding space  $v$  and opposite for the negative samples.

### Semi-Supervised Learning: Loss

Given two disjoint datasets as  $D_L$  (labeled dataset),  $D_U$  (unlabeled dataset), the model as  $f$  is optimized by the conjunction of the supervised and unsupervised loss function defined as

$$\min_{f} \mathbb{E}_{(x, y) \sim P(X, Y)} [l(y, f(x))] + \beta \mathbb{E}_{(x, y_{ps}) \sim P(X, Y_{ps})} [l_u(y_{ps}, f(x))] \quad (3)$$

First term is estimated over a small labeled subset, and the unsupervised loss over the unlabeled subset. For this framework, the  $l_u$  is defined to support the supervised objective on pseudo-labelling.

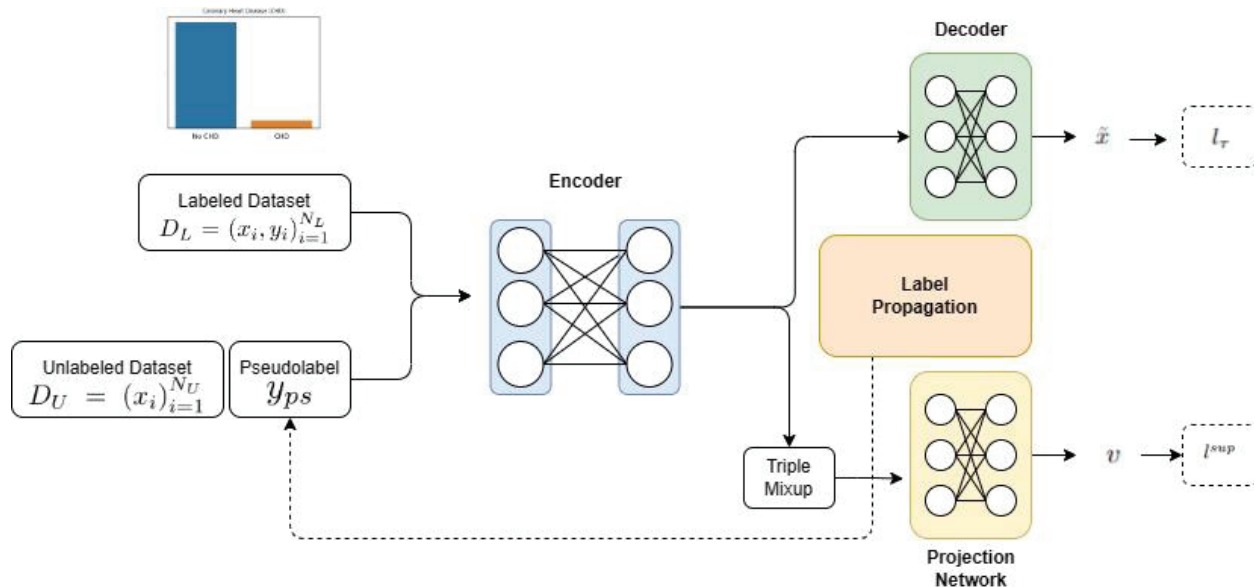
### Methodology

Triplet Mixup Data Augmentation was proposed by [12], where they used this data augmentation technique to generate minor examples to fit tabular domains:

$$\tilde{x} = \lambda_i x_i + \lambda_j x_j + (1 - \lambda_i - \lambda_j) x_k \quad (4)$$

where they used all the data from the minority (vulnerable) class because their goal was to alleviate the data imbalance problem when using continuous data. In our case, the context is different since we do have continuous and categorical features. On the other hand, we propose Triplet Mixup to interpolate between data belonging from the same class to create positive samples in the hidden space, instead of doing so in the input space as [12]. As mentioned by [11], low probable samples may be produced by Mixup in the input space due to the multi-modality of the data and the categorical features. More specifically, given an encoder  $E$ , that

contains  $f_T$  layers ( $T$  = total of layers), produces abstract representations of the input samples in the intermediate layers that are then interpolated with equation 4 to ensure high probable samples.



**Figure 1.** Contrastive Triple Mix Up in hidden space Framework.

Extending the definition of 1 for the interpolated intermediate layer samples generated by the encoder  $e$  that is composed of  $T$  layers  $f^t (t \in 1, \dots, T)$ . The Triple Mix up create the interpolation in the hidden layers as:

$$h^t_{ijk} = \lambda_i h^t_i + \lambda_j h^t_j + (1 - \lambda_i - \lambda_j) h^t_k \quad (5)$$

Where  $\lambda$  is a scalar from a uniform distribution  $U(0, \alpha)$  with  $\alpha \in [0, 0.5]$ .

From the equation (2) to maximize the distance for feature classes belonging to different classes and minimize the distance for same feature classes, the loss term is as described in [11].

## Future Work.

The priority of this work is to study the influence of the proposed triple mix-up on the hidden space as an augmentation technique to attack the tabular imbalanced dataset, using the framework by the authors in [11]; given the actual contrastive learning loss used [20]; benefits from larger batch sizes and longer training. Also, we want to investigate the impact of applying data augmentation on medical dataset in an early stage [12] mixed with Triplet Mixup in the inner stages of the model. The main intention is to generate more data given a single  $X$  keeping a good generalization in the classification of the Coronary Heart disease problem.

## Referencias

- [1] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 778–10 787.
- [2] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 386–397, 2020.

- [3] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020.
- [4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.
- [5] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 11, pp. 3212–3232, 2019.
- [6] D. Snow, "DeltaPy: A Framework for Tabular Data Augmentation in Python," SSRN Electronic Journal, pp. 1–3, 2020.
- [7] B. Sathianarayanan, Y. C. Singh Samant, P. S. Conjeevaram Guruprasad, V. B. Hariharan, and N. D. Manickam, "Feature-based augmentation and classification for tabular data," CAAI Transactions on Intelligence Technology, vol. 7, no. 3, pp. 481–491, 2022.
- [8] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, "Saint: Improved neural networks for tabular data via row attention and contrastive pre-training," 6 2021. [Online]. Available: <http://arxiv.org/abs/2106.01342>
- [9] J. Yoon, Y. Zhang, J. Jordon, and M. van der Schaar, "Vime: Extending the success of self-and semi-supervised learning to tabular domain," Advances in Neural Information Processing Systems, vol. 33, pp. 11 033–11 043, 2020.
- [10] M. Hyun, J. Jeong, and N. Kwak, "Class-imbalanced semi-supervised learning," 2 2020. [Online]. Available: <http://arxiv.org/abs/2002.06815>
- [11] S. Darabi, S. Fazeli, A. Pazoki, S. Sankararaman, and M. Sarrafzadeh, "Contrastive mixup: Self- and semi-supervised learning for tabular domain," 2021. [Online]. Available: <http://arxiv.org/abs/2108.12296>
- [12] X. Li, L. Khan, M. Zamani, S. Wickramasuriya, K. W. Hamlen, and B. Thuraisingham, "Mcom: A semi-supervised method for imbalanced tabular security data" in IFIP Annual Conference on Data and Applications Security and Privacy. Springer, 2022, pp. 48–67.
- [13] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," 10 2020. [Online]. Available: <http://arxiv.org/abs/2011.00362>
- [14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2 2020. [Online]. Available: <http://arxiv.org/abs/2002.05709>
- [15] P. M. Tripathi, A. Kumar, R. Komaragiri, and M. Kumar, A Review on Computational Methods for Denoising and Detecting ECG Signals to Detect Cardiovascular Diseases. Springer Netherlands, 2022, vol. 29, no. 3. [Online]. Available: <https://doi.org/10.1007/s11831-021-09642-2>
- [16] A. Subas, E. Alickovic, and J. Kevric, "Diagnosis of chronic kidney disease by using random forest," IFMBE Proceedings, vol. 62, no. 3, pp. 589–594, 2017.
- [17] W. Deng, Z. Huang, J. Zhang, and J. Xu, "A Data Mining Based System for Transaction Fraud Detection," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering, ICCECE 2021, pp. 542–545, 2021.
- [18] D. Krishnani, A. Kumari, A. Dewangan, A. Singh, and N. S. Naik, "Prediction of coronary heart disease using supervised machine learning algorithms," IEEE Region 10 Annual International Conference, Proceedings/TENCON, vol. 2019-Octob, pp. 367–372, 2019.
- [19] H. Yang, "Coronary heart disease historical data," 2022. [Online]. Available: <https://dx.doi.org/10.21227/eapx-t883>
- [20] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," CoRR, vol. abs/2002.05709, 2020. [Online]. Available: <https://arxiv.org/abs/2002.05709>
- [21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2019. [Online]. Available: <https://arxiv.org/abs/1911.05722>