CAMELOT++: Enhancing Patient Phenotype Discovery through Time-Series Clustering and Survival Analysis

Katarina Vukosavljević[1], Henrique Aguiar[1], Julia Hippisley-Cox[2], Tingting Zhu[1]
[1]Department of Engineering Science, University of Oxford, [2]Department of Primary Care Sciences, University of Oxford

ABSTRACT

Understanding intensive care outcomes is essential for improving intensive care unit (ICU) management. In this paper, we present a novel deep clustering method for multi-task learning on temporal electronic health records (EHRs), CAMELOT++. Firstly, we extend a current state-of-the-art clustering model to form CAMELOT+. Here we introduce mixed-data modelling, extending the encoder to encode static and dynamic features separately and, as a result, enhance clustering performance and predictive accuracy. Secondly, we extend this model to a multi-task setting, CAMELOT++, and integrate survival analysis into the clustering framework, allowing for cluster-specific survival curve visualisation, thus improving both interpretability and outcome prediction. Using the MIMIC-III dataset, our models are evaluated on the separability of the cluster, the classification of primary diagnoses, and the prediction of survival. Our results demonstrate significant improvements in predictive ability and survival prediction over existing models, as well as effective multi-task learning.

Key Contributions

1. Novel Model : Extension of the CAMELOT model to CAMELOT+ for mixed-data encoding and to CAMELOT++ for multi-task learning.

2. Enhanced Interpretability: Integration of survival analysis for visualisation of cluster-specific survival curves.

3. Empirical Validation: Significant improvements in prediction tasks using the MIMIC-III dataset.

Methods

Dataset: MIMIC-III Extract
- **Cohort**: 23,422 patients, with age ≥18 years, ICU stays between 12 hours and 10 days
- **Features**: 4 static variables, 29 vital signs, and 10 interventions

Tasks:
- Primary diagnosis classification (ICD9 Code Grouping)
- Survival prediction for the first 48 hours of ICU stay

CAMELOT++ Architecture

- **Encoder**: Encodes temporal data with an LSTM attention encoder and encodes static data with an MLP, combining them in the latent space.

- **Phenotype Discovery Net (PDN)**: Assigns cluster probabilities and samples cluster phenotypes using the Gumbel-Softmax estimator.

- **Phenotype Assignment (P)**: Contains a set of trainable cluster representation vectors and passes the assigned cluster representation to the downstream task.

- **Task-specific Networks**
  - **OutcomeNet (O)**: MLP for multi-class classification
  - **SurvNet (S)**: DySurv's MLP for survival probability over time estimation.

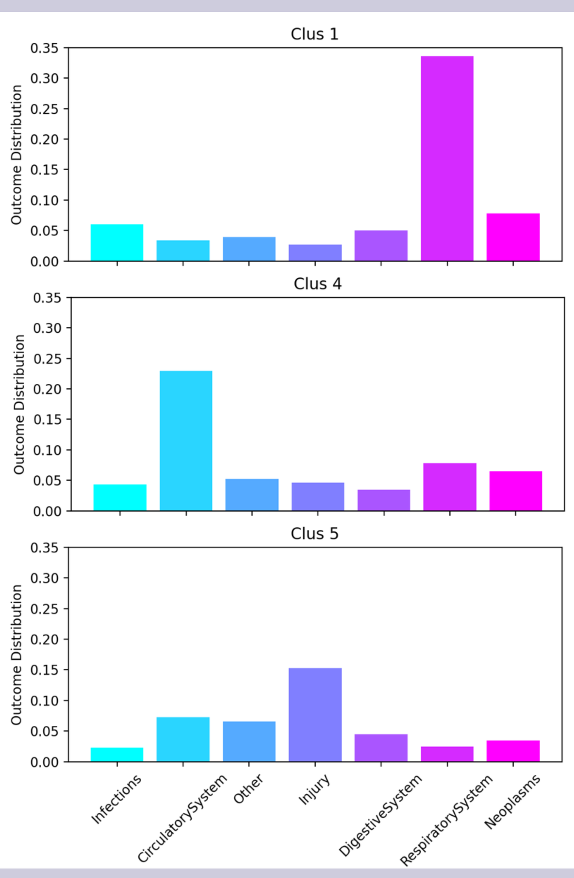
- **Loss Functions**:
  - Combines clustering, prediction, and survival losses.
  - Joint optimisation ensures robust feature-time pair discovery, robust cluster assignments and prevents cluster collapse.

Phenotypes

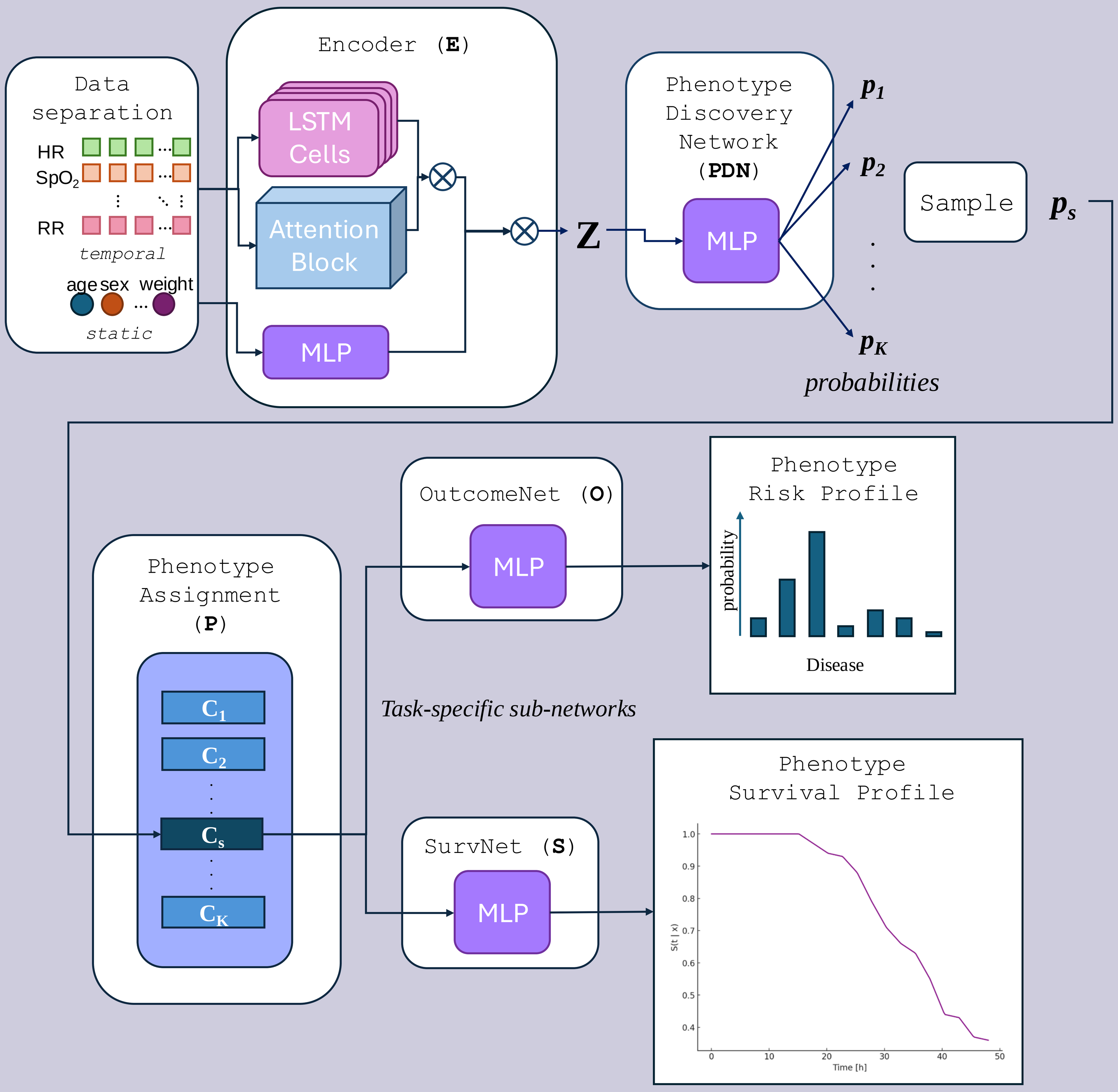
Both CAMELOT+ and CAMELOT++ find up to 20 phenotypically distinct clusters.

Each cluster contains patients with a mixture of different outcomes.

Both models successfully identify predictive clusters of minority classes (Injury).



CAMELOT++

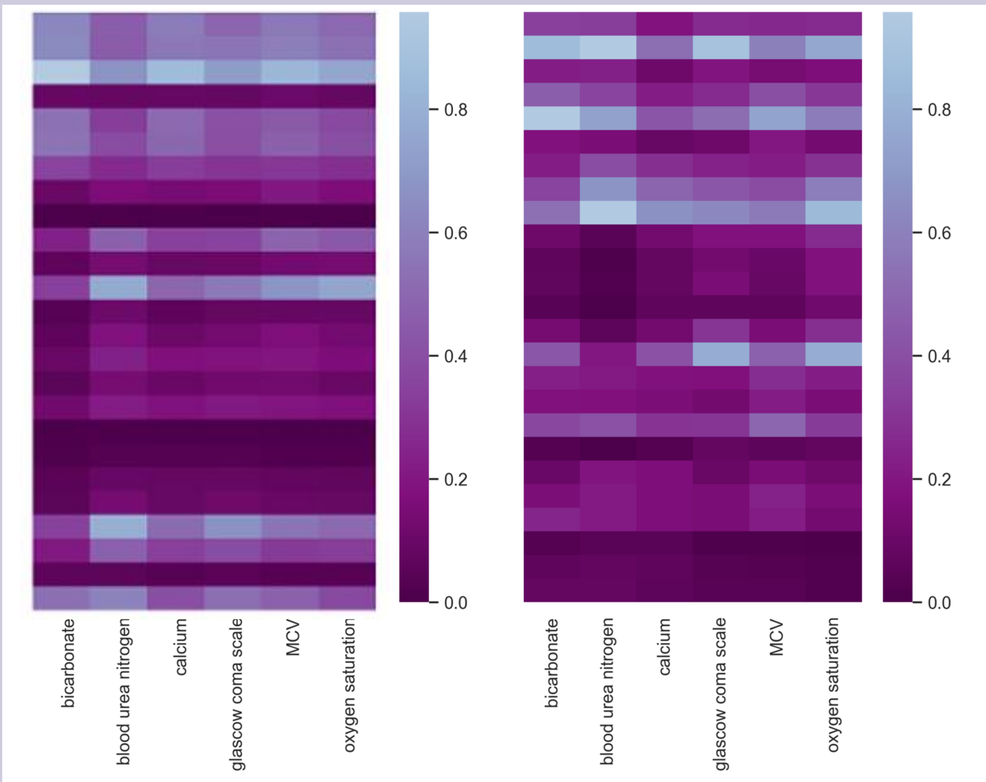


Visualisation

Attention Block weight visualisation for CAMELOT (left) and CAMELOT+ (right) for a selected few temporal features.

CAMELOT+'s encoder leads to higher quality latent embeddings as:

- CAMELOT+ identifies specific important feature-time pairs
- CAMELOT rather finds important time-points



Example feature attention maps for one cluster when all data is passed through the LSTMs (CAMELOT, left) compared to temporal data only (CAMELOT+, right)

Results

- CAMELOT+ outperforms other clustering methods on disease classification.
- CAMELOT++ (multi-task setting) performs comparably.

Method	AUROC	AUPRC	F1	Recall
SVM	0.797 (±0.003)	0.403 (±0.007)	0.375 (± 0.007)	0.375(±0.022)
XGB	**0.833 (±0.002)**	**0.518 (±0.008)**	**0.437 (±0.002)**	**0.423 (±0.002)**
TSKM	0.590 (±0.002)	0.229 (±0.001)	0.093 (±0.014)	0.148 (±0.005)
AC-TPC	0.624 (±0.021)	0.329 (±0.015)	0.230 (±0.024)	0.284 (±0.027)
CAMELOT	0.756 (±0.001)	0.352 (±0.001)	0.317 (±0.005)	0.332 (±0.004)
CAMELOT+	**0.790 (±0.004)**	**0.392 (±0.003)**	**0.283 (±0.014)**	0.271 (±0.013)
CAMELOT++	0.785 (±0.004)	0.384 (±0.006)	0.208 (±0.020)	0.224 (±0.013)

Disease classification results

- TSKM extension achieves highest C-index due to clustering patients according to length of stay
- CAMELOT++ achieves best calibration confirmed by IBS and NBLL scores

Method	$C^{td}_{ind}\uparrow$	IBS ↓	NBLL ↓
TSKM & DySurv	**88.3** (±1.4)	0.018 (±0.001)	0.072 (±0.011)
CAMELOT & DySurv	68.0 (±12.2)	0.023 (±0.006)	0.113 (±0.036)
CAMELOT+ & DySurv	78.5 (±0.05)	0.021 (±0.005)	0.103 (±0.030)
CAMELOT++	84.9 (±1.4)	**0.017 (±0.001)**	**0.072 (±0.003)**

Survival prediction results

References

1. H. Aguiar, M. Santos, P. Watkinson, and T. Zhu. "Learning of Cluster-based Feature Importance for Electronic Health Record Time-series". In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 161–179. PMLR, 2022. [Online]. Available: https://proceedings.mlr.press/v162/aguiar22a.html.
2. M. Mešinović, P. Watkinson, and T. Zhu. "DySurv: Dynamic Deep Learning Model for Survival Prediction in the ICU", 2023. *arXiv:2310.18681*.
3. C. Lee and M. Van Der Schaar. "Temporal Phenotyping using Deep Predictive Clustering of Disease Progression". In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5767–5777. PMLR, 2020. [Online]. Available: http://proceedings.mlr.press/v119/lee20h/lee20h.pdf.
4. S. Wang, M. B. A. McDermott, G. Chauhan, M. Ghassemi, M. C. Hughes, and T. Naumann. "MIMIC-Extract: a data extraction, pre-processing, and representation pipeline for MIMIC-III". In CHIL '20: Proceedings of the ACM Conference on Health, Inference, and Learning, pages 222–235, 2020. doi: https://doi.org/10.1145/3368555.3384469.

Acknowledgement

K. Vukosavljević is supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1).