Patient Stratification with Temporal Self-Supervised Learning

Dimitrios Proios[1], Alban Bornet[1], Anthony Yazdani[1], Douglas Teodoro[1],
{dimitrios.proios, alban.bornet, anthony.yazdani, douglas.teodoro}@unige.ch

[1]Department of Radiology and Medical Informatics, Faculty of Medicine, University of Geneva, Geneva, Switzerland

Abstract

Patient stratification, the task of defining precise patient cohorts, can improve healthcare research and lead to personalized and effective treatments. Electronic Health Records (EHRs) contain temporal data on vital signs, lab results, clinical assessment, demographics, and administered medications recorded during hospital care. Self-supervised learning can exploit sequences of medical events composing clinical trajectories to represent patients in a high-dimensional latent space. This latent representation enables modeling patient characteristics for large-scale stratification analyses. Despite extensive research on machine learning methods for supervised learning problems, there are very few replicable multi-centric benchmarks focused on unsupervised learning problems. To address this gap, we introduce a reproducible pipeline aimed at evaluating self-supervised and unsupervised learning algorithms using open-source data from four publicly available Intensive Care Unit (ICU) repositories. First, we compare statistical and self-supervised models to generate temporal embeddings based on patient characteristics. Then, we evaluate unsupervised learning in a stratification task aimed at rediscovering hierarchical levels of annotated ICD codes, providing a structured approach for patient stratification research. Our experiments are fully replicable using the open-source software repository available at: https://github.com/ds4dh/AAAI_cohort_stratification.

Introduction

Identification of clinically relevant subtypes, a task known as patient stratification, promotes personalized medicine to support translational medicine research, improve clinical markers systems, and reduce healthcare costs (Teschendorff et al. 2006). Researchers have demonstrated that machine learning methods can enhance patient stratification by mining electronic health records (EHRs). For example, Fahad Shabbir et al. (Ahmed et al. 2020) developed a deep neural network approach to predict mortality in trauma patients admitted to the intensive care unit.

To create representation from heterogeneous information, researchers extensively employed statistical and deep learning methods for patient stratification. Santero et al., encoded patient's trajectories in chronological order to generate sequence-like representations, which were fed to a Word2vec model to automatically learn concept representation , while Choi et al., modeled sequences of ICU events with Graph Transformer on the same problem (Jaume-Santero et al. 2022),(Choi et al. 2020). Despite their success, these models cannot address the large number of codes in existing clinical terminologies, which are continuously evolving.

For that reason, researchers employed unsupervised machine learning creating patient representations enhancing existing ontologies. Landi et al., demonstrated the usage of embeddings to perform stratification for multiple diseases at a large scale, reusing the same Convolutional Autoencoder model (Landi et al. 2020). However, most efforts focused on experiments with private datasets from one clinical center. Furthermore most evaluations focused disease-specific stratification, which hinders the adoption of large-scale EHR-based stratification analysis.

While in the supervised formulation several research works provided open datasets and formed benchmarks for supervised tasks (Harutyunyan et al. 2019), (van de Water et al. 2024). Specifically, Harutyunyan et al., (Harutyunyan et al. 2019) proposed four supervised benchmarks for mortality, length-of-stay, acute decompensation, and phenotyping inference with MIMIC-III dataset machine learning application frameworks in the supervised setting (Johnson et al. 2016), (Goldberger et al. 2000). Van de Water et al., used the *ricu* R package, to formalize multicentric supervised tasks of clinical relevance (van de Water et al. 2024), (Bennett et al. 2023). Despite promising results in the supervised domain, there is a lack of unsupervised learning benchmarks in open datasets for patient stratification due to extrinsic label unavailability (Alexander 2023).

Researchers have mitigated the lack of labels for clustering evaluation; re-discovering existing knowledge as a way to validate new methods has served the biomedical domain (Teschendorff et al. 2006), (Kim et al. 2023), (Bradshaw et al. 2023), (Vrbik et al. 2015). Specifically, Teschendorff et al., (Teschendorff et al. 2006) identified existing major cancer classifiers and proposed new clinical markers for complex disease identification; Hyunkyung et al., (Kim et al. 2023) clustered type 2 diabetes loci, uncovering distinct mechanistic pathways as clinical markets. Bradshaw et al., (Bradshaw et al. 2023) integrated phenotype and protein interaction networks, providing hypotheses for undiagnosed

1. Pre-processing	2. Self-Supervised Models	3. Unsupervised Clustering
• Categorical Encoding • Feature dimension Imputation • Split train/test/stay • Grouping by stay	• Temporal dimension imputation • STAT • LSTM • GRU	• Stratification • Label naming assignment • Hierarchical clustering • t-SNE • k-Means

Figure 1: Overview of architecture pipeline.

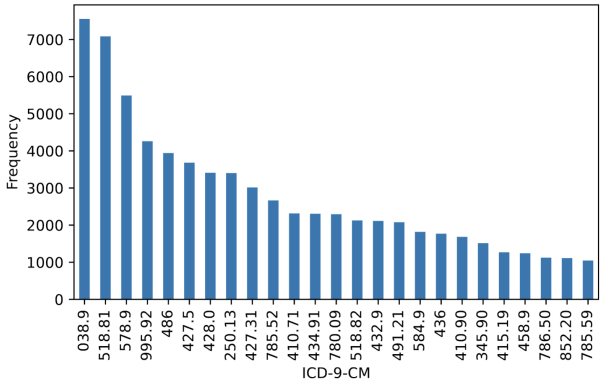

Figure 2: Frequency histogram of top 25 ICD-9-CM labels across datasets.

diseases, while I. Vrbik et al., (Vrbik et al. 2015) rediscovered breast cancer subtypes that matched existing phylogenetic profiles as true labels. These rediscoveries validate known curated terminologies, demonstrating their capacity to identify relevant subtypes.

In order to address the lack of unsupervised benchmarks for patient representation learning, we propose an evaluation framework using as external labels the International Classification of Diseases (ICD) terminology and the Clinical Classification Software (CCS) hierarchical taxonomies (WHO 2019). In summary, we propose an unsupervised stratification benchmark using standard multi-centric ICU data (Bennett et al. 2023), with the following contributions:

- Replicable unsupervised benchmark allowing for model evaluation for patient stratification using multicentric ICU data repositories;
- Imputation-free data representations for the architectures comprised of statistical methodologies as well as LSTM and GRU deep learning models;
- Interpretability analysis using distinct problem formulation allowing us to revise the capacity of each model across disease CCS and ICD disease ontologies.

Methods

Dataset description

We used data from four distinct ICU data repositories (Johnson et al. 2023; Sadeghi et al. 2024; Hyland et al. 2020; Pollard et al. 2018), which include multiple ICU units and encompass diseases from almost all ICD chapters, albeit with varying frequencies. The feature space consists of 114 features, from which 108 are hourly-sampled time series and 6 are static (i.e., not varying over time), as described in Table 1. These features include demographic characteristics, vital signs, mechanical support indicators, and clinical assessments, representing critical data from ICU monitors that continuously assess patients at risk of deterioration. For each ICU stay, distinct ICD-9-CM and ICD-10 codes are reported across three of the four datasets. Given the irregular sampling of some diseases, we focused on the top-25 ICD-10 labels, mapping them to ICD-9-CM, to ensure sufficient sample size and statistical power since their distribution is non-uniform, right-skewed and long tailed, as illustrated in Figure 2. Moreover, we applied hourly time granularity for the time series features across all datasets, in order to harmonize data collection rates.

Dataset	Feat.	Stays	Codes	Time unit
eICU	114	173,109	919	Hour
HiRID	100	33,905	-	Minute
MIMIC-IV	113	73,175	37,690	Hour
SiC	86	27,386	2,169	Minute

Table 1: Descriptive statistical summary across datasets.

Data preprocessing

We used the ricu R package [4], an open-source library that allows to apply temporal, value-based filtering and normalization. Furthermore, we performed preprocessing steps, including ordinal and one-hot encoding of categorical variables. We applied both feature and time imputation to account for missing values and align feature values across datasets (e.g., eICU contains boolean flags for mortality for all patients, whereas other datasets may contain only one of the two states). Last we applied feature normalization using the training sets using Robust Scaler, and grouping data by ICU stay (Pedregosa et al. 2011).

Temporal self-supervised modeling

We employed both statistical and deep learning baselines to generate temporal embeddings from ICU data in the derived dataset. We used one statistical (STAT) and two recurrent neural network (RNN)-based deep learning self-supervised methods to generate patient representations with the aforementioned dataset.

The two deep learning baselines were trained using autoregressive unidirectional LSTM and GRU models, based on the work of (Zang and Wang 2021) and (Harutyunyan et al. 2019). Three distinct LSTM-based models, one per

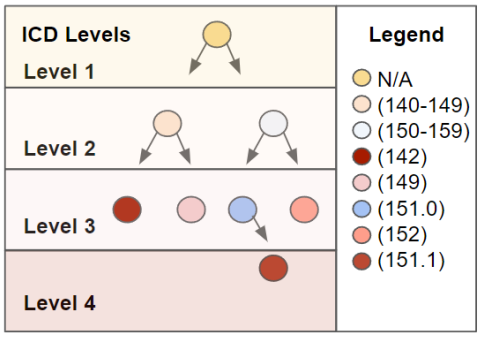

Figure 3: An example of an extrinsic label subset using the ICD-9-CM hierarchical structure.

dataset, were trained and optimized for the autoregressive task of inferring the feature values of the next time step. In contrast, the STAT baseline does not involve any learning process. For patient representations, we utilized the hidden states of the last timestep. In the case of STAT, we used concatenated statistical moments across time windows to create a statistical representation of each ICU stay, as described in (Proios et al. 2023).

For deep learning baselines, we used a single channel accepting a 108-dimensional feature vector. In addition, we experimented with distinct LSTM cells for each feature. Subsequently, we replaced the LSTM cells with gated recurrent unit (GRU) cells and repeated the experiments.

The generated embeddings were used as input for training unsupervised and semi-supervised clustering models to evaluate the capacity to identify patient cohorts. Our overall pipeline is summarized in Figure 1.

ICD hierarchical tree levels

The ICD is a hierarchical coding system used to classify diseases across the four levels L_i ($i \in \{1, 2, 3, 4\}$). Let $\mathcal{L}_i$ be the set of codes at Level L_i. Let $x_p \in R^d$ represent the embedding vector for patient p, where d is the dimensionality of the embedding space. These embeddings are generated by the self-supervised model (e.g., LSTM, GRU, STAT) and encapsulate the patient's clinical trajectory information. Each patient $p \in \mathcal{P}$ is associated with a true label $y_{p,i} \in \mathcal{L}_i$, where $y_{p,i}$ represents the patient's disease code at a specific level in the ICD hierarchy.

The mapping $f_i : \mathcal{L}_i \to \mathcal{L}_{i-1}$ ensures that each code $c_i \in \mathcal{L}_i$ has exactly one parent in $\mathcal{L}_{i-1}$. This hierarchical structure defines a rooted tree $L = (V, E)$:

$$V = \mathcal{L}_1 \cup \mathcal{L}_2 \cup \mathcal{L}_3 \cup \mathcal{L}_4, \tag{1}$$

$$E = \{(c_i, c_{i+1}) \mid c_{i+1} \in \mathcal{L}_{i+1},\ c_i = f_{i+1}(c_{i+1})\}. \tag{2}$$

Patient cohort stratification

In this task, we formulate patient stratification as an unsupervised problem. The ICD hierarchy is used to extrinsically evaluate a model's ability to align with $\mathcal{L}_i$. Formally, a clustering model (e.g. k-Means) maps each patient p to a cluster $K_j^{(i)}$, where $K^{(i)}$ represents the set of clusters produced for L_i. In this setting, we evaluate the derived clusters with respect to the j-th ICD code at the i-th level. In the ideal case, each individual cluster $K_j^{(i)}$ is comprised by a single code j, i.e., composed of all patient with identical $y_{p,i}$.

Hierarchy rediscovery

In this second problem formulation, we define an iterative clustering problem, where the goal is to approximate with a clustering model from broad to specific levels. Let

$$g^{(i+1)} : (\mathcal{P}, K^{(i)}) \to K^{(i+1)}, \tag{3}$$

where $K^{(i)}$ represents the hierarchical cluster assignments of $\mathcal{P}$ at level i, using the prior set of clusters $K^{(i)}$ with $K^0 = \emptyset$. The process can be described in the following steps:

1. **Initial clustering at level L_1:** Cluster all patients $\mathcal{P}$ into k clusters corresponding to the broadest ICD categories in L_1:
$$K^{(1)} = g^{(1)}(\mathcal{P}, \emptyset). \tag{4}$$
2. **Iterative refinement for subsequent levels:** For each subsequent level L_{i+1}, we refine each cluster from the previous level by clustering the subset of patients belonging to that cluster:
$$K^{(i+1)} = g^{(i+1)}(\mathcal{P}, K^{(i)}), \tag{5}$$

This hierarchical approach aims to mirror the ICD tree structure by partitioning data into increasingly granular subcategories.

Cluster Label Assignment

In this task, we assign a label to each cluster $K_j^{(i)} \in K^{(i)}$ using the true labels $y_{p,i}$ of the patients belonging to that cluster. Although labels $y_{p,i}$ exist for each patient p, they remain unseen during the clustering process, preserving the unsupervised nature of the task. Instead, labels are utilized post-clustering to evaluate the quality of the cluster assignments in a transductive classification setting. Let

$$\mathcal{P}_{K_j^{(i)}} = \{p \in K_j^{(i)}\} \tag{6}$$

denote the set of patients in cluster $K_j^{(i)}$. We define the cluster labeling function ℓ as follows:

$$\ell(K_j^{(i)}) = y_{p^*,i} \tag{7}$$

where $\ell(K_j^{(i)})$ assigns to cluster $K_j^{(i)}$ the most representative label $y_{p,i}$, that is, $y_{p^*,i}$, among the available labels in $\mathcal{P}_{K_j^{(i)}}$. We consider three strategies to determine p^*, using only the true labels of the training set:

1. Centroid-based. Assign the label of the patient whose embedding is closest to the cluster centroid:

$$\mu_j = \frac{1}{|\mathcal{P}_{K_j^{(i)}}|} \sum_{p \in \mathcal{P}_{K_j^{(i)}}} x_p, \tag{8}$$

where

$$p^* = \text{argmin}_{p \in \mathcal{P}_{K_j^{(i)}}} \|x_p - \mu_j\|_2 \tag{9}$$

2. Medoid-based. Assign the label of the patient whose embedding minimizes the total distance to all other embeddings in the cluster:

$$p^* = \text{argmin}_{p \in \mathcal{P}_{K_j^{(i)}}} \sum_{q \in \mathcal{P}_{K_j^{(i)}}} \|x_p - x_q\|_2. \tag{10}$$

3. Majority-vote. Assign the most frequent true label among patients in $K_j^{(i)}$. Formally,

$$p^* = \text{argmax}_{c \in \mathcal{L}_i} \sum_{p \in \mathcal{P}_{K_j^{(i)}}} \delta\left(y_{p,i} = c\right), \tag{11}$$

where c represents a candidate label from the set of possible labels and δ the indicator function summing the number of patients with a particular label defined as:

$$\delta\left(y_{p,i} = c\right) = \begin{cases} 1, & \text{if } y_{p,i} = c, \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

These strategies assign a single representative label $\ell(K_j^{(i)})$ to each cluster $K_j^{(i)}$, enabling a consistent comparison of the cluster assignments against the true labels $y_{p,i}$.

For evaluation, we employ V-Measure (Rosenberg and Hirschberg 2007), Adjusted Mutual Information (AMI) (Vinh, Epps, and Bailey 2010), and Accuracy. We optimize hyperparameters (including t-SNE parameters) using Optuna (Akiba et al. 2019) with 1,000 trials for each formulation of the problem.

Results

Temporal self-supervised learning results

Our results aim to enable comparison between distinct self-supervised and statistical methods—specifically STAT, LSTM-based, and GRU-based models in three problem formulations: clustering, label assignment, and hierarchical clustering.

Using our metrics, we evaluate the ability of each model to derive patient representations without relying on true labels during training. For the statistical method, we used a fixed-size concatenated vector of statistical moments across distinct time windows. This methodology was applied without any parameter adjustments, thus requiring no training.

For the second and third sets of models, we trained autoregressive deep learning baselines using RNN architectures with LSTM or GRU layers, each followed by a fully connected layer. These architectures were implemented in PyTorch and trained with a learning rate of 10^{-4} using the AdamW optimizer to minimize the Mean Squared Error (MSE) loss (Paszke et al. 2019), (Loshchilov and Hutter 2019).

We retained the hidden state from the last timestep as the patient representation embedding for both architectures. We observed a gradual decrease in validation loss across the three datasets, indicating that at least some information could be inferred for predicting the next timestep values in the autoregressive self-supervised setting.

Patient cohort cluster evaluation

Using the statistical and deep learning baselines we generated vector representation of patient trajectories. We trained and evaluated distinct models on three out of four datasets excluding HiRID due to the absence of diagnoses information. At the inference time, we retrieved statistical and neural embeddings from each model. In our experiments self-unsupervised training neural baselines outperform in most cases the statistical ones, demonstrating that machine learning is capable of learning the dataset's non-linearities more robustly.

Using the neural embeddings from the two machine learning baselines and the STAT patient temporal representations, we compared clustering algorithms based on feature vectors using k-Means with varying k values for each clustering task and applying t-SNE for dimensionality reduction. Finally, we evaluated the different clustering evaluation formulations (i.e., $L_1 - L_4$) with performance indicating the model was more capable of rediscovering higher levels of hierarchy. We summarize the results for V-measure in Figures 4, 5 and 6.

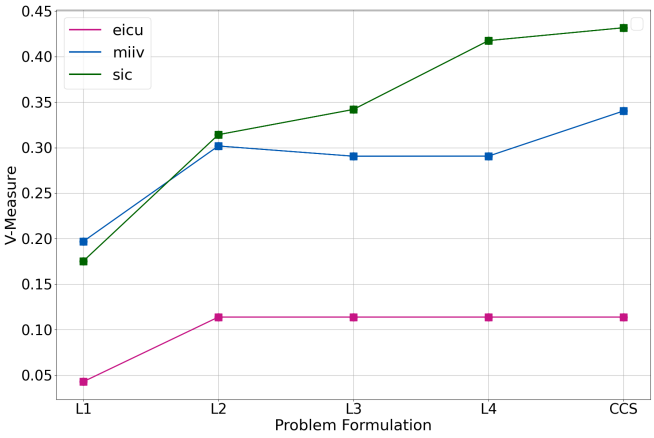

Figure 4: STAT clustering evaluation using V-Measure metric comparison across datasets and problem formulations.

We note that while both metrics managed to partially rediscover parts of ICD hierarchy LSTM models surpassed significantly the statistical baseline across all problem definitions. In all models we observe an upward trend as the task became more complex aligning proportionally with the number of clusters to be rediscovered, with the notable exception of eICU dataset for GRU-based architecture.

Finally regarding performance across datasets, we improved significantly for eICU dataset indicating that the high number of features was leveraged by the LSTM baseline more effectively. On the other hand, models perfomance on SiC dataset was worse using LSTM and GRU baselines in the unsupervised and last MIMIC-IV had consistent performance.

Hierarchical rediscovery evaluation

We averaged the accuracy scores for the clustering problems $L_k \rightarrow L_{k+1}$ for the embeddings derived by self-supervised

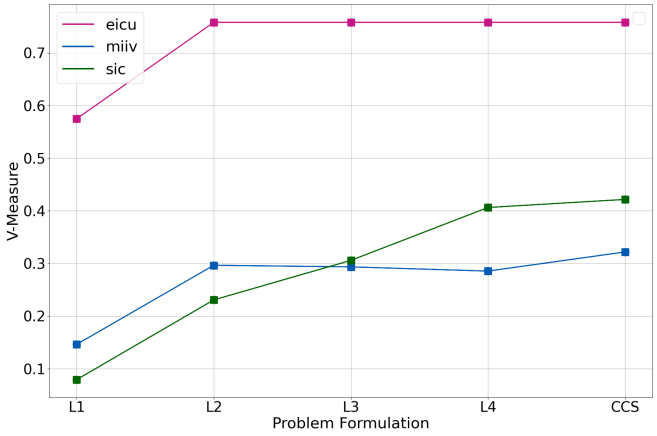

Figure 5: LSTM clustering evaluation using V-Measure metric comparison across datasets and problem formulations.

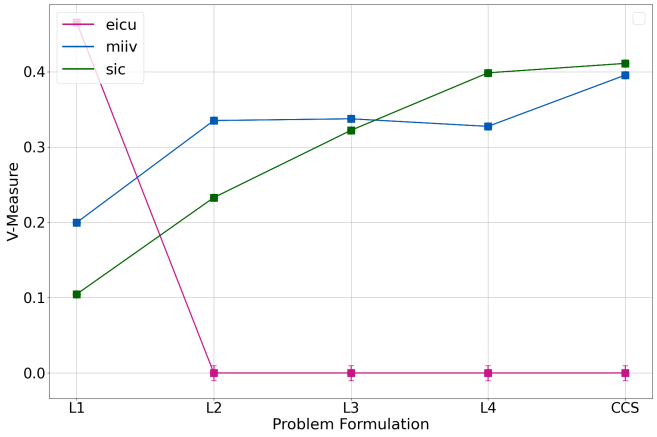

Figure 6: GRU clustering evaluation using V-Measure metric comparison across datasets and problem formulations.

methods, as illustrated in Figures 7, 8 and 9. We observe inconsistent trends, and missing points for the lower levels of ICD hierarchy which can be accounted to the smaller number of samples among categories with respect to our criteria for sufficiently large clusters comprised more than 10 samples.

Once relaxing this assumption to the minimal number to form a distinct cluster to a higher setting we can retain sufficiently large groups though not qualifying for comparison. For all the datasets we observe an initial trend to discover more effectively the higher levels of of our hierarchies while performance drops in $L_2 \rightarrow L_4$ and $L_3 \rightarrow L_4$, in the more granular ICD hierarchy level .

Cluster label assignment evaluation

In this problem formulation, we used our training set to assign labels to the each cluster. We use the the true label of the centroid, medoid cluster points in addition to the majority based to assign a cluster label to each dataset and assessed the models' embeddings with respect to the most representative sample of the cohort.

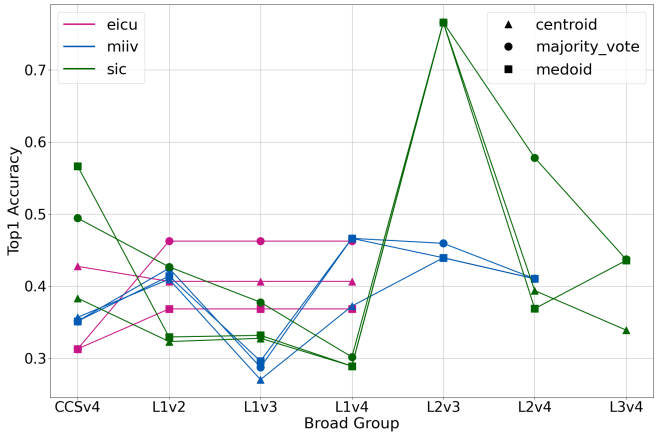

Figure 7: STAT Average accuracy per problem definition $L_n \rightarrow L_{n+1}$ in the clustering evaluation comparison across datasets and problem formulations.

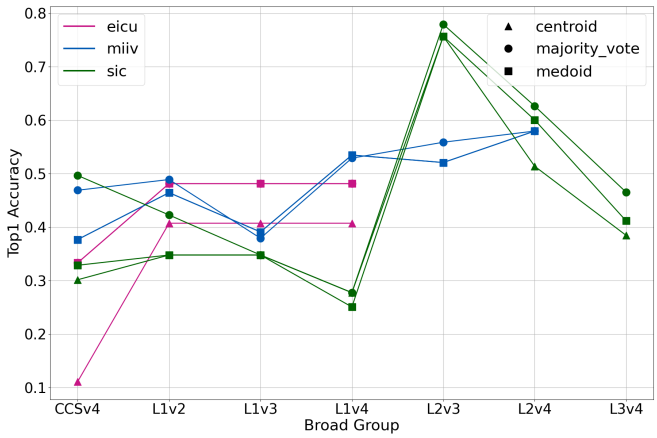

Figure 8: LSTM Average accuracy per problem definition $L_n \rightarrow L_{n+1}$ in the clustering evaluation comparison across datasets and problem formulations.

First, we observe that the performance is reversed. Regarding performance across datasets, we improved significantly for the eICU dataset, indicating that the high number of features was leveraged by the LSTM baseline more effectively, while the opposite is true for SiC, as illustrated in Figures 10, 11, and 12. We denote that the majority vote outperformed both metrics across all datasets, while the Medoid strategy performed consistently the worse.

Regarding performance across datasets we observe that some benefited more than others from the creation of nonlinear vectorial embeddings, furthermore it seems to be proportional to the number of features as illustrated in Figure 11. However for the SiC dataset the trend seems to be reversed achieving more accurate performance using the statistical method. Finally there was unanimous agreement for the difficulty of the task with respect to model performance.

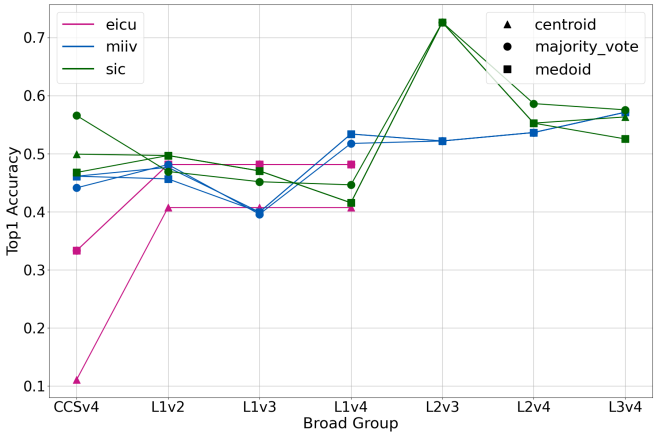

Figure 9: GRU Average accuracy per problem definition $L_n \rightarrow L_{n+1}$ in the clustering evaluation comparison across datasets and problem formulations.

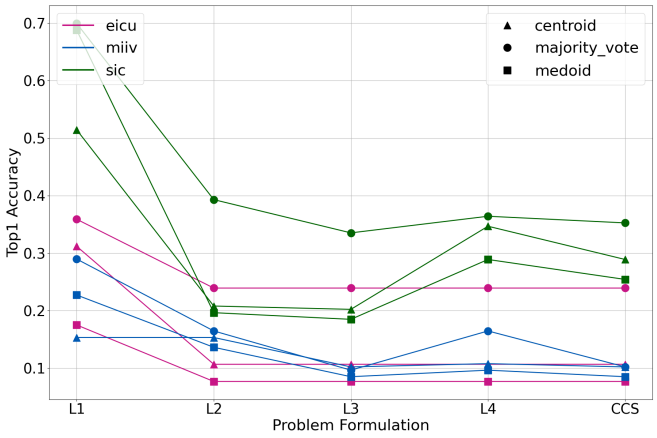

Figure 10: STAT Performance of top 1 Accuracy for $L_1 - L_4$, ICD-9-M labels across datasets.

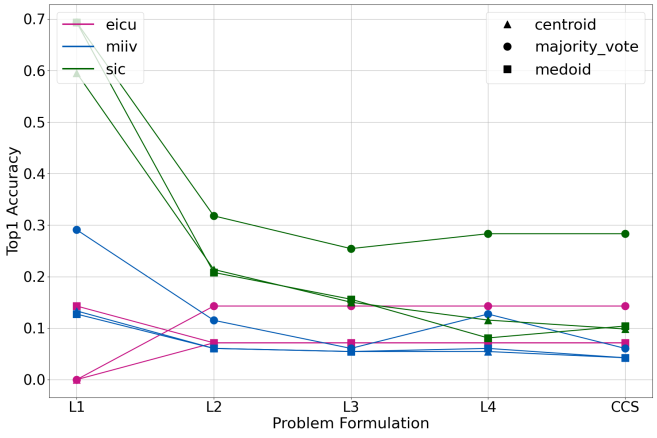

Figure 11: LSTM Performance of top 1 Accuracy for $L_1 - L_4$, ICD-9-M labels across datasets.

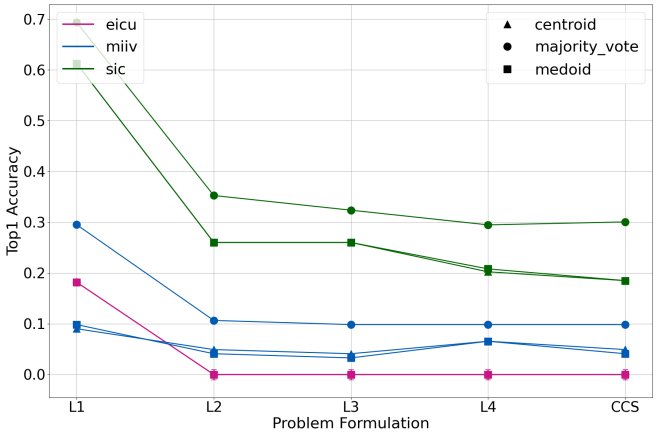

Figure 12: GRU Performance of top 1 Accuracy for $L_1 - L_4$, ICD-9-M labels across datasets.

Discussion

Self-supervised learning in multicentric ICU datasets presents several challenges influencing model performance, particularly concerning the loss function and the rediscovery of the stratification hierarchy. The LSTM baseline assumes regularly sampled data. We observe large differences between the STAT and LSTM baselines, indicating that potential adjustments could be made to define more lenient tasks.

In the clustering problems' definitions we observed significant performance deviations along the long-tailed of true labels frequency distribution. The selection process allowed to transform the problem into less skewed distributions.

We also observed that intrinsic measures like Silhouette score, while easier to optimize for, do not guarantee better performance for extrinsic evaluation. Specifically, AMI proved to be the most robust metric in terms of quantitative analysis. Furthermore, concerning the performance difference among clustering problem formulations we hypothesize this is due to the more generic cluster definitions encompassing multiple clusters versus smaller cluster problem definitions in L_3 and L_4 of our true labels set. Nonetheless, both models were able to partially rediscover CCS, L_1 and L_2 categories, demonstrating their capability to recover broad clinical patterns.

Conclusions

In this work, we demonstrate how temporal representations of ICU stays can be used to rediscover patient cohorts, a problem known as patient stratification. We evaluate the agreement of clusters derived from temporal self-supervised models' embeddings with respect to ICD-9-CM terminology. We use the stratification framework as an interpretability test to assess agreement with existing knowledge when diagnoses within any ontology or terminology are associated with a dataset. To support further advancements in patient stratification, we provided a reproducible framework using a multicentric, open ICU datasets for training and extrinsic evaluation for hierarchical unsupervised learning, partially

rediscovering ICD terminology using patient temporal embeddings.

This work addresses a critical gap in unsupervised learning for healthcare by leveraging and integrating available public ICU datasets, providing a reproducible benchmark for patient stratification. By combining temporal embeddings with hierarchical taxonomies like CCS and ICD, it bridges the methodological gap between data-driven models and clinically interpretable outcomes. This framework lays a foundation for future advancements in personalized medicine, enabling broader adoption and innovation in clinical research using open-access data.

References

Ahmed, M., Fahad Shabbir; Ali, M., Liaqat; Joseph, F., MD; Ikram, M., Asad; Ul Mustafa, M., Raza; and Bukhari, P., Syed Ahmad Chan. 2020. A statistically rigorous deep neural network analysis. *Journal of Trauma and Acute Care Surgery*.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2623–2631. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.

Alexander, N. 2023. *Unsupervised learning methods for identifying and evaluating disease clusters in electronic health records*. Doctoral, UCL (University College London). Pages: 1-436 Publication Title: Doctoral thesis, UCL (University College London).

Bennett, N.; Plečko, D.; Ukor, I.-F.; Meinshausen, N.; and Bühlmann, P. 2023. ricu: R's interface to intensive care data. *GigaScience*, 12: giad041.

Bradshaw, M. S.; Gibbs, C. P.; Martin, S.; Firman, T.; Gaskell, A.; Fosdick, B. K.; and Layer, R. M. 2023. HYPOTHESIS GENERATION FOR RARE AND UNDIAGNOSED DISEASES THROUGH CLUSTERING AND CLASSIFYING TIME-VERSIONED BIOLOGICAL ONTOLOGIES.

Choi, E.; Xu, Z.; Li, Y.; Dusenberry, M. W.; Flores, G.; Xue, Y.; and Dai, A. M. 2020. Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer. ArXiv:1906.04716 [cs, stat].

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C. K.; and Stanley, H. E. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23): E215–220.

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): 96. Publisher: Nature Publishing Group.

Hyland, S. L.; Faltys, M.; Hüser, M.; Lyu, X.; Gumbsch, T.; Esteban, C.; Bock, C.; Horn, M.; Moor, M.; Rieck, B.; Zimmermann, M.; Bodenham, D.; Borgwardt, K.; Rätsch, G.; and Merz, T. M. 2020. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26(3): 364–373. Publisher: Nature Publishing Group.

Jaume-Santero, F.; Zhang, B.; Proios, D.; Yazdani, A.; Gouareb, R.; Bjelogrlic, M.; and Teodoro, D. 2022. Cluster Analysis ofnbsp;Low-Dimensional Medical Concept Representations fromnbsp;Electronic Health Records. In *Health Information Science: 11th International Conference, HIS 2022, Virtual Event, October 28–30, 2022, Proceedings*, 313–324. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-20626-9.

Johnson, A. E.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T. J.; Hao, S.; Moody, B.; Gow, B.; et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1): 1.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1): 160035. Publisher: Nature Publishing Group.

Kim, H.; Westerman, K. E.; Smith, K.; Chiou, J.; Cole, J. B.; Majarian, T.; von Grotthuss, M.; Kwak, S. H.; Kim, J.; Mercader, J. M.; Florez, J. C.; Gaulton, K.; Manning, A. K.; and Udler, M. S. 2023. High-throughput genetic clustering of type 2 diabetes loci reveals heterogeneous mechanistic pathways of metabolic disease. *Diabetologia*, 66(3): 495–507.

Landi, I.; Glicksberg, B. S.; Lee, H.-C.; Cherng, S.; Landi, G.; Danieletto, M.; Dudley, J. T.; Furlanello, C.; and Miotto, R. 2020. Deep representation learning of electronic health records to unlock patient stratification at scale. *npj Digital Medicine*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct): 2825–2830.

Pollard, T. J.; Johnson, A. E. W.; Raffa, J. D.; Celi, L. A.; Mark, R. G.; and Badawi, O. 2018. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1): 180178.

Proios, D.; Yazdani, A.; Bornet, A.; Ehrsam, J.; Rekik, I.; and Teodoro, D. 2023. Leveraging patient similarities via graph neural networks to predict phenotypes from temporal data. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10. IEEE.

Rosenberg, A.; and Hirschberg, J. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. *Journal of Computational Linguistics*.

Sadeghi, S.; Hempel, L.; Rodemund, N.; and Kirsten, T. 2024. Salzburg Intensive Care database (SICdb): a detailed exploration and comparative analysis with MIMIC-IV. *Scientific Reports*, 14(1): 11438. Publisher: Nature Publishing Group.

Teschendorff, A. E.; Naderi, A.; Barbosa-Morais, N. L.; and Caldas, C. 2006. PACK: Profile Analysis using Clustering and Kurtosis to find molecular classifiers in cancer. *Bioinformatics*, 22(18): 2269–2275.

van de Water, R.; Schmidt, H.; Elbers, P.; Thoral, P.; Arnrich, B.; and Rockenschaub, P. 2024. Yet Another ICU Benchmark: A Flexible Multi-Center Framework for Clinical ML. ArXiv:2306.05109 [cs].

Vinh, N. X.; Epps, J.; and Bailey, J. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*.

Vrbik, I.; Stephens, D. A.; Roger, M.; and Brenner, B. G. 2015. The Gap Procedure: for the identification of phylogenetic clusters in HIV-1 sequence data. *BMC Bioinformatics*, 16: 355.

WHO. 2019. International statistical classification of diseases and related health problems (11th ed.).

Zang, C.; and Wang, F. 2021. SCEHR: Supervised Contrastive Learning for Clinical Risk Prediction using Electronic Health Records. *CoRR*, abs/2110.04943.