MIEO: encoding clinical data to enhance cardiovascular event prediction

Davide Borghini[1], Davide Marchi[1], Angelo Nardone[1], Giordano Scerra[1], Silvia Giulia Galfrè[1], Alessandro Pingitore[2], Giuseppe Prencipe[1], Corrado Priami[1], Alina Sîrbu[3*]

1. Department of Computer Science, University of Pisa, Pisa, Italy. 2. Clinical Physiology Institute, CNR, Pisa, Italy. 3. Department of Computer Science and Engineering, University of Bologna, Bologna, Italy.

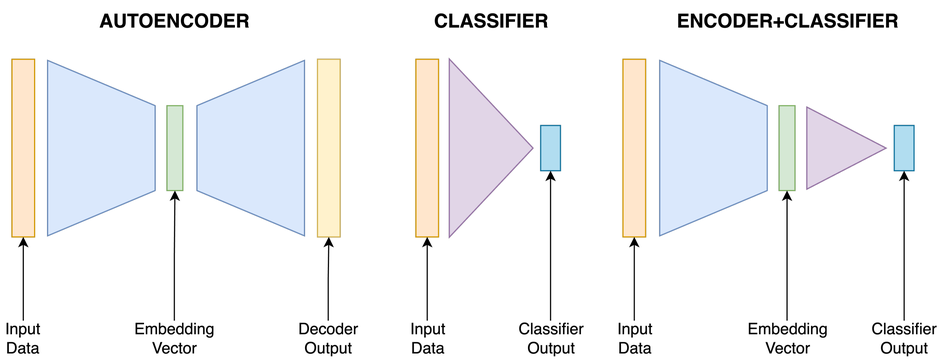
Abstract

- Clinical data is increasingly available, facilitating the use of machine learning (ML) in clinical decision-making. However, two major challenges persist:
  - **Limited labeled** data for supervised learning
  - Heterogeneous records with **missing values**
- We present a new method to address both issues. We learn a patient **representation** that can **handle missing data** from **unlabelled clinical data** in a self supervised fashion. These representations are used to train an ANN for predicting cardiovascular death.
- Key results:
  - Latent embeddings improve balanced accuracy compared to training directly on raw data.
  - The approach should become particularly effective when large amounts of unlabelled data become available.

Models

The study was based on three models:

- An **autoencoder** modified to fit our needs (**MIEO**), used to generate embeddings of the clinical data.
- A **neural network classifier** serving as a baseline. It directly classified the raw data; for a fair comparison, the null mask was also provided as input.
- A **neural network classifier trained on the autoencoder embeddings**. This approach allows the use of unlabelled data and may benefit from a more informative representation.

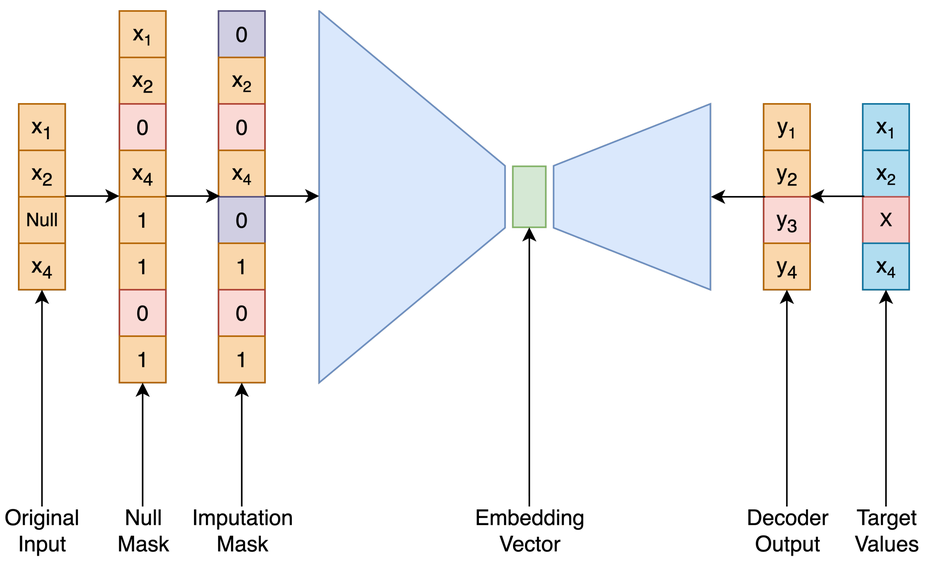


MIEO Autoencoder

- Follows the structure of a classical **autoencoder**. We modified the base architecture to handle null values and optimize for binary data.
- **Null values:** we pass to the encoder a mask indicating which values are not present for a given patient. To make the model able to infer the missing features we randomly masked additional elements during training.
- **Binary data and continuous data** are handled by the same encoder, but in the decoder the error is computed separately (binary cross entropy for binary, and mean squared error for continuos). The two errors are then **weighted by a user defined hyperparameter** λ.
- The final **loss equation** is:

$$\mathbf{Loss}_\lambda(m, x) = \lambda \sum_{i \in \text{bin}} m_i \cdot \text{MSE}(x_i, \hat{x}_i) + (1 - \lambda) \sum_{i \in \text{con}} m_i \cdot \text{BCE}(x_i, \hat{x}_i)$$

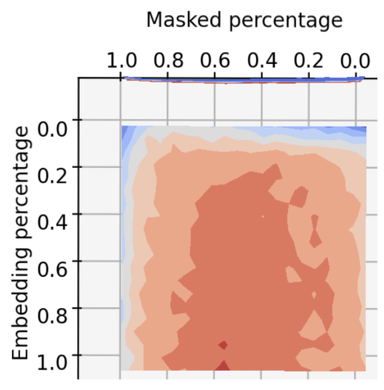
where m_i is either 1 or 0 if the value is respectiveley not null or null and λ is the balancing hyperparameter.

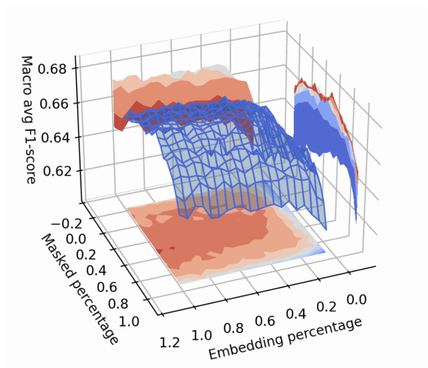


Dataset

- **8067 IHD patients** hospitalized at the CNR Clinical Physiology Institute in Pisa between 1977 and 2011.
- **68 clinical variables**, of which most are binary (e.g., smoking, diabetes) and the remainder are continuous measures (e.g., creatinine, cholesterol).
- After preprocessing the overall percentage of **missing data was 2.98%**, with only two variables showing more than 50% missing values.
- The **classification target** was **cardiovascular death within 8 years** from the first visit, resulting in a **labelled dataset of 3770 patients** and an **unlabelled set of 4297 patients**.

Results





- MIEO was trained by **combining labelled and unlabelled data**.
- ANN models were trained using only **labelled data**.
- **Balanced accuracy** (macro-average recalls) was adopted as the main evaluation metric, due to the class imbalance in the dataset.
- Test and validation results were **consistent**, showing good generalisation.
- The MIEO+ANN model achieved slightly **higher balanced accuracy** compared to the standard ANN, indicating a better ability to recognise CVD events.

		MIEO +ANN			ANN			
	Class	Precision	**Recall**	F1	Precision	**Recall**	F1	Support
Validation dataset	0.0	0.908	0.790	0.845	0.867	0.898	0.882	472
	1.0	0.484	0.710	0.576	0.579	0.5038	0.539	131
	Accuracy			0.773			0.813	603
	Macro avg	0.696	**0.750**	0.710	0.723	0.701	0.711	603
	Weighted avg	0.816	0.773	0.786	0.805	0.813	0.808	603
Test dataset	0.0	0.88	0.80	0.84	0.85	0.91	0.88	578
	1.0	0.49	0.65	0.56	0.62	0.47	0.54	176
	Accuracy			0.76			0.81	754
	Macro Avg	0.69	**0.72**	0.70	0.73	0.69	0.71	754
	Weighted Avg	0.79	0.76	0.77	0.80	0.81	0.80	754

Conclusions

- The method can learn from unlabelled data with missing values, making it **useful in real clinical settings**.
- While current improvements are small, using larger datasets could create richer embeddings that cover different conditions.

Future work will investigate:

- The use of **different models** for the downstream task, and other self supervised architectures.
- The integration of **additional data** in the self supervised phase and a thorough analysis of imputation performance.

Acknowledgements

This work was supported by the European Union – Next Generation EU, within the National Recovery and Resilience Plan (PRIN 2022 – MEDICA Project, CUP I53D23003720006; Tuscany Health Ecosystem, ECS00000017, Spoke 3, CUP B83C22003920001); by the European Union – Horizon 2020 Program (SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics, Grant Agreement 871042); and by the University of Pisa through the SPARK project.

*Corresponding author, alina.sirbu@unibo.it