Friedrich-Alexander-Universität
Fachbereich Wirtschafts- und
Sozialwissenschaften | WiSo

Exploring Generalizability of Time Series Classification Models for Predicting Mortality in Intensive Care Units

Goals of the Meeting

What is this meeting about

Agenda:

- **Overview**
- **Preliminary Results**
- **Technical view**
- **Outlook**

Overview

Overview

Main Idea

Models

Beth Israel - Medical Center
Boston, Massachusetts, USA

University Hospital Carl Gustav Carus
Dresden, Saxony, Germany

Overview

Main Idea

Beth Israel - Medical Center
Boston, Massachusetts, USA

Models

University Hospital Carl Gustav Carus
Dresden, Saxony, Germany

In ICU Mortality after first 24h measurement
⇨ features measured as time series

LSTM

Performance on different population?
⇨ Insights to performance difference through XAI

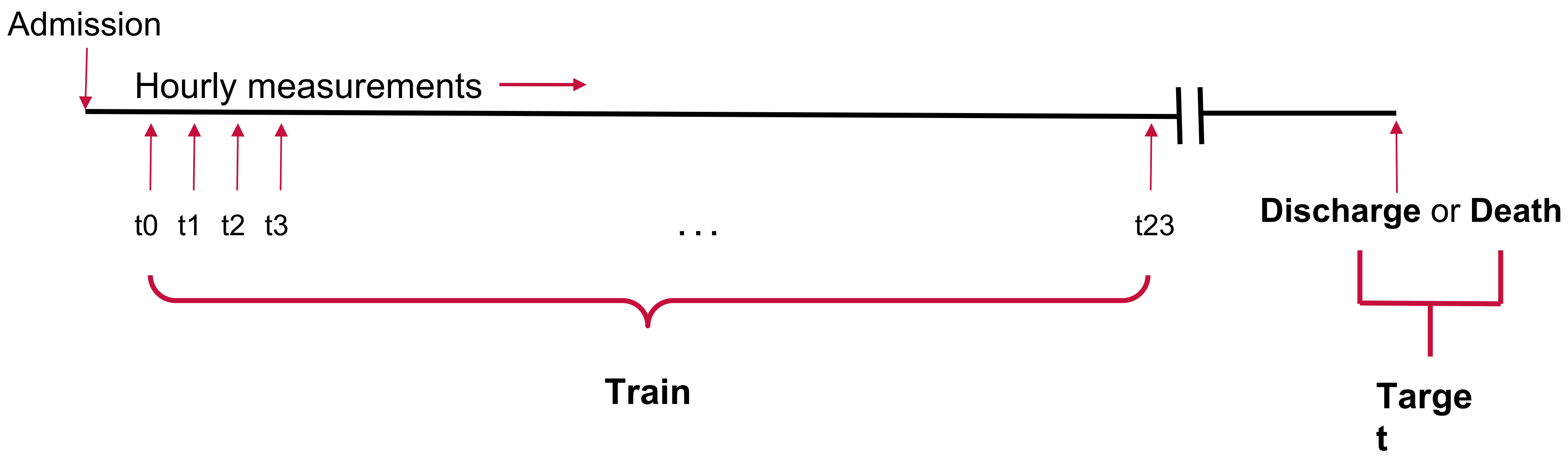
Overview

Why Is This Interesting?

- **Datasets are often centralized or commercialized**
  - smaller hospitals have to rely on external data/ models
- **High stake environment**
  - Trust & understanding of model behaviour is crucial
- **Generalization is not a given**
  - Past events show that population shift can lower performance
- **Time series may hold a high information density and are suitable for event detection**

Overview

Target



Overview
Experiments

Generalizability

1 Simple Transfer

Hospital A → Hospital B

Train | Test

2 Combined Transfer

Hospital A 50%
Hospital B 50%
→ Hospital B

Train | Test

3 Low Data Simulation

Hospital A | Hospital A | Hospital A | ⋮ → Hospital B

Train 1, Train 2, Train 3, …, Train m | Test

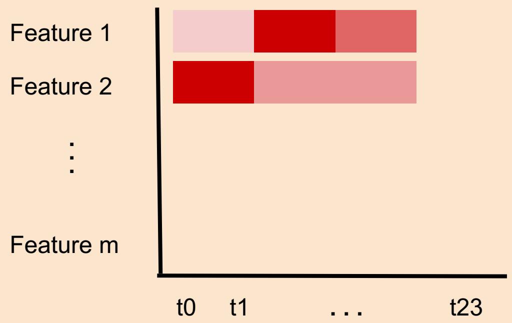
(4 Time Shift Simulation)

Hospital A t0 → Hospital A t1

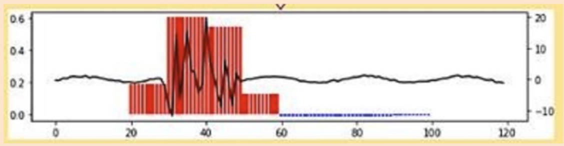
Train | Test

Interpretability

1 Time-wise Feature Importance Heatmap



2 Time- & Sample-wise Contribution



Nayebi Et al. 2023 https://www.sciencedirect.com/science/article/pii/S1532046423001594

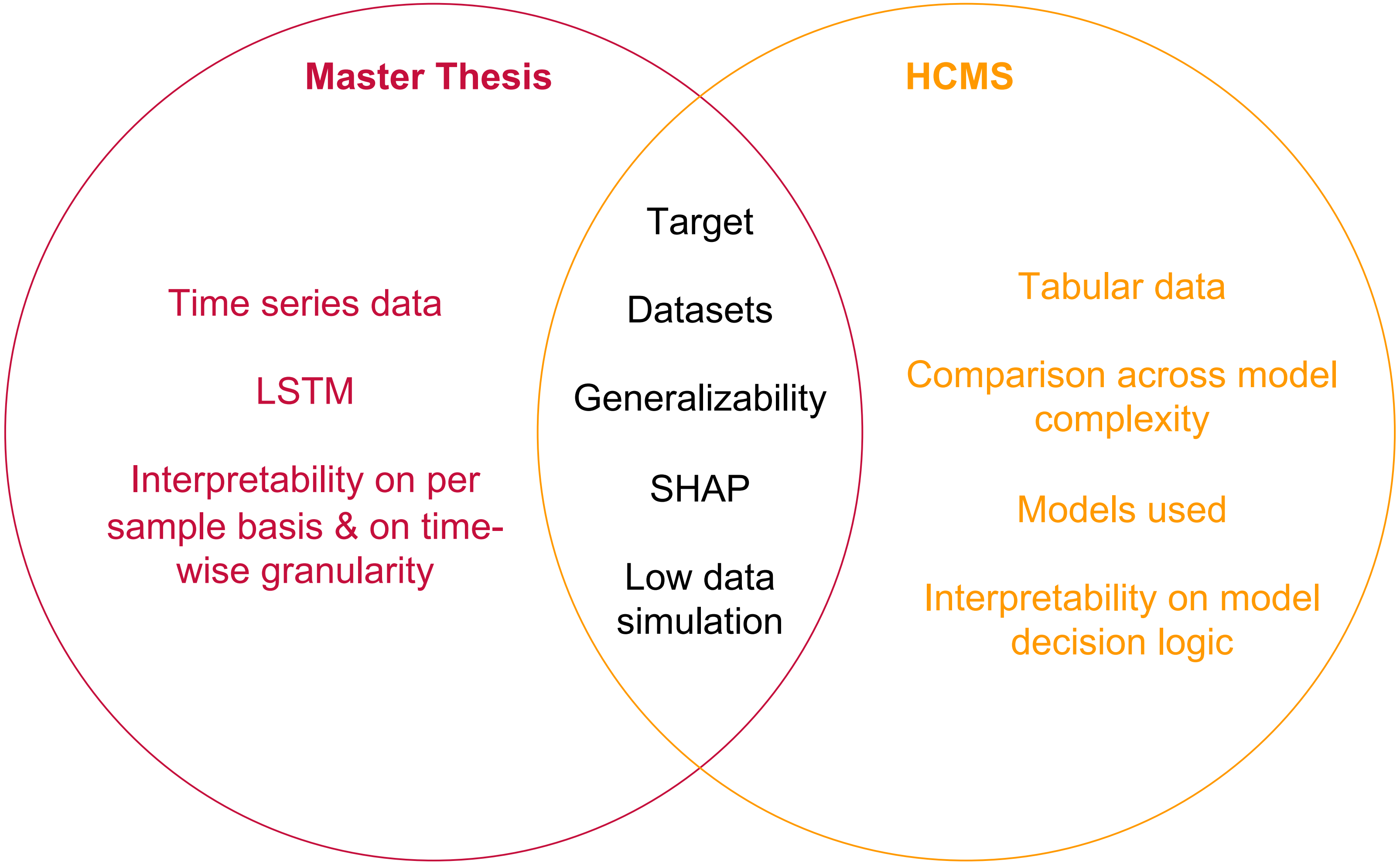
Overview

Main Contribution

- Exploring the unique challenges of applying time series classification models to ICU mortality prediction across different populations
- Visualizing the models behavior when transferred from one dataset to another in the in order to understanding the underlying behaviour of the model.
- Providing insights into how varying data availability impacts performance of time series classification models.

Overview

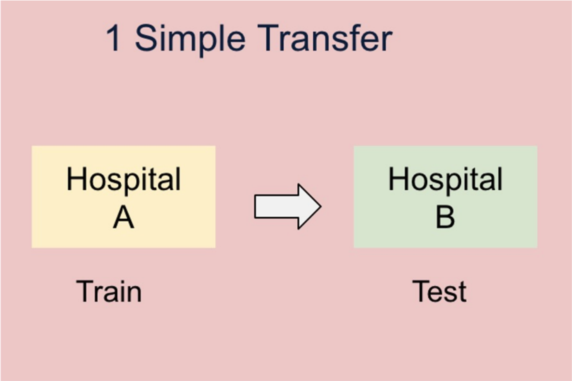
Differentiation and Similarities to HCMS Paper



Preliminary Results

Preliminary Results

Simple transfer



Train Set	Model	Test Set		
		TUDD (N = 1465)	MIMIC-IV (N = 14 558)	Combination (N = 18 346)
TUDD (N =)				
MIMIC-IV (N = 58 231)	S LSTM	66.31	87.72	
	MC LSTM	74.32	86.54	
Combination (N = 73 395)				

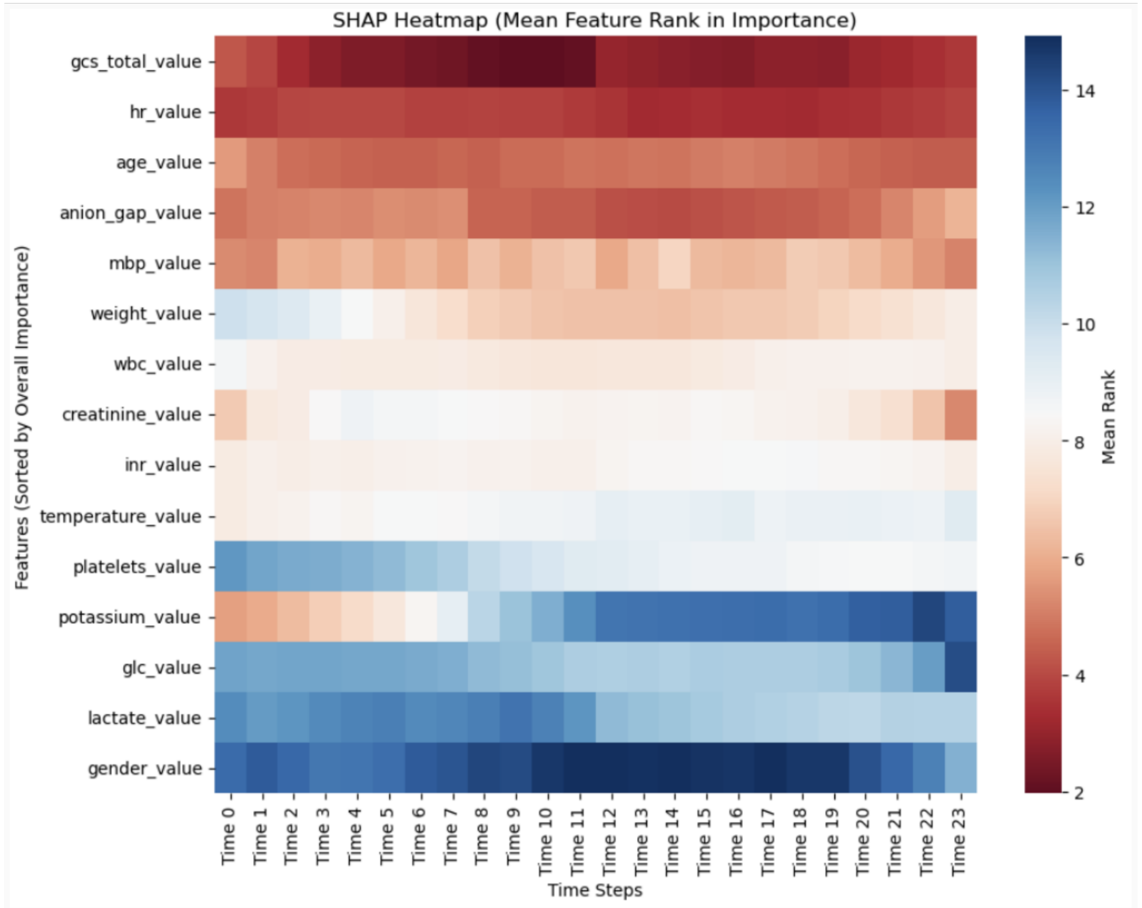
Table 2. caption

Preliminary Results

Feature importance heatmap with SHAP

- Idea: compare heatmap from different trained models
- Sampling of background data (10% of total test set)
  - background data excluded from test
- Concat per sample absolute SHAP values with **DeepExplainer** (1000 samples)
  - from explainer: (num_samples,time_steps,num_features)
  - so each (time_step, feature) cell has a SHAP value
- Downside: time steps are not independent due to memory states (sequential nature of the lstm)

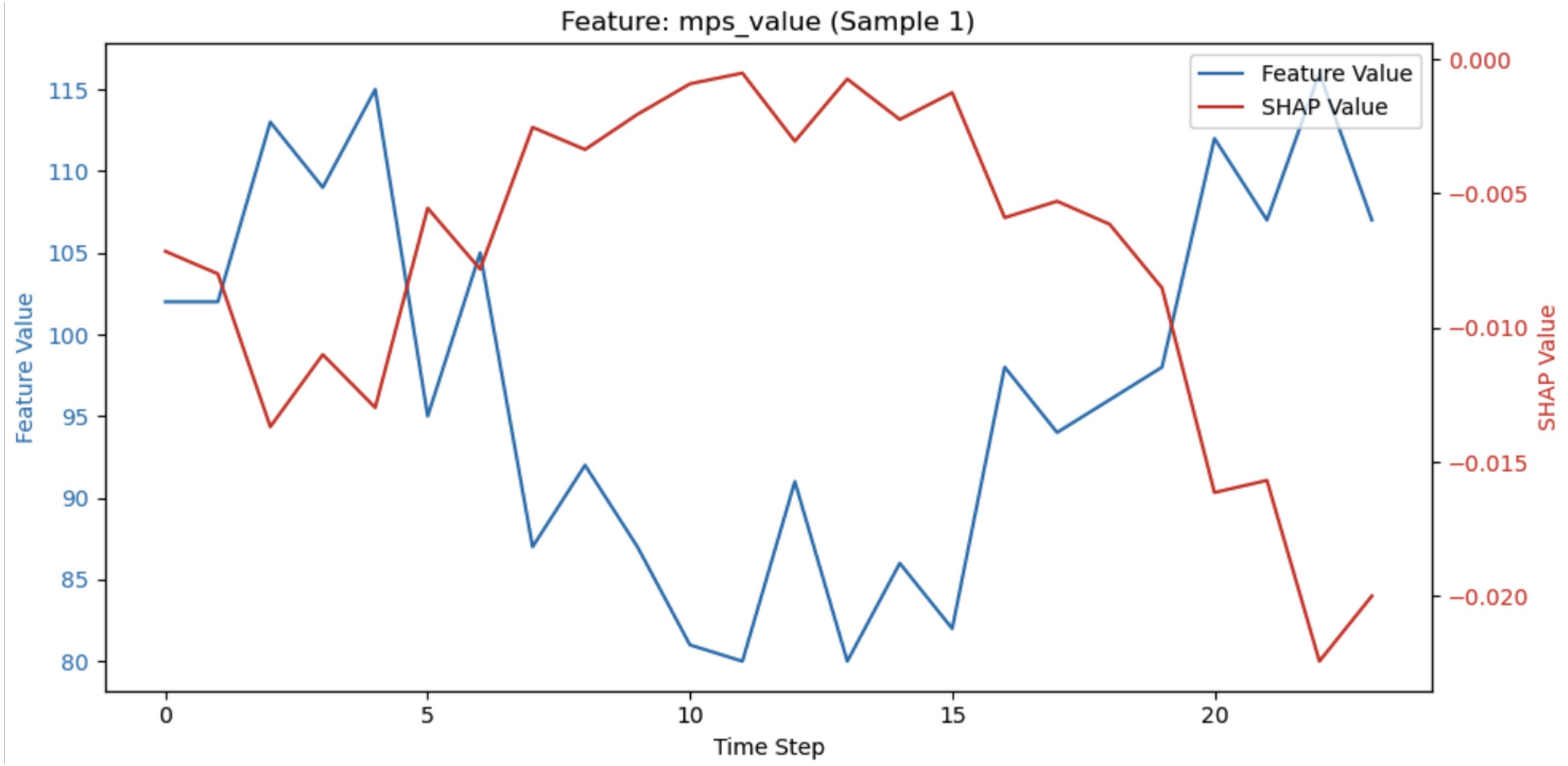
Idea from Thorsen-Meyer Et al. 2020
https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30018-2/fulltext



Preliminary results

Sample- & feature-wise contribution

Mean Blood Pressure



Technical View

Technical View

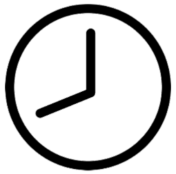
Data

- 15 features used
- At first usual pre-processing steps

Feature	MIMIC-IV			EUH		
	Mean	Std	Missing (%)	Mean	Std	Missing (%)
Sequential Features						
Temperature	36.84	0.48	3.60	36.76	0.75	6.82
Heart Rate	84.93	15.88	0.13	83.40	16.81	0.27
Glucose	136.12	49.12	5.26	138.39	35.54	0.27
Mean Blood Pressure	78.67	11.48	0.55	83.80	12.20	17.16
Glasgow Coma Scale Total	12.52	3.26	0.91	11.56	4.52	56.31
Potassium	4.18	0.55	4.51	4.13	0.42	0.37
Leukocytes	11.91	7.70	5.58	11.50	5.93	4.90
Thrombocytes	207.29	105.80	5.13	223.84	110.03	4.83
Prothrombin Time	1.44	0.61	20.81	1.37	0.35	5.16
Anion Gap	14.08	3.30	5.39	6.16	2.91	71.97
Lactate	2.24	1.81	46.83	1.48	1.57	0.25
Creatinine	1.42	1.51	4.49	1.07	0.99	5.78
Static Features						
Gender (Female %)	44.22	–	0.00	40.24	–	0.00
Age	63.16	16.79	0.00	64.39	15.96	0.00
Weight	82.00	23.60	3.50	77.88	19.55	37.53
Mortality (%)	7.43	–	0.00	5.72	–	0.00

Technical View

Sequencing - create common table



Extract 24h after first measurement

stay_id: xxxx

Time	Feature1	Feature2
13:45	-	3
14:10	3	-
14:20	1	5
14:35	-	9
14:50	4	-

Aggregate to full hour

stay_id: xxxx

Time	Feature1	Feature2
14:00	2	4
15:00	4	9

⋮ 24 rows

+ concatenate

stay_id: yyyy

Time	Feature1	Feature2
19:00	25	43
20:00	155	56

⋮ 24 rows

+ concatenate

stay_id: zzzz

Time	Feature1	Feature2
9:00	45	65
10:00	99	76

⋮ 24 rows

stay ID	Time Index
xxxx	0
xxxx	1
…	…
xxxx	23
yyyy	0

⋮

Merge

stay ID	Time Index	Feature1	Feature2
xxxx	0	2	4
xxxx	1	4	9
…	…		
xxxx	23	5	4
yyyy	0	25	56

⋮

Technical View

Sequencing - create input sequences

stay ID	Time Index	Feature1	Feature2
xxxx	0	2	4
xxxx	1	4	9
….	….		
xxxx	23	5	4
yyyy	0	25	56
….	….	….	….

Grouped by stay_id, time index

stay ID	Label
xxxx	1
yyyy	0
….	….

2	4
4	9
….	….
5	4

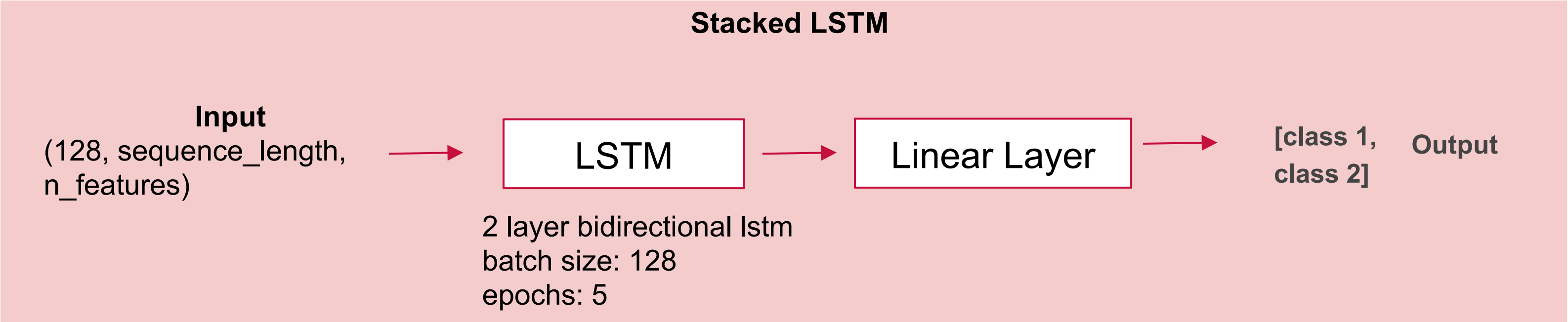
1

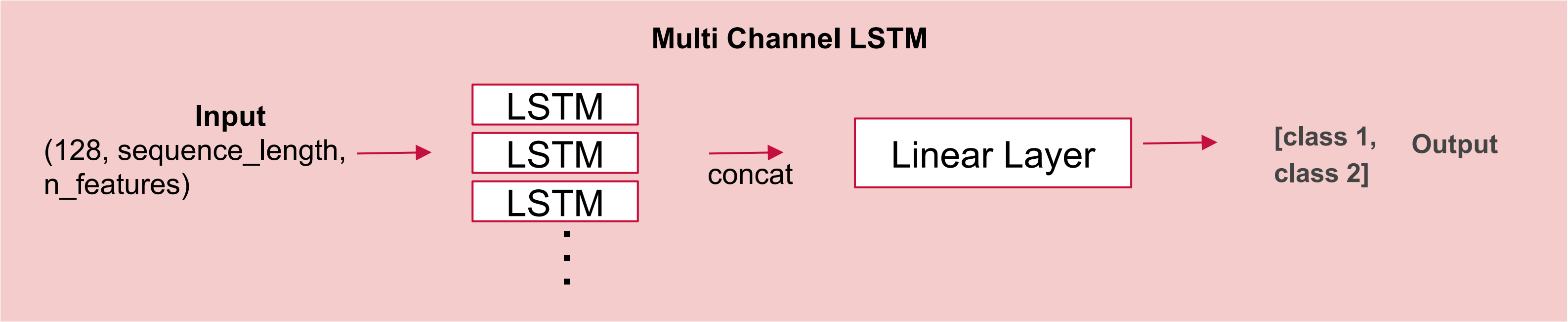
sequences = [[(24, 15), label], …]

Technical View

Model Architecture

Hyperparameters
hidden state: 100
dropout: 0.75
optimizer: Adam
lr: 1e-4
loss: cross entropy
class weights: 3.0 on non survival





batch size: 128
epochs: 5

Outlook

Outlook

- Comparison of models trained on MIMIC and Dresden dataset
- Test for SHAP variability
- Add attention layer to lstm and visualize attention weights
- Other models?