

Final Presentation

Machine Learning-Based ICU Stay Prediction



Leibniz ScienceCampus
**Digital Transformation
of Research**

Student: Tristan Heil

Advisor: Prof. Harald Sack

CO-Referee: PD Dr.-Ing. Axel Loewe

Supervisors: M.Sc. Silvia Becker, Dr. Anna Jacyszyn, Dr.-Ing. Genet Asefa Gesese



Location: online, 18.06.2026

Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

Motivation

5-8%

ICU beds

Percentage of total hospital
beds dedicated to intensive
care unit

Rhodes, A., et al. 2012

3000€

Daily costs

Staff, equipment, diagnostics

Tan, S. S., et al. 2008

11d

Median length
of stay

50% of patients discharged
within 11d

Chauhan, V. K., et al. 2022



ICUs represent specialized units for critically ill patients requiring 24/7 monitoring by specialized staff and sophisticated equipment

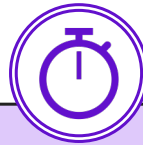
Motivation



ICU capacity limits exceeded
during Covid-19



www.klinikum.uni-heidelberg.de



Physician decisions based on
experience and limited data



www.saarbruecker-zeitung.de

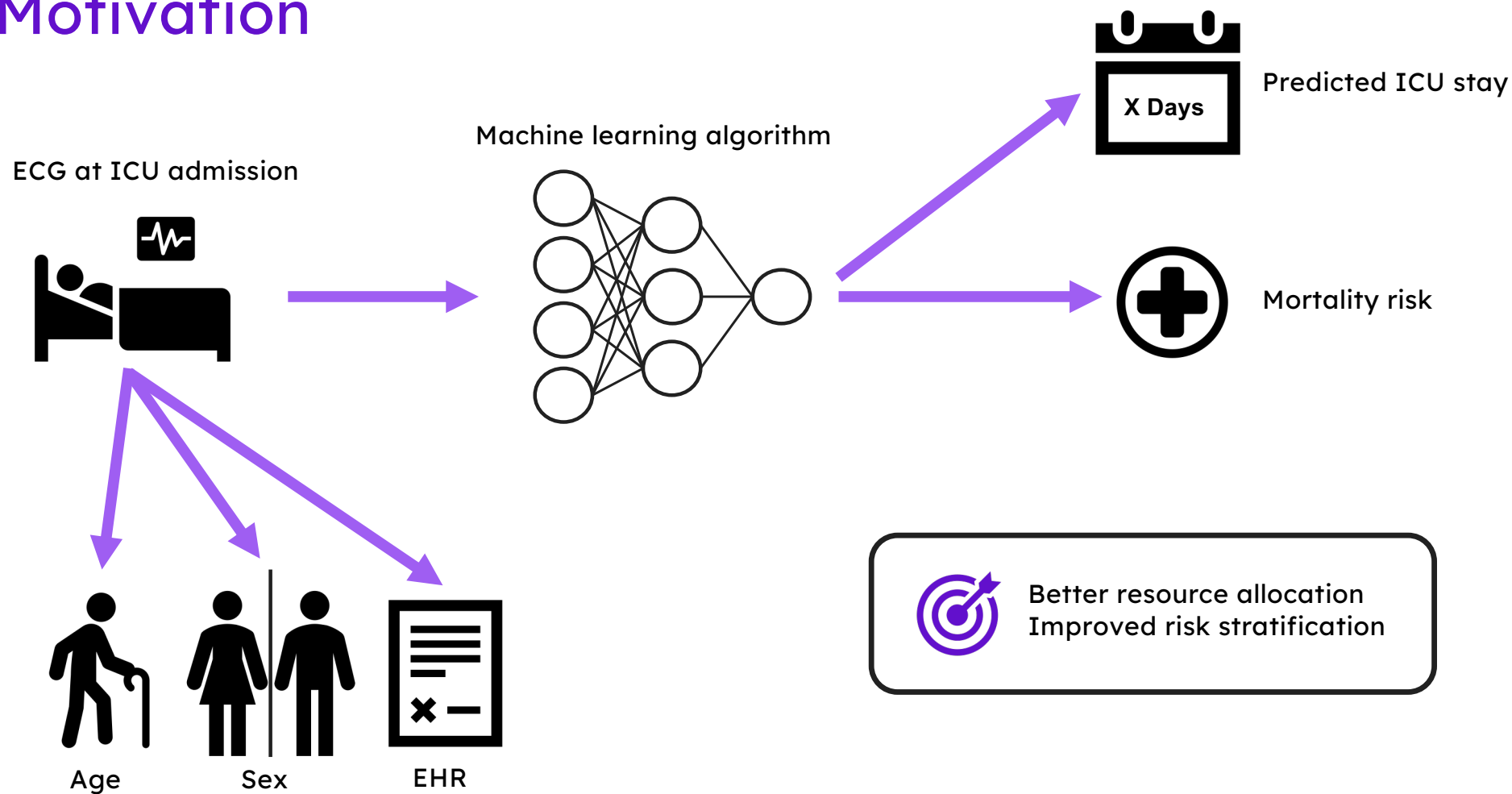


ECG data is routinely available
at admission



canadiem.org/interpreting-critical-vital-signs

Motivation



Research Questions

Q1

What level of predictive performance and calibration do DL models achieve when predicting ICU LOS and mortality from the first raw ECG recorded at ICU admission?

Q2

To what extent does integrating ECG with structured EHR and demographic data improve predictive performance compared to ECG-only models?

Q3

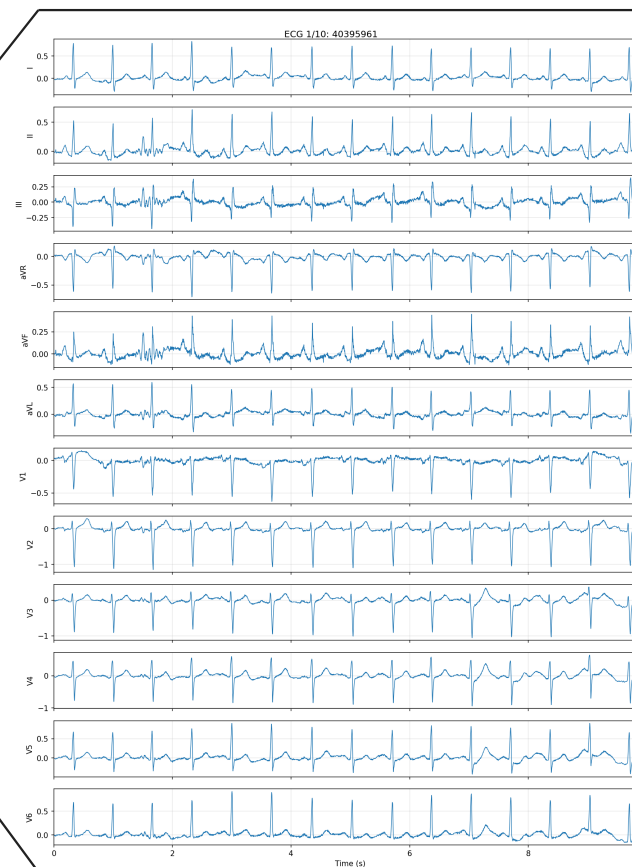
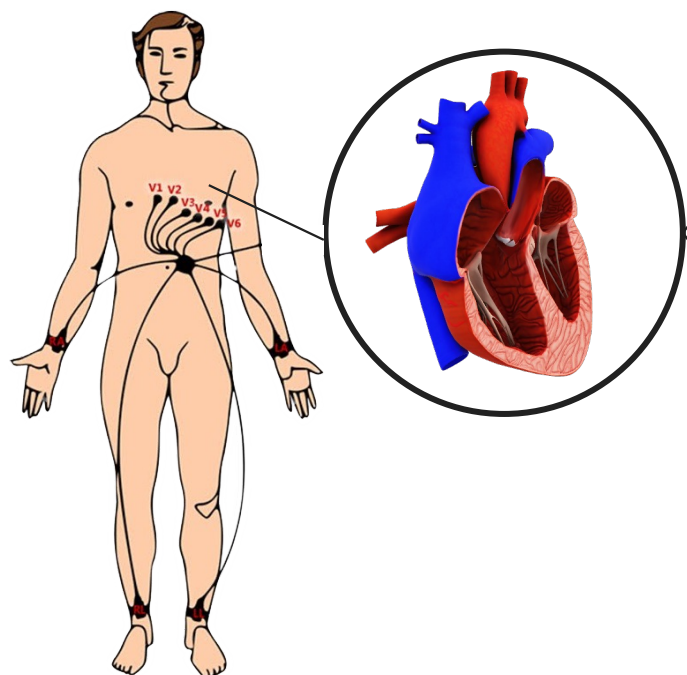
Which ML architectures are most suitable for modeling ECG-based ICU outcome prediction in terms of accuracy and computational cost?

Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

Fundamentals and State of the Art

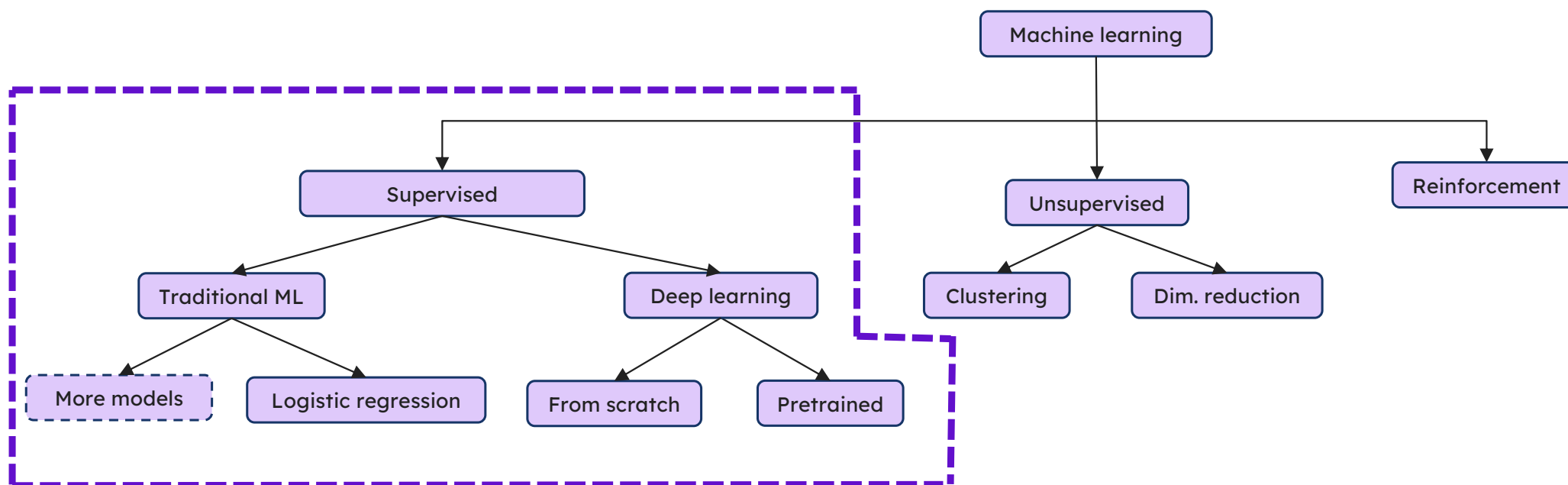
Electrocardiogram



12-lead
500 Hz
10s

Fundamentals and State of the Art

Machine learning



Fundamentals and State of the Art

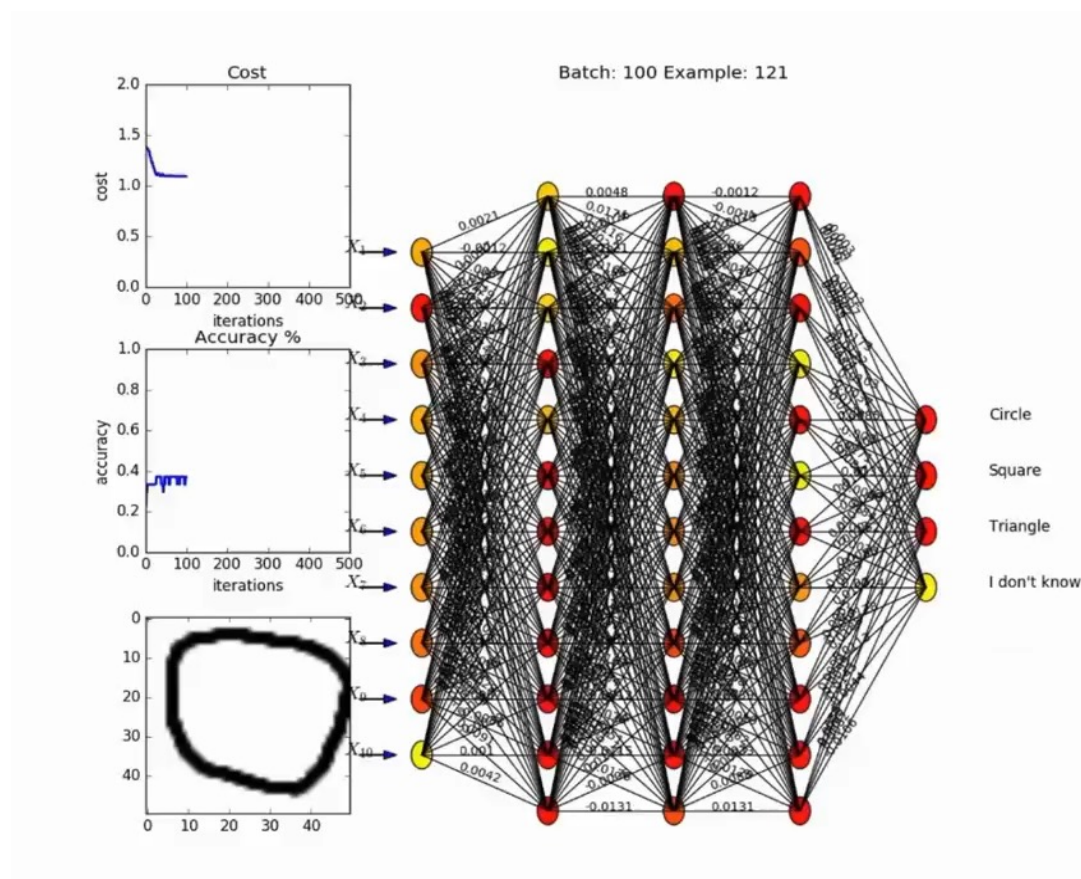
Deep Learning

Input → Hidden Layers → Output

Weight connections

Cost (loss) decreases over time

Accuracy increases over time

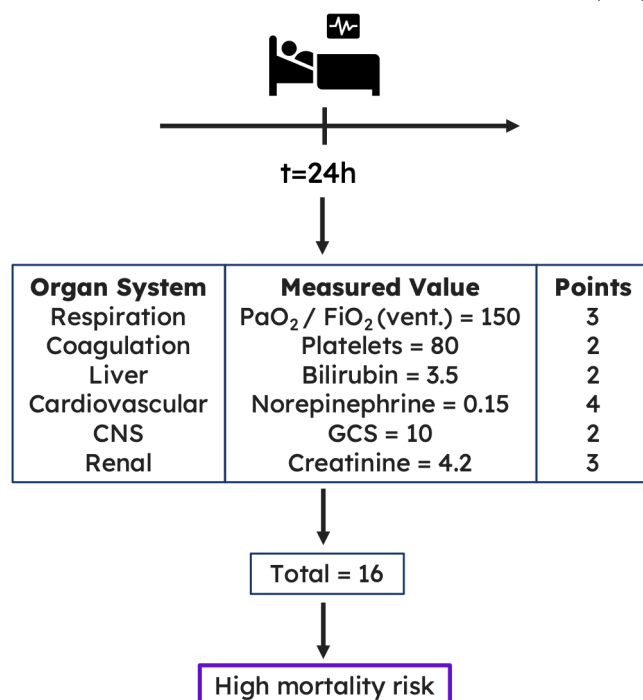


[youtube.com/shorts/fQ376G-Ek1E](https://www.youtube.com/shorts/fQ376G-Ek1E)

Fundamentals and State of the Art

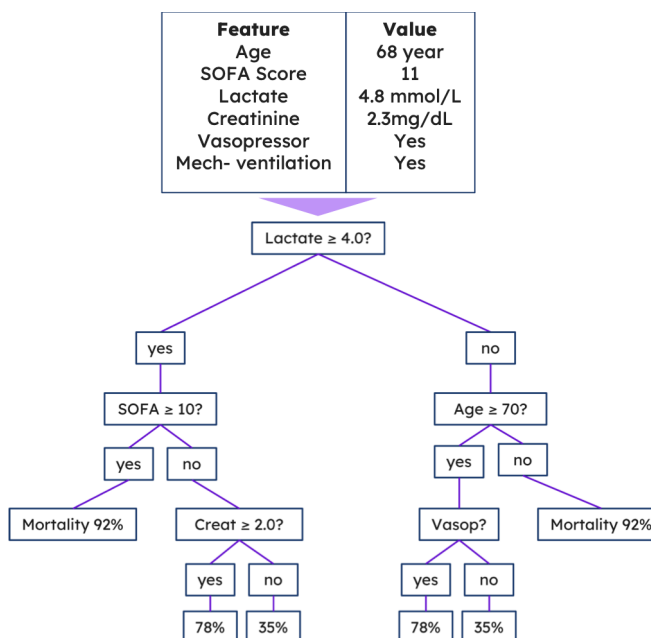
Classical Scoring Systems

Vincent JL, et al., 1996



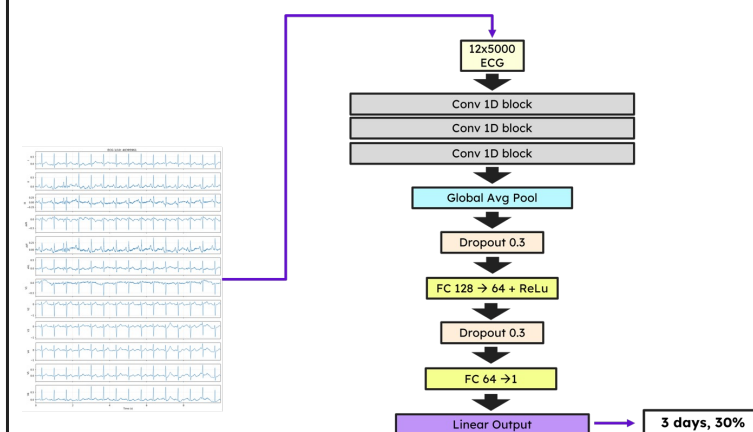
ML on static EHR data

Shillan, et al., 2019



DL on time-series data

Harutyunyan, et al., 2019



Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

Data

Datasource: MIMIC-IV Dataset



MIMIC-IV Database

- Deidentified dataset
- Recorded at Beth Israel Deaconess Medical Center in Boston
- 365,000 patients
- 940,00 ICU stays



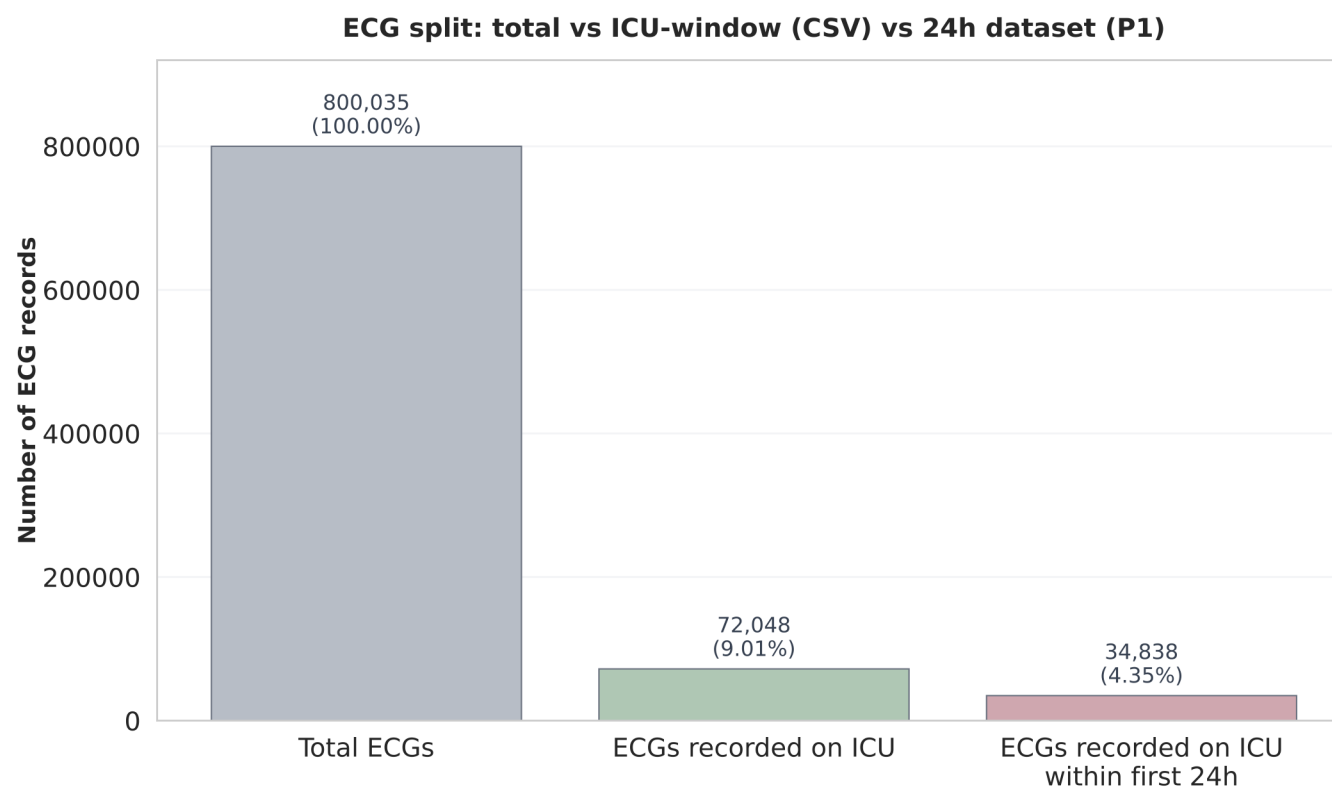
MIMIC-IV-ECG

- 160,000 unique patients
- 95,000 ICU stays
- 800,000 ECGs

Data types: ECG signals, demographics (age+sex), electronic health care records (EHR), ICU Unit

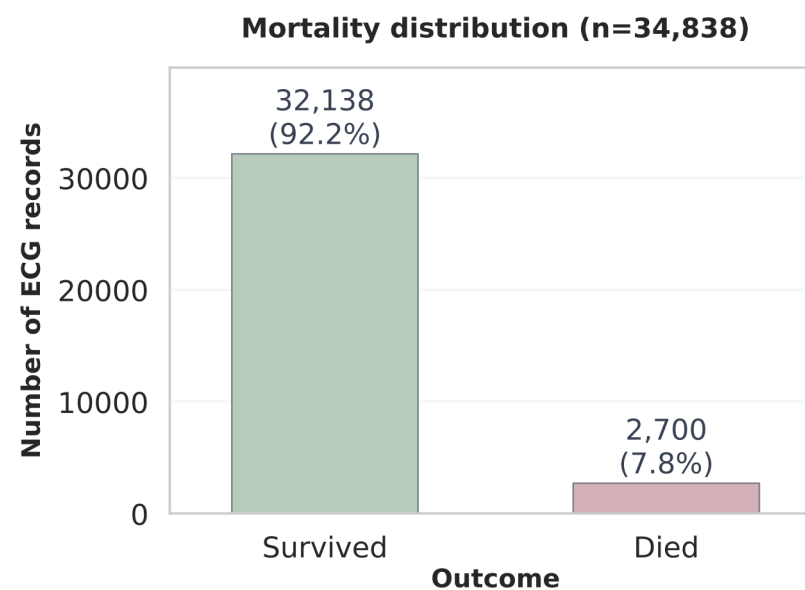
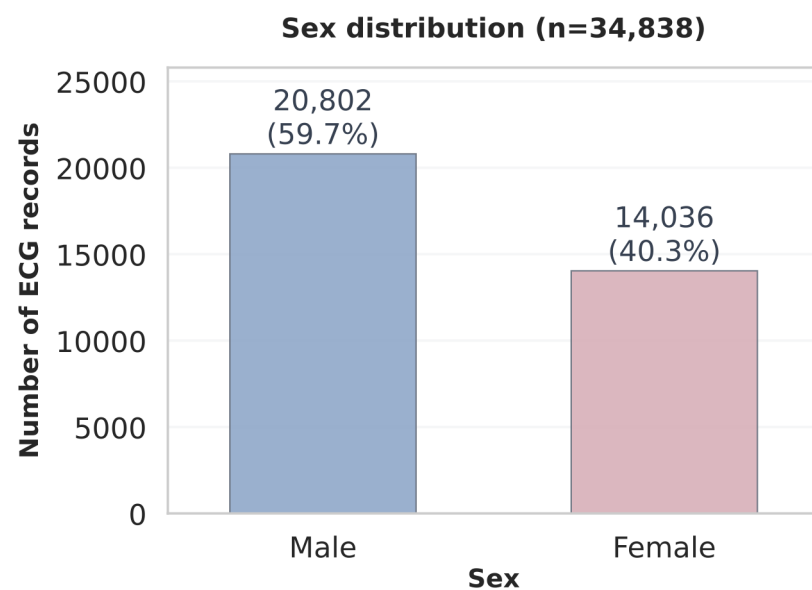
Data

MIMIC-IV ECG dataset



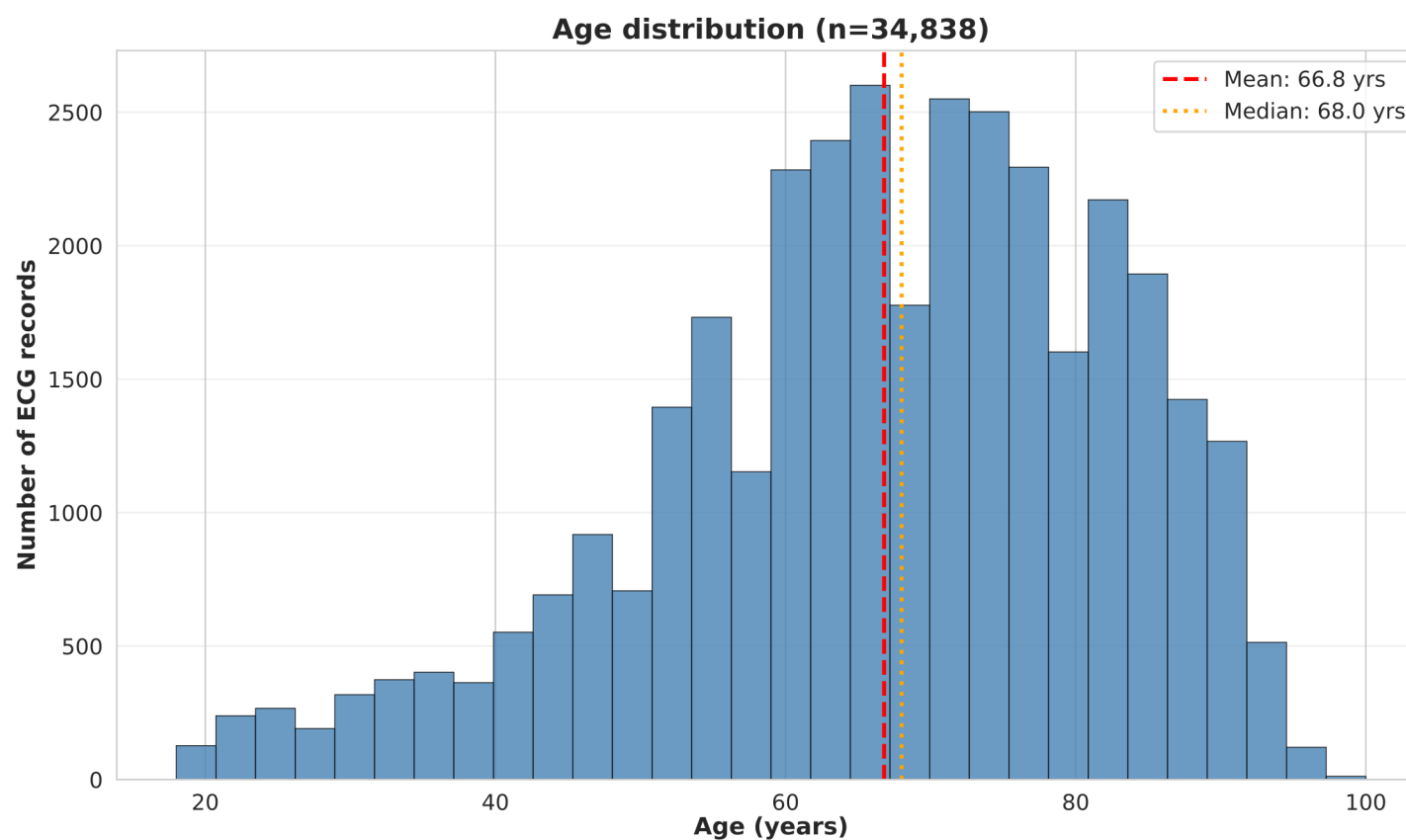
Data

MIMIC-IV ECG dataset



Data

MIMIC-IV ECG dataset

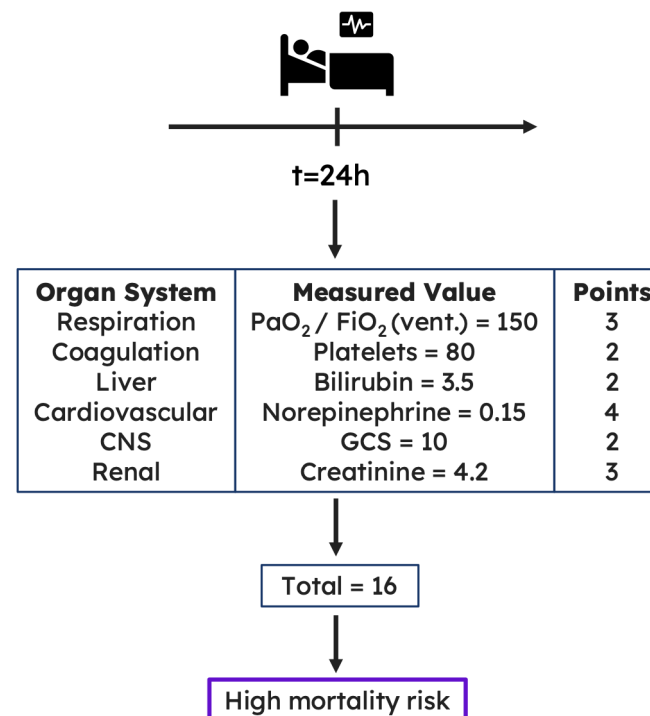


Data

Variant Included variables

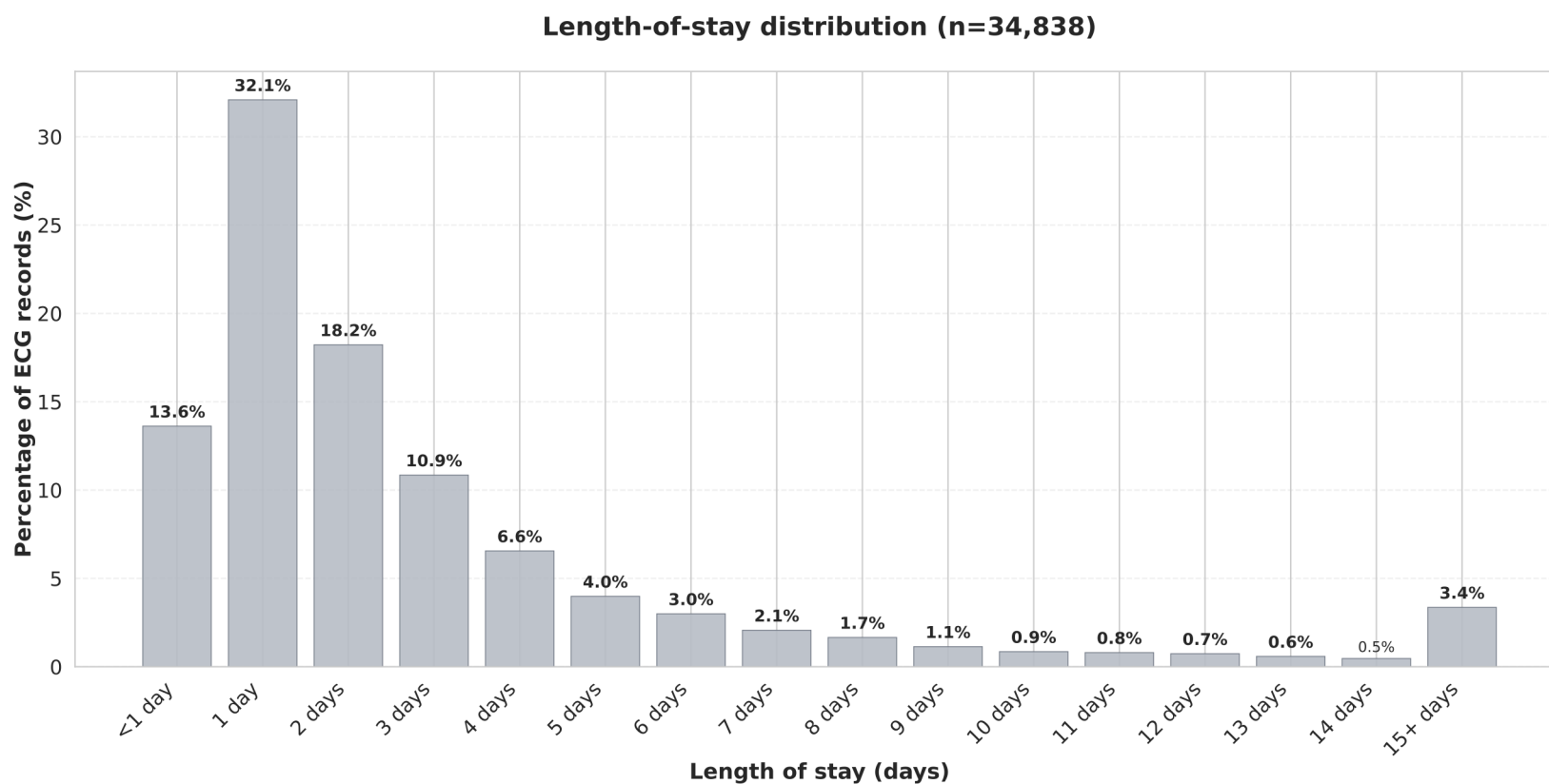
Full EHR	Respiratory rate, oxygen saturation, mean arterial pressure (MAP), systolic blood pressure (SBP), diastolic blood pressure (DBP), body temperature, Glasgow Coma Scale (GCS), eye and motor subscores, platelets, creatinine, urine output, weight
Small EHR	Respiratory rate, oxygen saturation, SBP, DBP, body temperature, weight

SOFA score



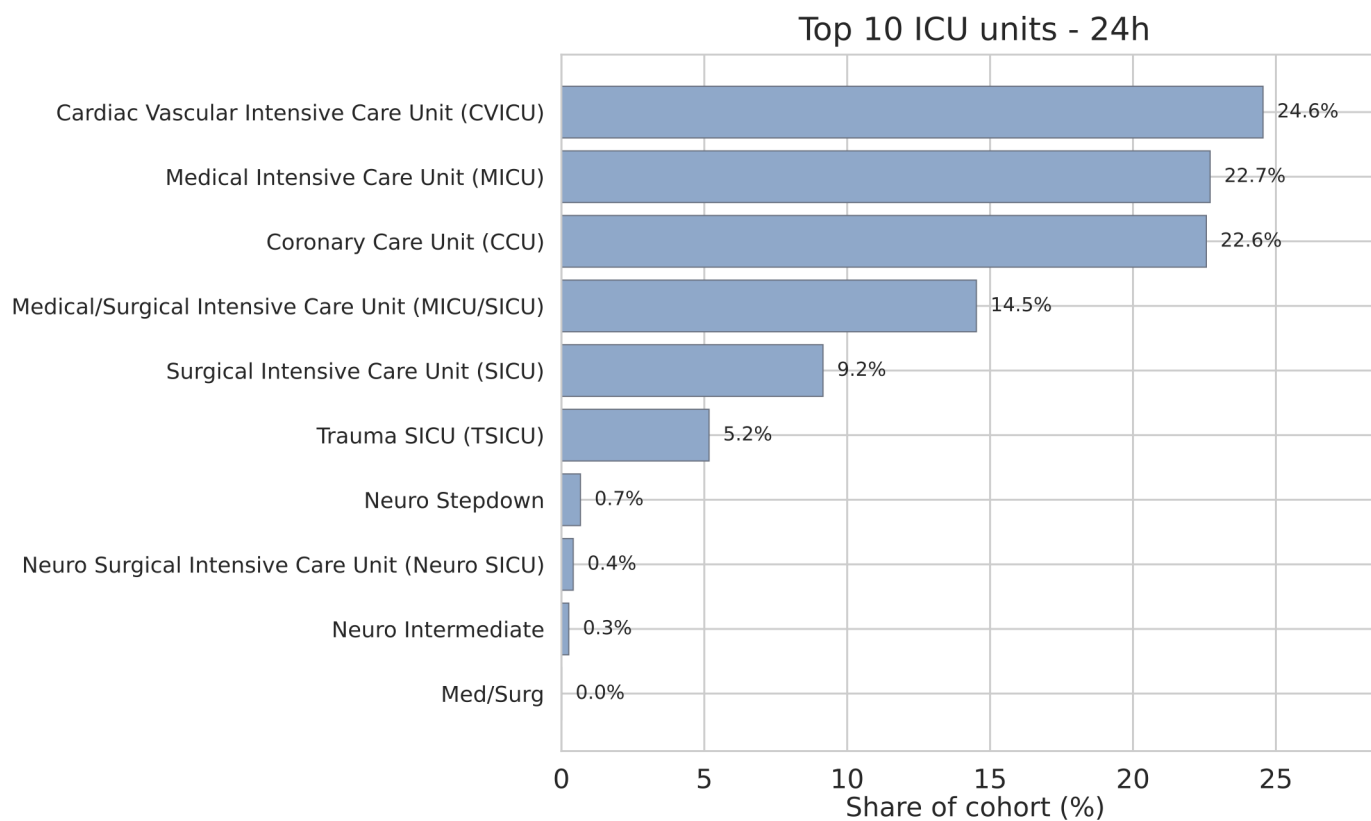
Data

MIMIC-IV ECG dataset



Data

MIMIC-IV ECG dataset



Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

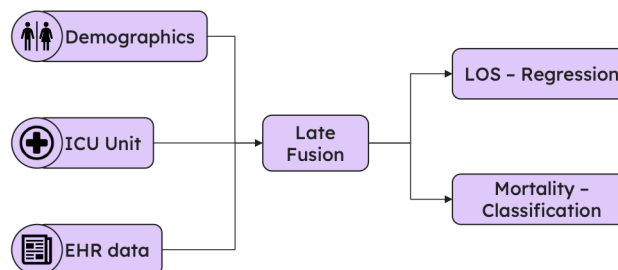
Experimental Design and Study Setup

Compute



- BW Uni Cluster 3.0
- PyTorch framework
- SLURM scheduling

Multimodal strategy



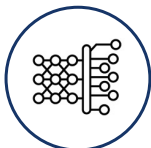
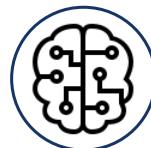
- Multitask setup
- Regression task for LOS
- Classification task for mortality

Experiment Identifiers

Prefix	Category
A	Architecture
P	Preprocessing
H	Hyperparameters
AG	Augmentation
D	Dataset variant

- Controlled experiment Ids
- Structured naming scheme
- Full documentation
- Ensures reproducibility and traceability

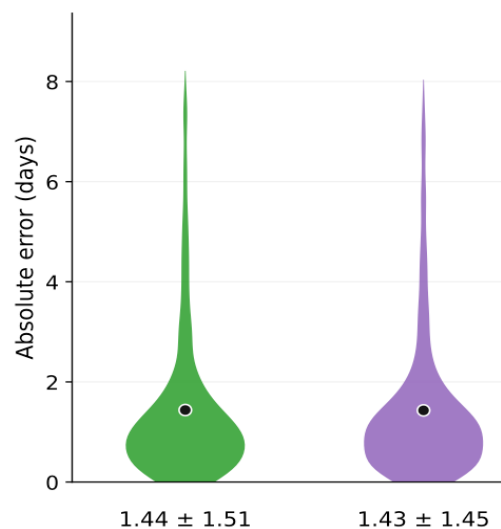
Compared Model Architectures

XGBoost	CNN scratch	Uni LSTM	Bi LSTM	Hybrid CNN LSTM	DeepECG	HubertECG
						
Classical ML Handcrafted Features	Local CNN 3x Conv1D ~47k param.	Sequential unidirectional LSTM 2x LSTM layer ~230k param.	Sequential bidirectional LSTM 2x BiLSTM layer ~600k param.	CNN + BiLSTM Conv blocks → LSTM ~475k param.	Pretrained Transformer WCR encoder ~91mio param.	Pretrained Transformer SSL encoder ~90mio param.

Evaluation Framework

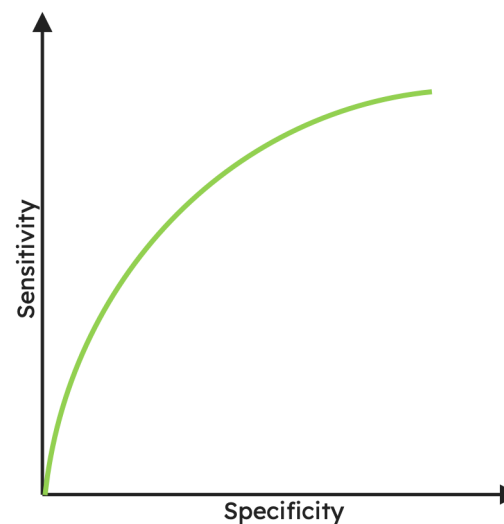
Length of Stay (LOS)

- Task: Regression
- Primary metric: Mean Absolute Error (MAE)
- Shows Error distribution



Mortality

- Task: Classification
- Primary: AUC-ROC
- Sensitivity vs specificity



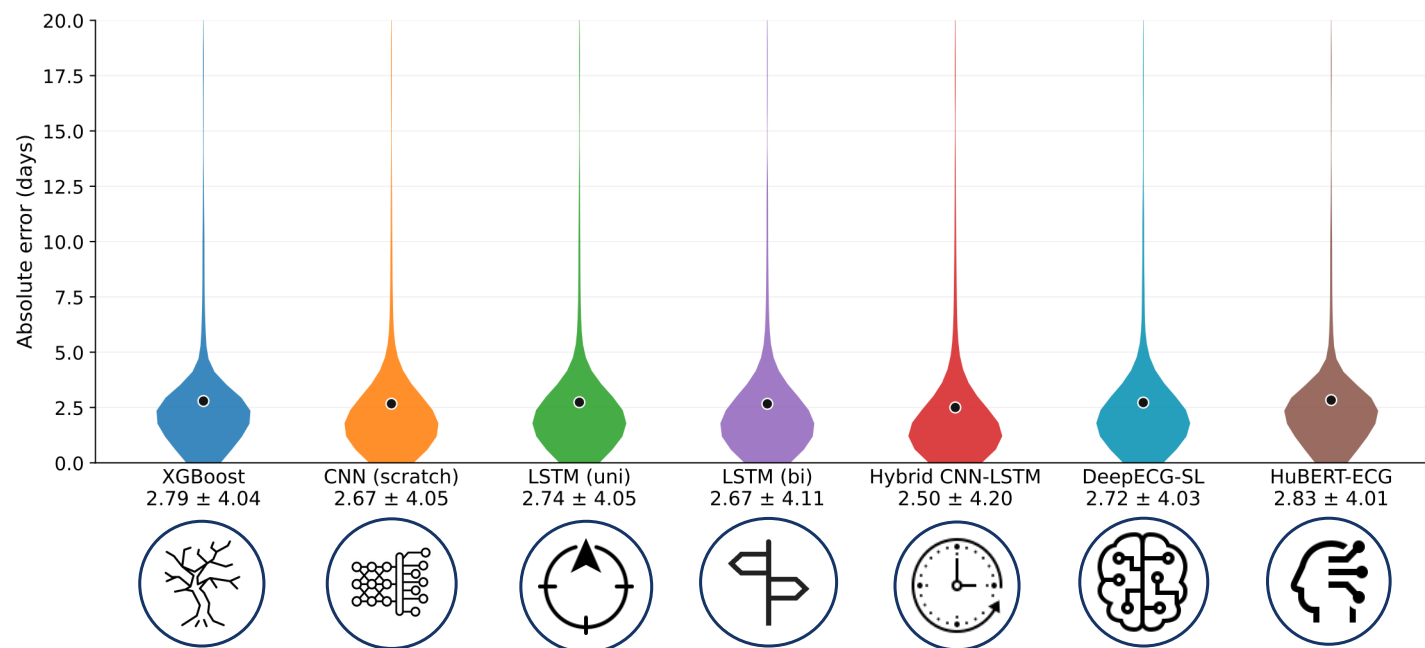
LOS prediction

ECG only baseline

MAE similar across models

Hybrid performed best
HuBERT-ECG performed
worst

Right-skewed errors with
long tails



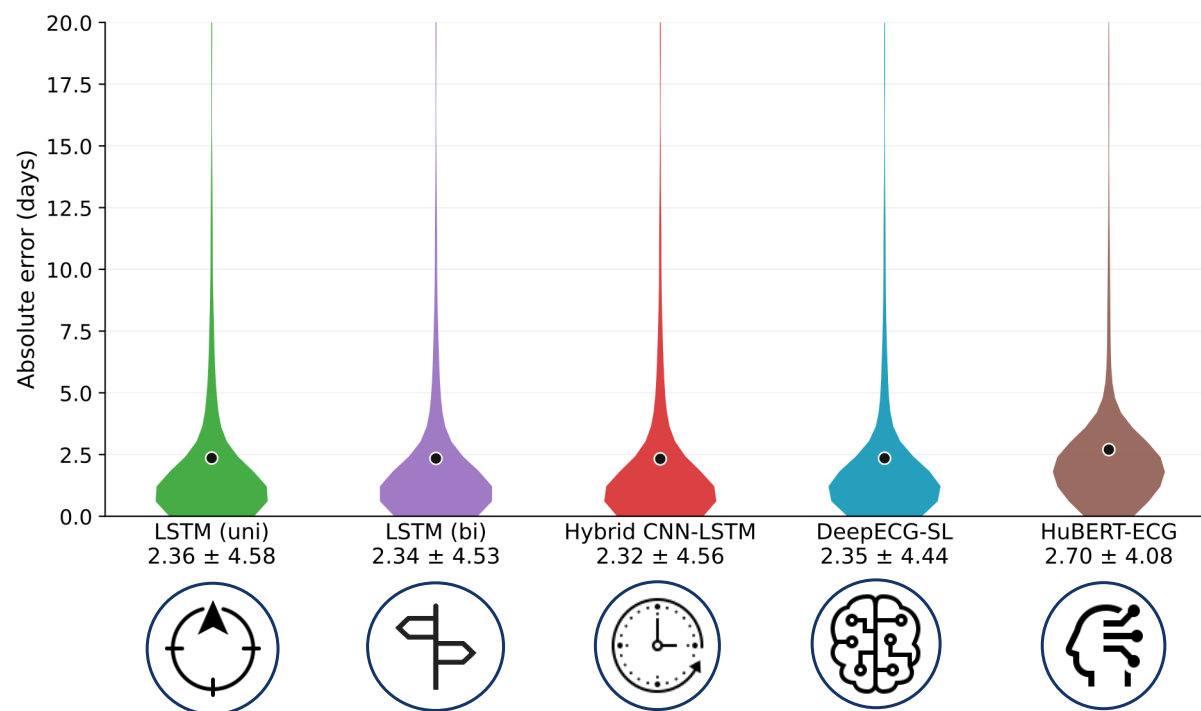
LOS prediction

Focus on 5 architectures

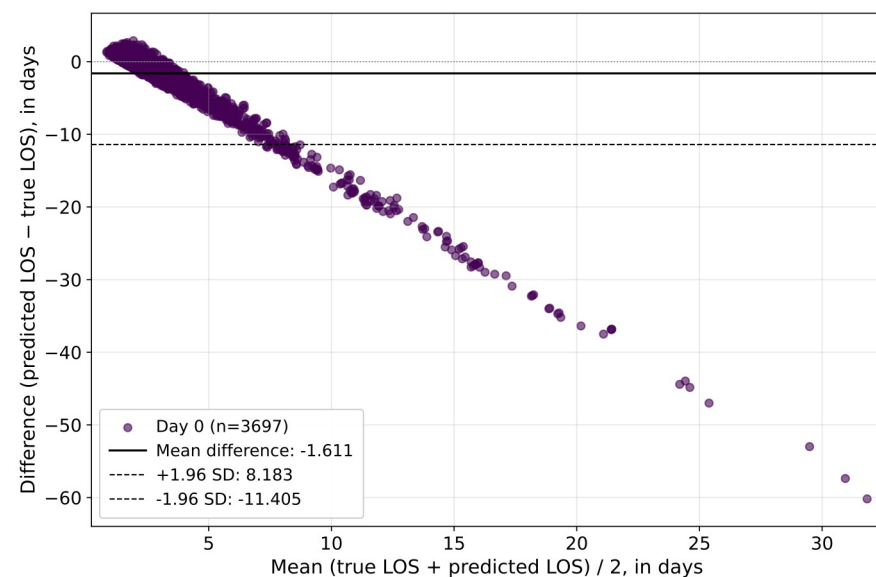
MAE improved vs baseline

Hybrid still best
HuBERT-ECG still worst

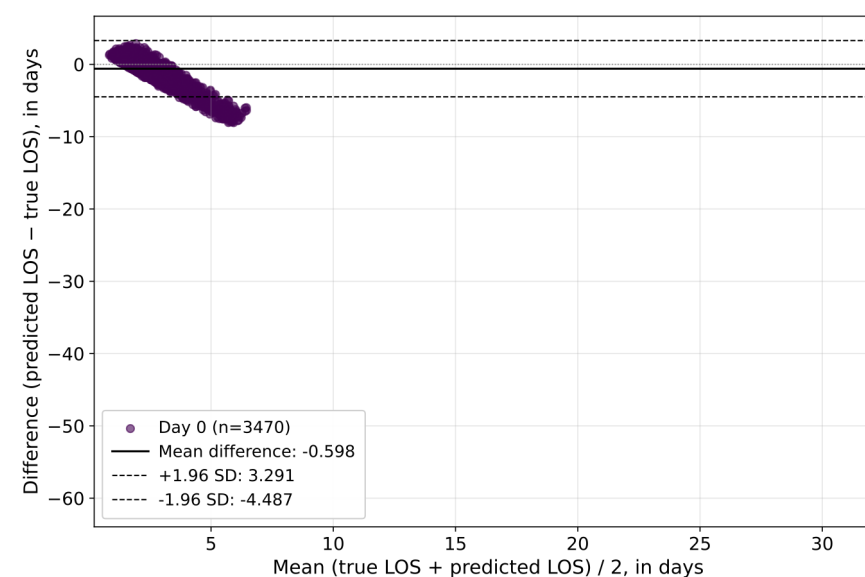
No clear benefit from complex
architectures



LOS Analysis with Bland Altman Plots



Full test cohort with n=3697



≤10 days cohort with n=3470

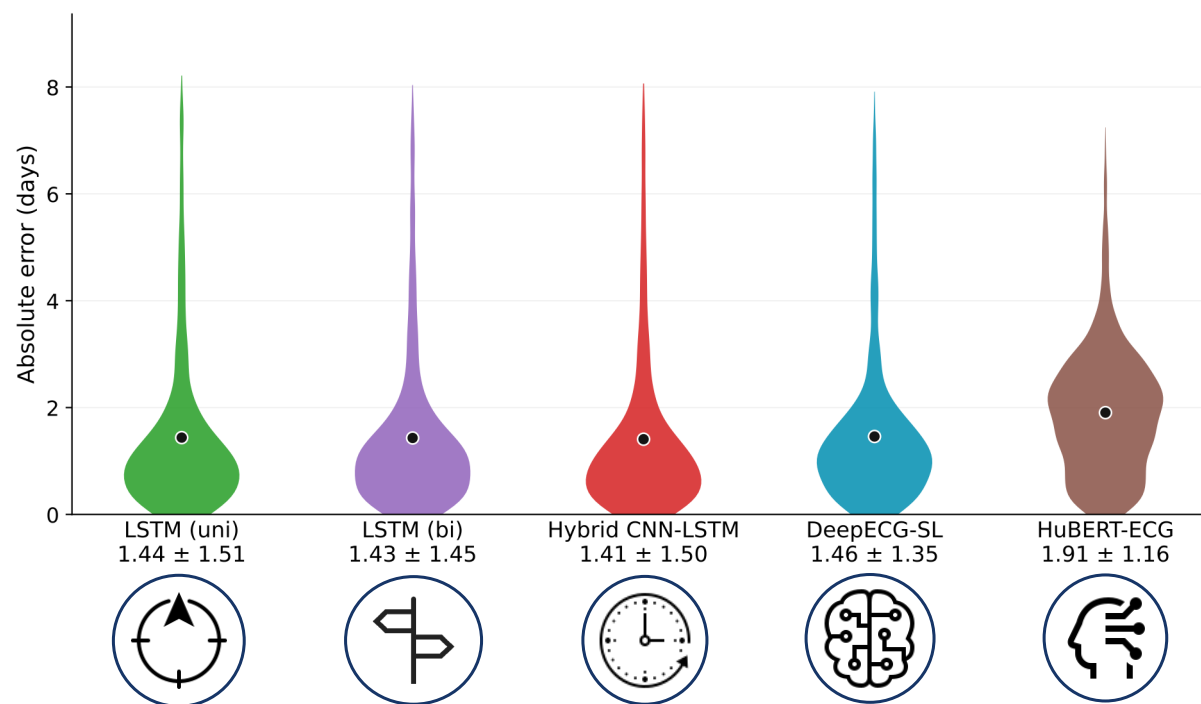
LOS Subgroup Analysis ≤ 10 days

Focus on 5 architectures

MAE drops by ~1d

STD drops significantly

Hybrid still best
HuBERT-ECG still worst



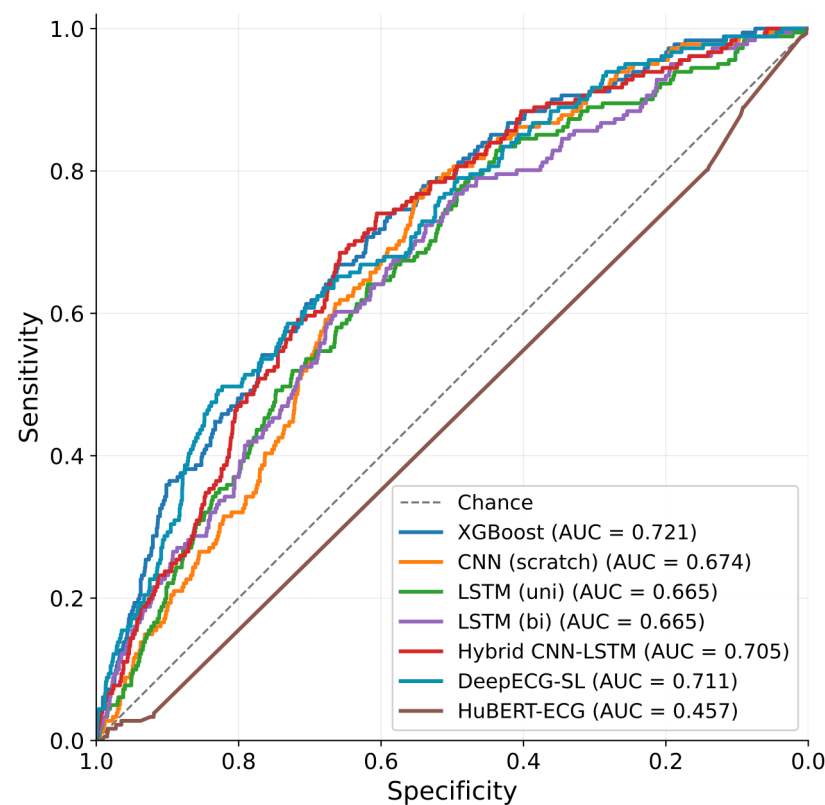
Mortality prediction

Mortality as binary classification

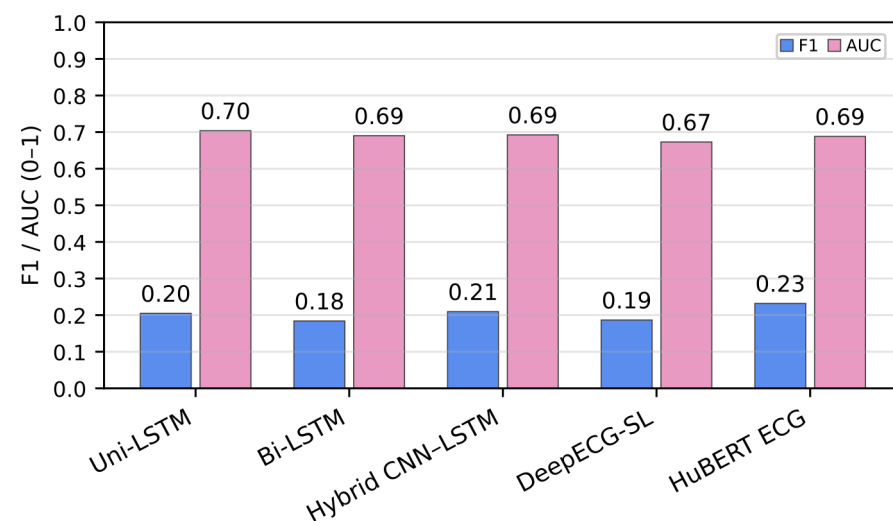
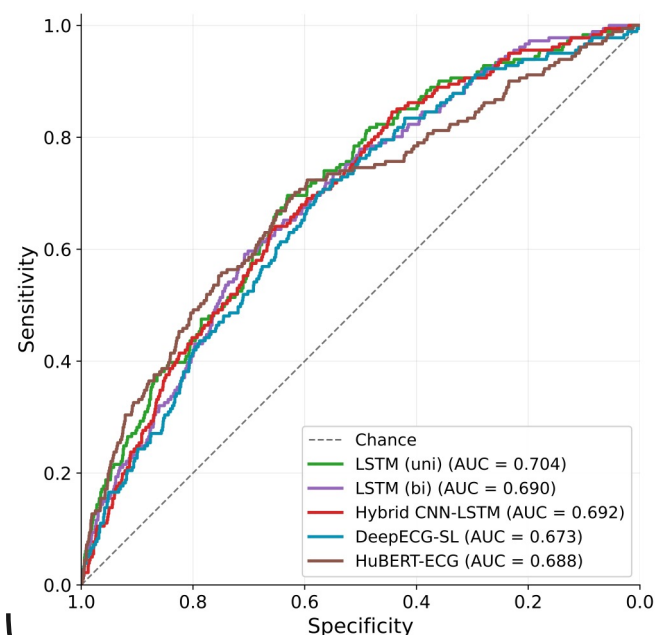
Architectures performed similar

Hubert-ECG performed worst

Default threshold $\tau = 0.5 \rightarrow F1 = 0$



Mortality prediction

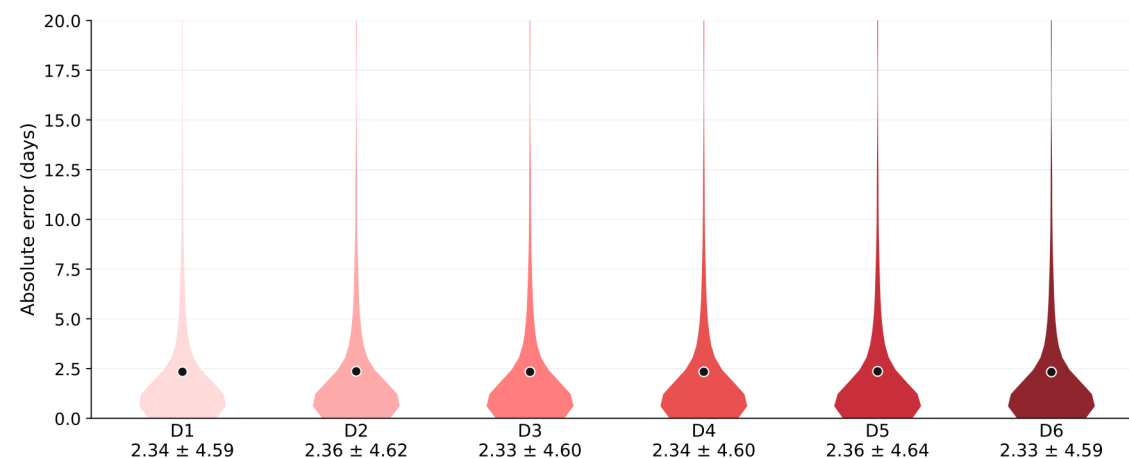


Decision threshold τ chosen to maximize validation F1

Multimodal Fusion – LOS

ID	Description
D1	ECG only
D2	Demographics (age + sex)
D3	ICU unit
D4	Demographics + ICU Unit
D5	EHR + Demographics + ICU Unit
D6	EHR subset + Demographics + ICU Unit

Implemented via late fusion

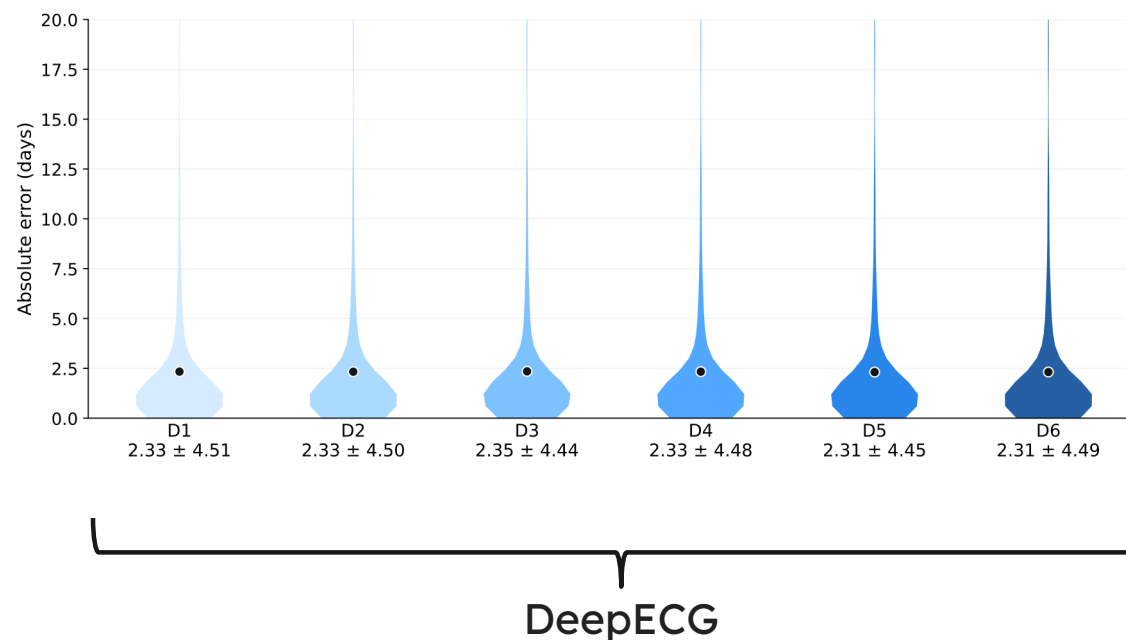


Hybrid CNN LSTM

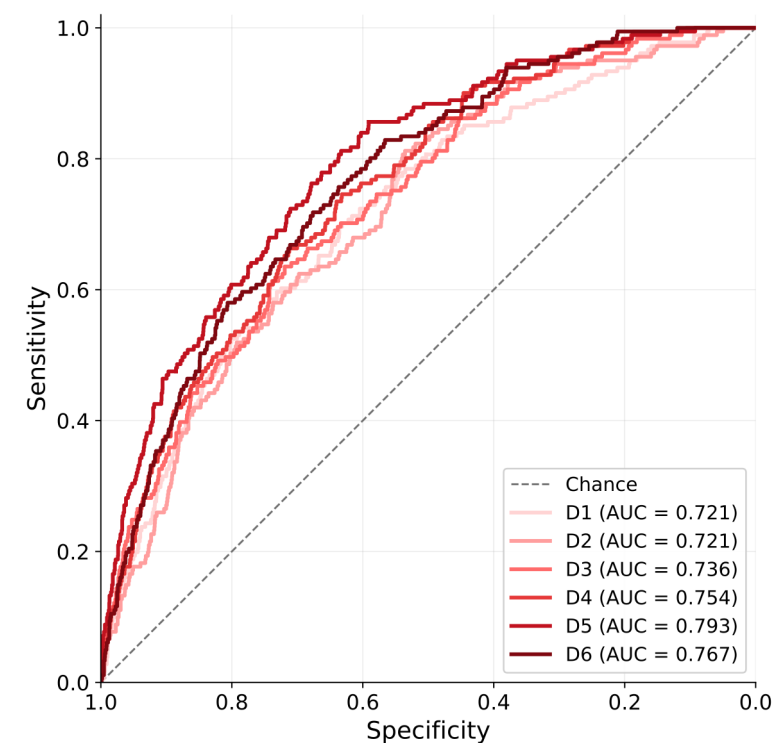
Multimodal Fusion – LOS

ID	Description
D1	ECG only
D2	Demographics (age + sex)
D3	ICU unit
D4	Demographics + ICU Unit
D5	EHR + Demographics + ICU Unit
D6	EHR subset + Demographics + ICU Unit

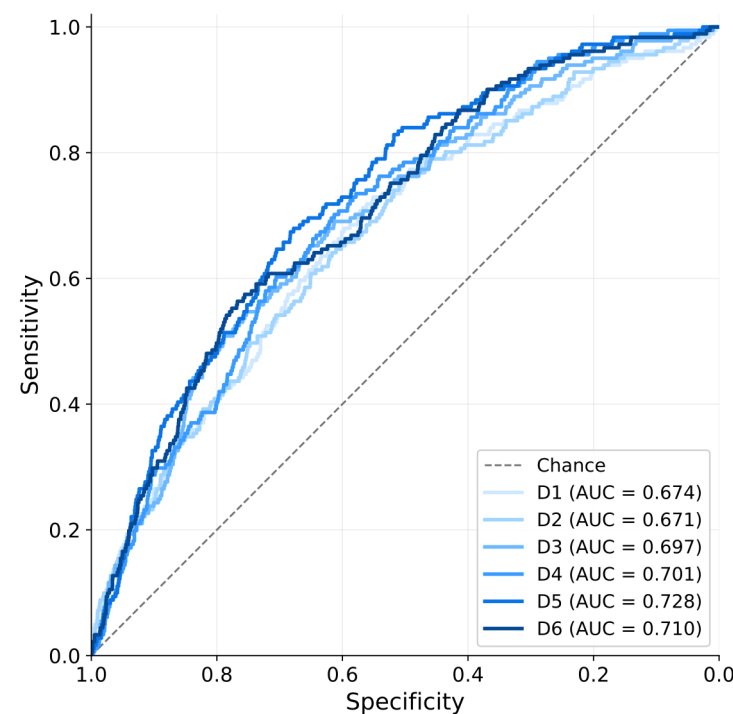
Implemented via late fusion



Multimodal Fusion – Mortality



Hybrid CNN LSTM
Best **D5** and **D6**

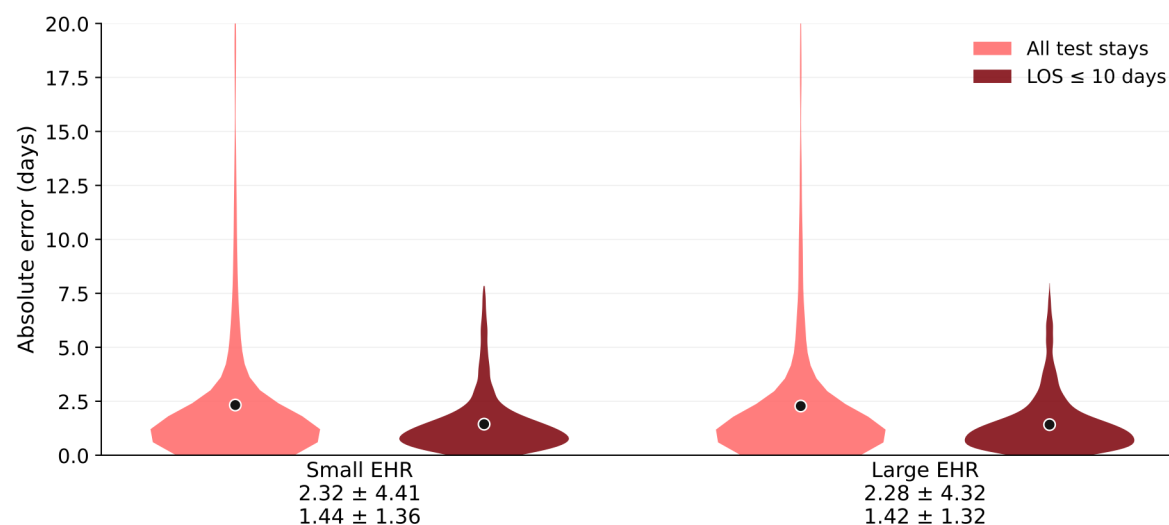


DeepECG
Best **D5** and **D6**

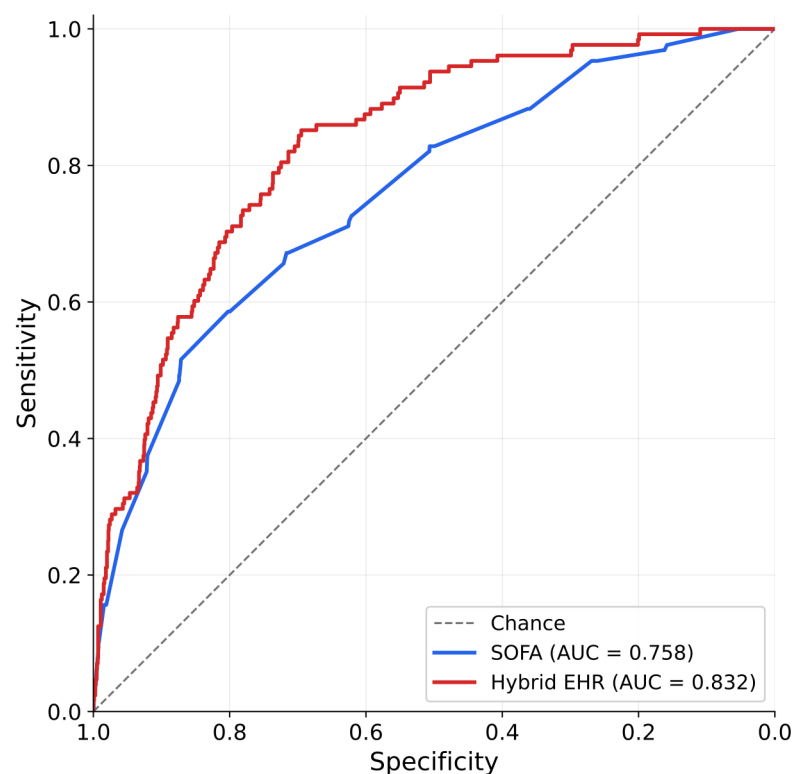
Hyperparameter Tuning with Optuna

Hyperparameter	Range
Learning rate	10^{-4} to 3×10^{-3}
Weight decay	10^{-6} to 10^{-3}
Dropout rate	0,25 to 0.5
LSTM dropout	0.2 to 0.5
LSTM hidden preset	{64,64}, {128,128}, {256,256}

- 2 optuna studies
- 5 hyperparameters tuned
- Limited server capacity
- Started 100 trails: 20/27 completed → timeout after 48 hours



SOFA Score vs Hybrid CNN-LSTM



Baseline: SOFA (0-24), leakage-free early window

Subset: $n = 1980$ with $\geq 4/7$ inputs available (72% of test set)

Fair comparison: evaluated on the same patient cohort

AUC: 0.832 (hybrid) vs 0.758 (SOFA)

Clinical trade-off: higher sensitivity (0.46 vs 0.35), ~equal specificity (0.91)

Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

Discussion

Main findings

LOS prediction

-  ECG signals provided measurable predictive value for LOS
-  Additional clinical context provided only marginal LOS improvements
-  Prediction accuracy highest for stays $\leq 10d$
-  Long-stay cases caused increasing underestimation

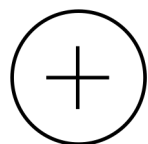
Mortality prediction

-  ECG captures mortality-related physiological information
-  Default threshold produced poor clinical classification
-  Threshold optimization improved F1 score
-  Threshold selection was critical despite good AUC

 Admission ECGs contain clinically relevant prognostic information and are most effective when combined with compact clinical context for mortality prediction.

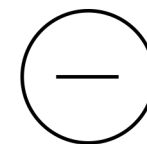
Discussion

Strengths & Limitations



Study strengths

- ★ Transparent experimental pipeline
- ★ Patient-level data splitting
- ★ Consistent preprocessing
- ★ Systematic comparison
- ★ Reproducible optimization workflow



Main limitations

- ⚠ Single-center cohort
- ⚠ Class imbalance
- ⚠ Single ECG snapshots
- ⚠ Limited compute budget

Discussion

Clinical implications

What the model **can** do

Early risk stratification
Support clinical decision making
Improve resource awareness



What the model **cannot** do

-  Autonomous triage
-  Replace clinical judgment
-  Generalize without external validation

ECG-based DL models can support ICU outcome prediction, but should be deployed as a decision-support tool with transparent risk estimates rather than as an autonomous clinical system.

Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

Outlook

Future Work



1. Model Development

- ❏ Alternative LOS representations
- ❏ Larger hyperparameter search under more compute budget
- ❏ Analysis of inputs and fusion strategies



2. Robust Validation

- ❏ Systematic handling of class imbalance
- ❏ Investigate regression-to-the-mean effects in LOS prediction
- ❏ External validation on an independent ICU-ECG cohort



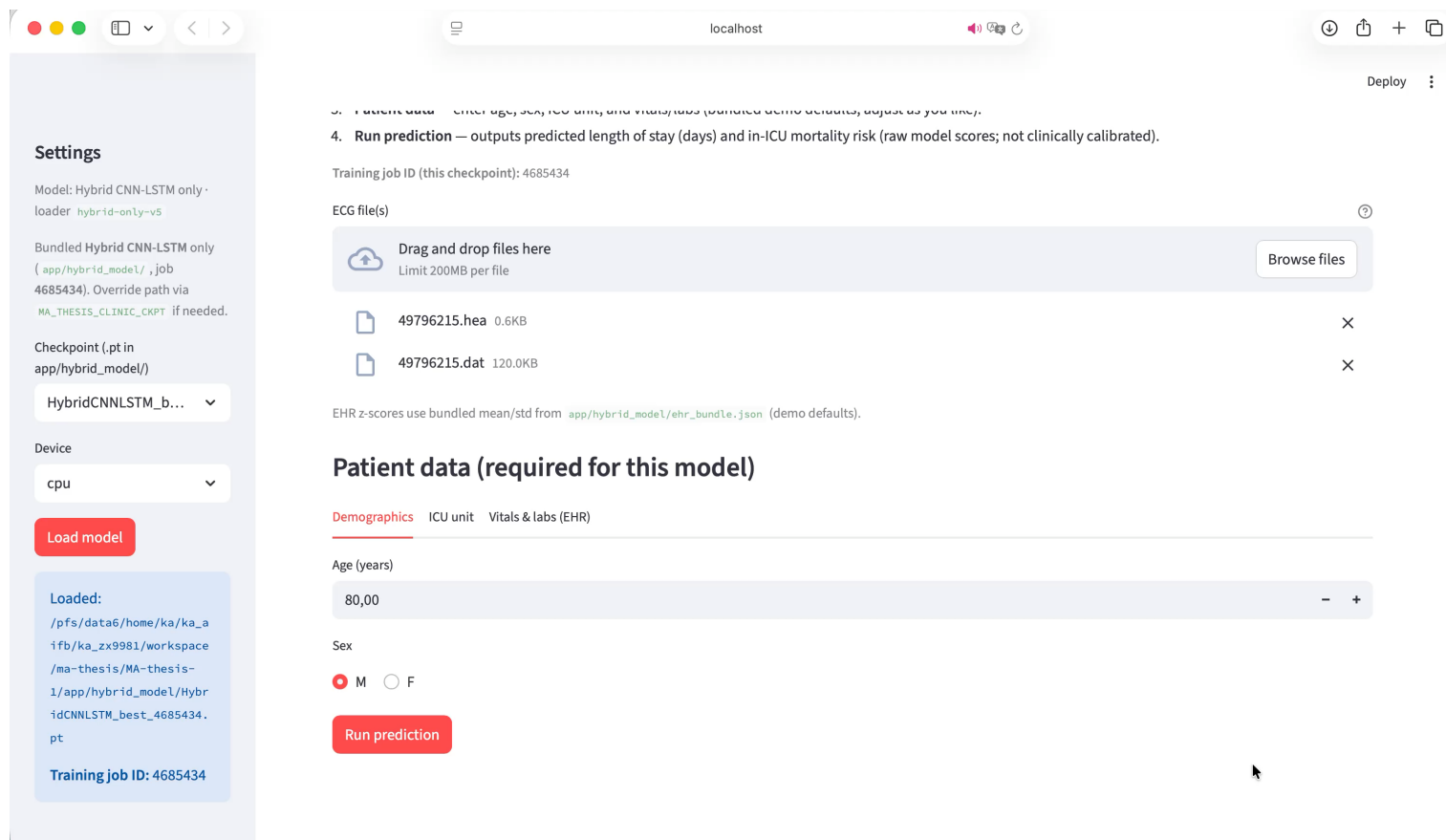
3. Clinical Deployment

- ❏ Human-centered design with ICU clinicians
- ❏ Evaluation of clinical utility and interpretability
- ❏ Validation of the prototype web interface



Goal: Integrate ICU risk prediction models into clinical decision support tools

Outlook



Settings

Model: Hybrid CNN-LSTM only · loader: hybrid-only-v5

Bundled Hybrid CNN-LSTM only (app/hybrid_model/ , job 4685434). Override path via MA_THESIS_CLINIC_CKPT if needed.

Checkpoint (.pt in app/hybrid_model/)

HybridCNNLSTM_b... ▾

Device

cpu ▾

Load model

Loaded:

```
/pfs/data6/home/ka/ka_a
ifb/ka_zx9981/workspace
/ma-thesis/MA-thesis-
1/app/hybrid_model/Hybr
idCNNLSTM_best_4685434.
pt
```

Training job ID: 4685434

3. **Train data** — enter age, sex, ICU unit, and vitals/labs (provided demo defaults; adjust as you like).

4. **Run prediction** — outputs predicted length of stay (days) and in-ICU mortality risk (raw model scores; not clinically calibrated).

Training job ID (this checkpoint): 4685434

ECG file(s)

Drag and drop files here
Limit 200MB per file

Browse files

49796215.he 0.6KB ×

49796215.dat 120.0KB ×

EHR z-scores use bundled mean/std from app/hybrid_model/ehr_bundle.json (demo defaults).

Patient data (required for this model)

Demographics ICU unit Vitals & labs (EHR)

Age (years)

80,00 - +

Sex

☒ M ☐ F

Run prediction

Agenda

	1	Motivation
	2	Fundamentals and State of the Art
	3	Data Understanding
	4	Methods and Results
	5	Discussion
	6	Outlook
	7	Conclusion

Conclusion

Research Question 1

Q1

What level of predictive performance and calibration do DL models achieve when predicting ICU LOS and mortality from the first raw ECG recorded at ICU admission?

Moderate

- LOS prediction reached MAE in low two-days range
- Mortality prediction achieved moderate AUC
- Performance strongest for common LOS ranges

Conclusion

Research Question 2

Q2

To what extent does integrating ECG with structured EHR and demographic data improve predictive performance compared to ECG-only models?

Partially

- Limited gains for LOS prediction
- Clearer benefits for mortality stratification
- Structured clinical context improved mortality prediction more than LOS prediction

Conclusion

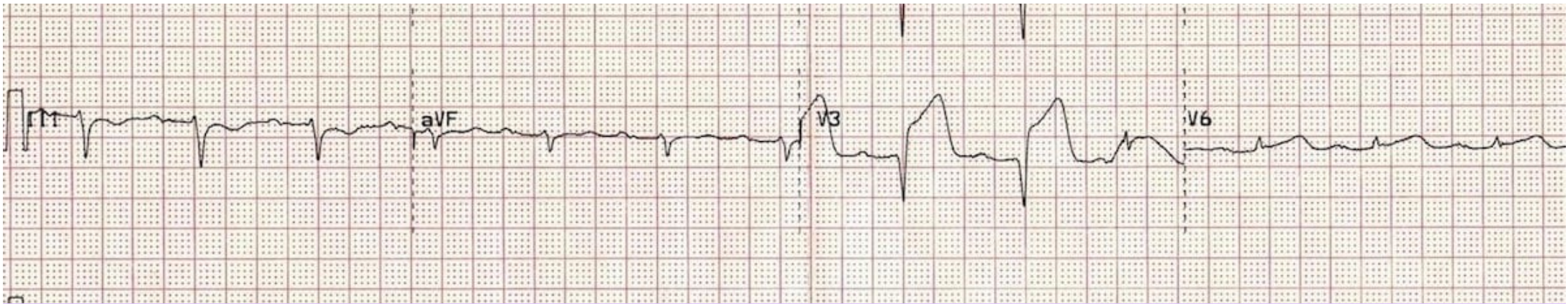
Research Question 3

Q3

Which ML architectures are most suitable for modeling ECG-based ICU outcome prediction in terms of accuracy and computational cost?

Hybrid CNN-LSTM

- Consistently competitive across tasks
- Best balance of performance and computational efficiency
- Selected as final architecture for the multimodal fusion strategy



Thank you! Questions?

