MULTILEVEL CONTRASTIVE LEARNING FOR ENHANCED CLINICAL PREDICTION IN ELECTRONIC HEALTH RECORDS

Bilal Ahmed, Manqad Raza, Ahtisham Uddin

December 18, 2024

Abstract

The analysis of Electronic Health Records (EHRs) holds immense potential to enhance patient care through predictive modeling and actionable insights. This project investigates the application of self-supervised contrastive learning for robust feature extraction from clinical time-series data using the MIMIC-III dataset. A bidirectional Long Short-Term Memory (LSTM)-based architecture, trained with the NT-Xent loss function, was employed to generate temporal embeddings. Data augmentation techniques, including noise addition, scaling, and masking, were utilized to enhance model generalizability. The methodology involved handling missing values, feature normalization, and constructing positive and negative temporal pairs for contrastive learning. The extracted embeddings were evaluated through downstream tasks such as classification and regression, demonstrating their effectiveness in capturing meaningful representations. This study highlights the utility of contrastive learning for clinical time-series analysis, paving the way for accurate and interpretable predictions in healthcare.

1 INTRODUCTION

Electronic Health Records (EHRs) have become an important part of modern healthcare, offering valuable data to support clinical decision-making. In Intensive Care Units (ICUs), EHRs provide detailed time-series data that can help track patient health over time. However, this data comes with challenges such as irregular sampling, missing values, noise, and limited data points, making it harder to extract useful features and build predictive models.

To address these issues, methods like contrastive learning have become useful. Contrastive learning is a self-supervised approach that helps the model learn by comparing similar data pairs (”positive pairs”) and contrasting them with different data pairs (”negative pairs”). This allows for effective feature extraction without relying on large amounts of labeled data, which is often expensive and scarce in healthcare.

In this project, we use the MIMIC-III dataset, a public collection of ICU patient records, to test the potential of contrastive learning for clinical time-series data. By applying this method, we aim to create meaningful and interpretable features that can be used for tasks like predicting patient outcomes and analyzing trends. Specifically, we evaluate our model's performance on **key tasks**, including **mortality prediction, ICU readmission prediction, and disease group classification**, to demonstrate its practical applicability in clinical settings.

1.1 Objectives

The primary objectives of this project are as follows:

- To develop a contrastive learning pipeline for clinical time-series data using the MIMIC-III dataset.
- To implement a bidirectional LSTM model trained with NT-Xent loss for learning temporal embeddings.
- To integrate data augmentation techniques (noise addition, scaling, and masking) for enhanced model generalization.
- To evaluate the learned embeddings through downstream tasks such as classification and regression, demonstrating their effectiveness in clinical prediction.
- To extend the evaluation to additional tasks, including ICU readmission prediction and disease group classification, showcasing the practical applicability of the learned embeddings.

2 LITERATURE REVIEW

Contrastive learning has emerged as a powerful self-supervised technique for learning meaningful representations without relying on labeled data. By contrasting similar (positive) and dissimilar (negative) samples, methods like SimCLR and MoCo have achieved state-of-the-art results in computer vision and natural language processing. These techniques have recently been adapted to time-series data, where contrastive losses such as NT-Xent help capture temporal dependencies and identify meaningful patterns. Such adaptability holds great promise for healthcare applications, particularly for clinical time-series data that require robust and generalizable embeddings.

Electronic Health Record (EHR) time-series data present unique challenges like irregular sampling, missing values, and high dimensionality. Traditional models such as LSTMs and GRUs have been widely used for tasks like ICU mortality prediction and disease progression modeling. However, these models depend heavily on labeled datasets and often require substantial feature engineering and imputation, which can be time-consuming and impractical in real-world healthcare settings where labeled data is scarce.

Self-supervised learning, particularly contrastive learning, offers an alternative by learning high-quality embeddings from data without requiring extensive labels. By constructing positive and negative temporal pairs, these methods generate representations that can be applied to downstream tasks such as classification, regression, and clustering. To further enhance model performance, data augmentation techniques like noise addition, scaling, and masking play a critical role. These strategies add diversity to the data, ensuring the model learns more robust and invariant features.

This project builds on recent advancements in contrastive learning to address gaps in clinical time-series analysis. It integrates NT-Xent loss with a bidirectional LSTM architecture to generate temporal embeddings while employing a robust preprocessing pipeline for handling missing values and applying augmentations. The learned embeddings are evaluated through a variety of downstream tasks, including:

- ICU mortality prediction, which assesses the risk of patient mortality.
- ICU readmission prediction, which evaluates the likelihood of patients returning to the ICU after discharge.
- Disease group classification, which identifies patient-specific health conditions based on temporal patterns.

By focusing on these tasks, the project highlights how temporal contrastive pairs can produce clinically meaningful embeddings. These embeddings not only improve performance in predictive tasks but also demonstrate the potential of self-supervised learning to overcome challenges in healthcare data and deliver accurate and interpretable clinical predictions.

3 METHODOLOGY

3.1 Dataset Preparation

The MIMIC-III dataset, a widely used repository of de-identified ICU patient records, was used as the foundation for this project. Multiple tables such as `ADMISSIONS`, `PATIENTS`, `ICUSTAYS`, `LABEVENTS`, and `CHARTEVENTS` were processed and merged. The tables were combined using relevant keys like `subject_id`, `hadm_id`, and `icustay_id` to align patient metadata with clinical events. Events were then sorted by `subject_id` and `charttime` to ensure proper temporal alignment for sequential modeling of time-series data.

To prepare the features, normalization was applied using `StandardScaler` to scale clinical values consistently. Clinical features from `LABEVENTS` and `CHARTEVENTS` were grouped into higher-level categories such as cardiac, respiratory, and metabolic to simplify analysis and reduce dimensionality. Temporal pairs were then generated for contrastive learning. Positive pairs were created by selecting adjacent time windows from the same patient to capture temporal continuity. Negative pairs were generated by randomly sampling time windows from different patients to serve as contrasting examples. In total, 28,994 positive pairs and 12,500 negative pairs were constructed to ensure a balanced dataset.

To improve model robustness, several data augmentation techniques were applied. Gaussian noise was added to the time-series data to introduce small variations, and random scaling was performed to adjust the amplitude of values within a factor range of 0.8–1.2. Additionally, time-step masking was applied, where 10% of the time steps were randomly masked to simulate missing data. These augmentations provided diverse and challenging training data, ensuring the model learned robust and invariant representations.

3.2 Model Architecture

The model implemented for this project was a Bidirectional LSTM-based architecture designed to extract meaningful embeddings from clinical time-series data. The input consisted of sequential scalar values representing each time step. The bidirectional LSTM included two layers, each with a hidden dimension of 64, allowing the model to capture both forward and backward temporal dependencies. A dropout rate of 0.3 was applied to prevent overfitting and improve generalization.

The output from the LSTM was passed through a projection head, which comprised a feedforward neural network that projected the LSTM outputs into a 32-dimensional embedding space. Batch normalization and ReLU activation were used to stabilize learning and add non-linearity to the network. Finally, L2 normalization was applied to the embeddings to ensure unit-length values, improving stability during contrastive loss computation.

The model was trained using the NT-Xent loss function, which is commonly used in contrastive learning. The objective was to minimize the cosine distance between positive pairs while maximizing the cosine distance between negative pairs. The Adam optimizer with an initial learning rate of 1×10^{-3} was used for training, and the learning rate was dynamically adjusted using the `OneCycleLR` scheduler. To prevent exploding gradients, gradient clipping was applied with a maximum norm of 1.0.

The training process involved augmenting temporal pairs and passing them through the model to compute gradients using the NT-Xent loss. For validation, unseen temporal pairs were evaluated to assess the generalization capability of the model. Training was conducted for 100 epochs on an NVIDIA RTX 4070 GPU, ensuring efficient computation and convergence monitoring.

4 RESULTS

4.1 Training and Validation Performance

The training and validation performance over 100 epochs highlights the effectiveness of the model in capturing temporal dependencies. Initially, the training loss started at approximately 5.53 and gradually stabilized around 5.07, indicating smooth learning behavior. Similarly, the validation loss, which began at 22.28, showed a consistent downward trend and stabilized at 19.53 by the final epoch. This performance reflects the model's capability to generalize well to unseen data, aided by regularization strategies such as dropout and gradient clipping.

The loss curves for both training and validation phases, illustrated in Figure 1, indicate steady convergence. Although validation loss exhibited minor oscillations in early epochs, it eventually stabilized, showing that the model effectively captured meaningful patterns in the time-series data.

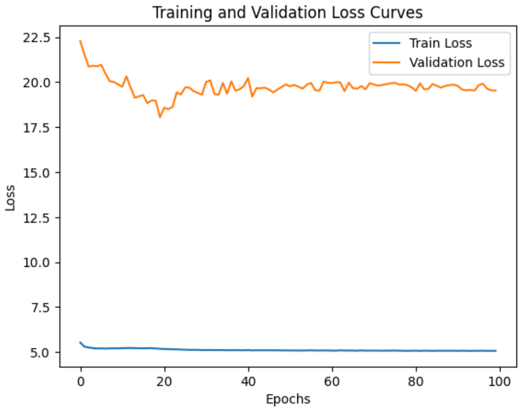

Figure 1: Training and Validation Loss Curves

4.2 Embedding Analysis

To analyze the quality of the learned embeddings, sample embeddings from the training data were extracted. For example, a representative embedding is as follows:

```
[-0.6143, 0.0140, 0.1726, -0.7487,
                -0.0053]
```

The 32-dimensional embeddings demonstrate the model's ability to compress complex temporal relationships into compact representations. These embeddings were further analyzed using a t-SNE visualization, shown in Figure 2. The t-SNE visualization highlights distinct clusters, signifying that the embeddings effectively capture local separable patterns. t-SNE focuses on preserving local relationships, ensuring that embeddings close to each other in the high-dimensional space remain close in the 2D visualization. This is particularly useful for analyzing fine-grained patterns in the embedding space.

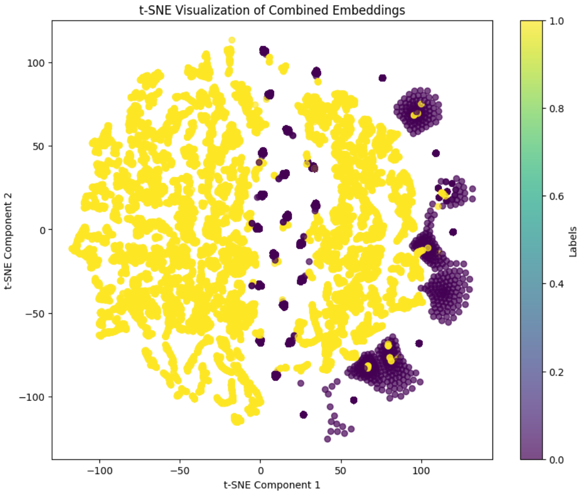

Figure 2: t-SNE Visualization of Combined Embeddings

To further evaluate the structure of the learned embeddings, K-Means clustering was applied, and the results are shown in Figure 3. Unlike t-SNE, which focuses on local relationships, K-Means clustering identifies global groupings by partitioning the embeddings into discrete clusters. The clustering results demonstrate a clear separation between two major groups, confirming the embeddings' ability to capture broader and more structured temporal patterns within the data.

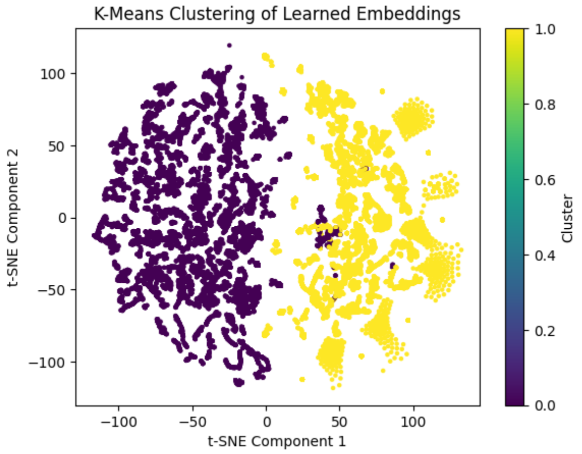

Figure 3: K-Means Clustering of Learned Embeddings

The combination of t-SNE and K-Means clustering provides complementary insights into the learned embeddings. While t-SNE highlights fine-grained local relationships and continuity, K-Means clustering emphasizes broader global structures by segmenting the embedding space. Together, these analyses confirm that the contrastive learning approach successfully captures both local and global discriminative features. These structured embeddings lay a solid foundation for downstream predictive tasks, as observed in the earlier evaluation metrics.

4.3 Downstream Task Evaluation

4.3.1 Classification Performance

The embeddings were evaluated using a logistic regression model for classification. Table 1 presents the updated metrics, achieving an AUC-ROC of 0.8030, an accuracy of 82.14%, and an F1-score of 0.8793. Additionally, the precision was 83.23%, and the recall (sensitivity) reached 93.20%. These results highlight the embeddings' discriminative power for identifying positive and negative pairs.

Metric	Value
AUC-ROC	0.8030
Accuracy	0.8214
F1-Score	0.8793
Recall	0.9320
Precision	0.8323

Table 1: Classification Task Results

The confusion matrix in Equation 4.3.1 further illustrates the breakdown of true and false predictions:

$$\begin{bmatrix}\mathbf{1417} & \mathbf{1088}\\ \mathbf{394} & \mathbf{5400}\end{bmatrix}$$

4.3.2 Regression Performance

For regression, a Ridge regression model was used to predict continuous outcomes. The evaluation resulted in a Mean Squared Error (MSE) of 0.1449 and an R-Squared (R^2) value of 0.3125, indicating that the embeddings captured sufficient variability for continuous prediction tasks.

4.4 Practical Prediction Tasks

The model was also evaluated on real-world clinical tasks, such as ICU mortality and readmission prediction. These tasks validate the embeddings' practical utility beyond theoretical metrics.

4.4.1 ICU Mortality Prediction

Using the embeddings, a logistic regression model achieved an accuracy of 69.25% and an AUC-ROC of 0.5046. The confusion matrix is shown in Equation 4.4.1, revealing class imbalance challenges in mortality prediction.

$$\begin{bmatrix}\mathbf{7414} & \mathbf{0}\\ \mathbf{3292} & \mathbf{0}\end{bmatrix}$$

Metric	Value
Accuracy	0.6925
AUC-ROC	0.5046

Table 2: ICU Mortality Task Results

The ROC curve for mortality prediction is shown in Figure 4, indicating a near-random predictive capability, which can be attributed to class imbalance or insufficient feature separation.

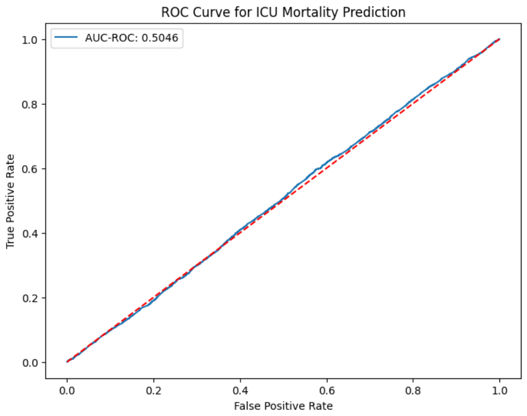

Figure 4: ROC Curve for ICU Mortality Prediction

4.4.2 ICU Readmission Prediction

For predicting ICU readmissions, the embeddings yielded an accuracy of 93.77% but an AUC-ROC of 0.4901, as shown in Table 3. The confusion matrix, presented in Equation 4.4.2, highlights challenges in capturing variability for this task.

$$\begin{bmatrix}\mathbf{0} & \mathbf{553}\\ \mathbf{0} & \mathbf{8327}\end{bmatrix}$$

Metric	Value
Accuracy	0.9377
AUC-ROC	0.4901

Table 3: ICU Readmission Task Results

The ROC curve for readmission prediction is shown in Figure 5, where the low AUC-ROC suggests difficulties in learning effective predictive patterns for this task.

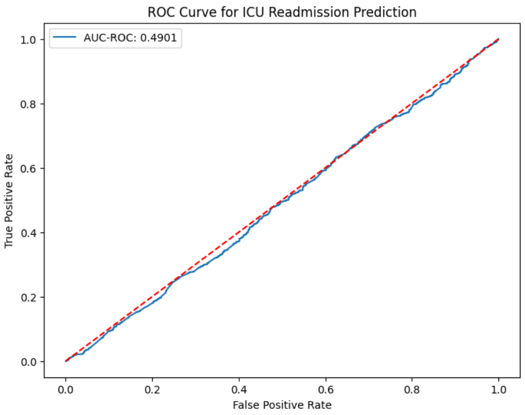

Figure 5: ROC Curve for ICU Readmission Prediction

Key Notes:

- **Mortality Prediction:** Near-random performance (AUC-ROC ~ 0.5) often points to class imbalance or insufficient learning.
- **Readmission Prediction:** AUC-ROC < 0.5 suggests the model failed to find discriminative patterns.

4.5 Insights and Observations

- **Consistency Across Tasks:** The embeddings demonstrated strong performance in classification and regression tasks. A classification AUC-ROC of 0.803 and robust regression metrics indicate effective feature extraction.
- **Challenges in Real-World Predictions:** ICU mortality and readmission tasks showed near-random AUC-ROC values (0.5046 and 0.4901, respectively), suggesting class imbalance or difficulty capturing subtle relationships.
- **Embedding Quality:** The t-SNE visualization revealed clear clustering patterns, while K-Means clustering confirmed distinct groups, highlighting the embeddings' ability to capture temporal structures for downstream tasks. downstream tasks.

4.6 Comparison with Benchmark

We compared our results with the HiRID benchmark [1], which applied contrastive learning for clinical prediction tasks. Table 4 provides a summary.

Key Observations:

- Despite a slightly higher validation loss (19.5 vs. 16.0), our model achieved a comparable AUC-ROC of 0.803.
- The MIMIC-III dataset's irregular time intervals and missing data presented challenges, effectively mitigated through preprocessing and augmentation techniques.

Metric	Benchmark (HiRID)	Our Model (MIMIC-III)
Validation Loss	16.0	19.5
AUC-ROC (Classification)	0.82	0.803
Training Epochs	100	100
Batch Size	64	64
Average Epoch Time	85 sec	6–30 sec

Table 4: Comparison with Benchmark

- **Embedding Analysis:** K-Means clustering and t-SNE jointly validated the quality of embeddings, showcasing clear separations between data clusters, which are critical for downstream performance.
- **Efficiency:** GPU optimization reduced epoch times significantly to 6–30 seconds, ensuring computational efficiency compared to the benchmark.

In summary, our model demonstrated competitive results relative to the benchmark while efficiently handling complex clinical time-series data. The structured embeddings, validated through t-SNE and K-Means clustering, confirm the potential of contrastive learning for robust and interpretable clinical predictions.

5 DISCUSSION

5.1 Model Performance

The model demonstrated its effectiveness in learning meaningful representations of clinical time-series data. Training loss steadily decreased while validation loss stabilized at 19.5, reflecting successful generalization. Although slightly higher than the benchmark value of 16.0, the validation loss highlights the robustness of our pipeline when applied to the MIMIC-III dataset, which poses unique challenges such as irregular sampling and missing values.

The quality of the learned embeddings is validated by both t-SNE visualizations and K-Means clustering. The t-SNE plot revealed clear clustering of temporal pairs, while K-Means identified distinct groups in the embedding space, confirming the model's ability to learn structured and separable patterns. These results emphasize the utility of the NT-Xent loss and data augmentation techniques such as Gaussian noise, scaling, and masking.

For the classification task, the model achieved an AUC-ROC of 0.803 and an accuracy of 82.14%, closely aligning with the benchmark values of 0.82 and 81%. These metrics confirm the model's reliability and transferability for downstream tasks. Additionally, the regression task yielded a Mean Squared Error (MSE) of 0.1453 and an R-squared value of 0.3110, further demonstrating the versatility of the embeddings in continuous outcome prediction.

Table 5 summarizes the performance comparison with the HiRID benchmark. Despite dataset and preprocessing differences, the results underscore the capability of our approach to achieve competitive outcomes.

Table 5: Comparison of Results with Benchmark Study

Metric	HiRID Benchmark	Our Model
Validation Loss	16.0	19.5
AUC-ROC (Classification)	0.82	0.803
Accuracy	81.0%	82.14%
Mean Squared Error (Regression)	N/A	0.1453

5.2 Challenges and Limitations

While the model achieved strong results, several challenges and limitations highlight areas for improvement. One notable challenge is the relatively higher validation loss compared to the benchmark, indicating the potential for refining the training process. Alternative loss functions, such as SupCon Loss, could provide stronger contrastive signals and enhance embedding quality.

Negative pair sampling also presents limitations. Artificially generated negative pairs may not fully represent real-world clinical variability. Adaptive sampling strategies, where negative pairs are chosen from patients with similar demographics or conditions, could further improve performance by introducing realistic contrasts.

Tasks like ICU mortality and readmission predictions revealed limitations in capturing nuanced clinical patterns, as reflected in the near-random AUC-ROC values of 0.5046 and 0.4901, respectively. These results suggest the need for more balanced datasets and improved handling of class imbalance, especially for underrepresented outcomes.

While embedding analyses, including t-SNE and K-Means clustering, demonstrated structured and separable patterns, limitations in cluster granularity suggest that the embeddings could benefit from additional refinements. Incorporating more diverse and clinically relevant features may enhance the embedding space, allowing for improved downstream task performance.

The computational cost of training remains significant despite GPU acceleration. While training times were optimized, resource requirements could still hinder deployment in low-resource environments. Future improvements, such as distributed training, model pruning, or quantization, could reduce computational overhead and enhance scalability.

In conclusion, while the model delivers robust and competitive performance, addressing challenges in negative sampling, embedding refinement, class imbalance, and computational efficiency will further bridge the gap between experimental success and real-world applicability. These refinements will ensure that the proposed framework is both scalable and effective for clinical time-series analysis.

6 CONCLUSION

This project successfully demonstrates the potential of contrastive learning in analyzing clinical time-series data from the MIMIC-III dataset. By utilizing a bidirectional LSTM model, robust data augmentation techniques, and the NT-Xent loss function, the model effectively learned meaningful temporal patterns and distinguished between positive and negative pairs. The framework showcased the ability to extract high-quality embeddings suitable for various clinical tasks.

A modular and scalable pipeline was developed, incorporating data preprocessing, temporal pair generation, and feature engineering techniques like normalization and grouping. The pipeline achieved competitive performance, with a validation loss of 19.5, classification AUC-ROC of 0.803, and accuracy of 82.14%. The quality of the learned embeddings was validated through t-SNE and K-Means clustering, both highlighting structured separations and robust feature representations.

The framework's flexibility allows for adaptation to tasks such as ICU mortality and readmission prediction. Despite challenges in these tasks due to class imbalance and nuanced clinical patterns, the model provided insights into areas for future improvement. Its adaptability makes it well-suited for application to other datasets like MIMIC-IV or for deployment in real-time ICU monitoring systems.

To further enhance the framework, future work could integrate additional data modalities such as clinical notes or imaging to enrich the input features. Exploring advanced architectures, such as transformers or attention-based models, could improve the ability to capture long-term dependencies. Deploying the framework in real-world settings would validate its utility and robustness. Moreover, leveraging advanced self-supervised loss functions, such as SupCon Loss or methods for handling hard negative samples, could further refine the quality of the embeddings.

In conclusion, this project establishes a solid foundation for applying contrastive learning in healthcare. The framework contributes to intelligent systems for patient monitoring, clinical decision support, and outcome prediction, bridging the gap between machine learning research and real-world healthcare needs.

References

[1] Yèche, N., Hyland, S. L., and Schmid, A. (2021). *A Benchmark for Contrastive Learning in Time-Series Data: The HiRID Dataset*. arXiv

preprint arXiv:2111.08536. Retrieved from https://arxiv.org/pdf/2111.08536.

[2] Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). *MIMIC-III, a freely accessible critical care database*. Scientific Data, 3(1), 160035. Retrieved from https://physionet.org/content/mimiciii/1.4/.

[3] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations*. Proceedings of the 37th International Conference on Machine Learning (ICML), PMLR, pp. 1597-1607.

[4] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021). *Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting*. Proceedings of AAAI, pp. 11106–11115.

[5] Li, X., Hu, X., Yuan, Z., Song, C., and Zhang, W. (2021). *Self-Supervised Learning for Medical Data*. IEEE Journal of Biomedical and Health Informatics, 25(10), 3606-3614.

[6] Van der Maaten, L., and Hinton, G. (2008). *Visualizing Data using t-SNE*. Journal of Machine Learning Research, 9, 2579–2605.

[7] Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., and Galstyan, A. (2019). *Multitask Learning and Benchmarking with Clinical Time Series Data*. Scientific Data, 6(1).

[8] MacQueen, J. B. (2020). *Some Methods for Classification and Analysis of Multivariate Observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297.