Robustness analysis of recurrent networks for clinical prediction tasks

Marton **Csutora** • BME

1. Abstract

Robust validation of deep learning models remains critical in healthcare, as their promising research performance often degrades in deployment. Using the MIMIC-IV database [14-16], this study investigates the robustness of three extended long short-term memory (xLSTM, [8]) models trained to predict sepsis, myocardial infarction, and vancomycin administration. Through analysis of scaling behavior and validation patterns, a reliable framework is established for estimating real-world model performance. The findings provide crucial insights into dataset characteristics and model reliability, offering guidance for future clinical deep learning research where accurate performance estimation is essential for patient safety.

2. Ethical considerations

Clinical deployment of these models requires careful consideration of **algorithmic fairness**, **liability** frameworks, and **patient consent** protocols, particularly given the life-critical nature of ICU predictions.

3. Methods

1 Raw time-series patient health data collected from MIMIC-IV is transformed into standard data for deep learning through:

- **pivoting** to regularize the spatial format
- **daily aggregation** to regularize the temporal format
- **outlier handling** through robust aggregation functions (median, IQR, minimum, maximum), as outliers can hold clinical significance in patient data
- **KNN imputation of missing data** chosen for its robustness and ability to preserve the data distribution [25]
- patient data is processed into overlapping **14 day sequences** to allow models to learn as much as possible from the data

Patient ID	Timestamp	Measurement	Value
00000001	2024.01.02 12:31	Heart rate	95.0
00000001	2024.01.02 12:31	Sodium	133.2
00000002	2024.01.03 08:10	Bands	2.4

Fig 1: MIMIC-IV patient data in a narrow format [author's]

Patient ID	Day	Heart rate	Sodium	Bands
00000001	2024.01.02	95.0	133.2	2.2
00000002	2024.01.03	83.0	127.5	2.4

Fig 2: Preprocessed dataset in a wide format [author's]

2 The xLSTM architecture is trained and optimized using the prepared dataset to predict vancomycin administration, myocardial infarction, and sepsis. Of the two model variants introduced by xLSTM, sLSTM is used for this study. The architecture is as seen below:

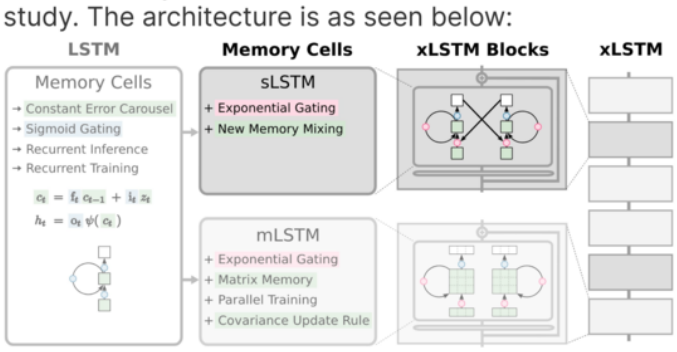


Fig 3: Architectural improvements of the two variants of xLSTM over the original LSTM. [8]

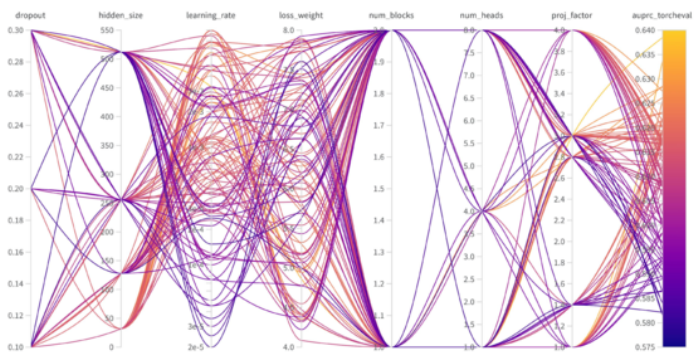
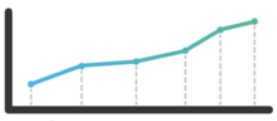


Fig 4: Different hyperparameter configurations found through the optimization process [author's]

3 The optimized models are analyzed through a rigorous validation process involving scaling law investigation, k-fold cross-validation, and prospective validation. This allows us to understand what the main performance bottlenecks are, how robustly the models learn patient health patterns, and to see how good actual deployment performance is likely to be.



4 Scaling law investigation is a way to assess how model performance changes based on, among others, the data amount, or model size. This study focuses on data scaling, as model size did not appear to significantly affect performance. Six different data configurations are assessed, with the data amount ranging between 10% and 100%.

5 K-fold cross-validation replaces the standard train-test split with k trials. In each trial, a different subset of the data is used as the test set, and the rest is used for training. In this study, I employed 5-fold cross-validation. The variations between the trials provide insights into e.g. training stability, while the average across all of them gives us more reliable performance values.

6 Prospective validation involves testing model performance on a previously unseen dataset that has been collected later than the training data. Most commonly, this would be done using a dataset produced much later, but here the training data was already very new, so I opted to use the latest 10% of the dataset instead.

4. Results

1 The data scaling shows interesting results: Performance, as expected, generally **increases with data amount**. The sepsis results are quite unusual though, which I mainly attribute to two key issues:

- MIMIC-IV does not have time-series diagnoses, so **targets** had to be **manually constructed**
- The sepsis dataset is **heavily imbalanced**, with the positive sample ratio being less than 2%

These factors reduce the data quality heavily, which makes it very difficult for the model to learn from the data. (see Fig 5 on the left)

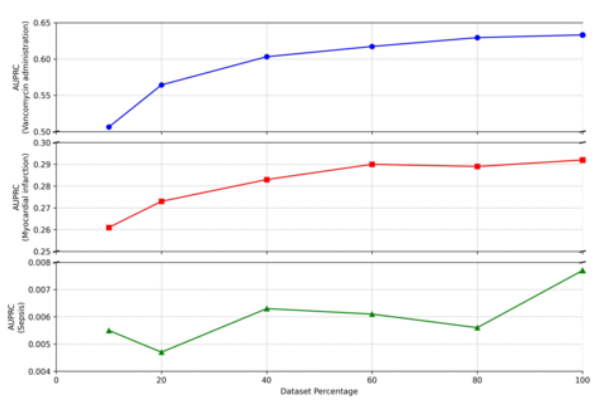


Fig 5: Data scaling results averaged across the k trials (vancomycin administration in blue, myocardial infarction in red, sepsis in green) [author's]

2 The k-fold cross-validation trials show that in addition to performance, **training stability improves with more data** as well. The odd patterns with sepsis can be seen here too, the model training appears to be very unstable due to the low data quality and quantity. This can be seen on the figures 6-8 below, each showing the training curves of the k folds for 20%, 60%, and 100% of the data for the three different targets. Related detailed tables are available in my thesis.

3 During prospective validation, the models' performance only dropped by at most about 3%, suggesting **good generalizability** and **stable real world performance**. Detailed tables can be found in my thesis.

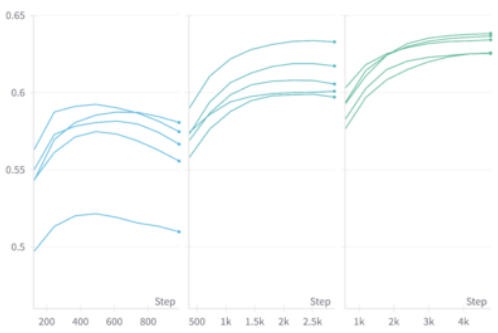


Fig 6: Vancomycin adminstration k-fold trials across 20% (left), 60% (middle), and 100% (right) of the dataset [author's]

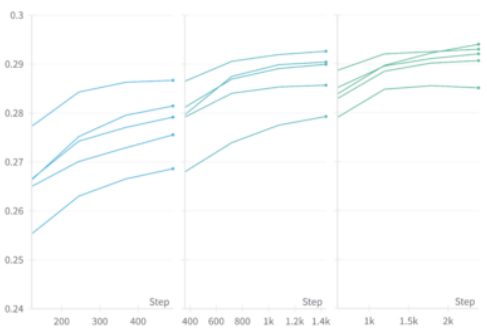


Fig 7: Myocardial infarction k-fold trials across 20% (left), 60% (middle), and 100% (right) of the dataset [author's]

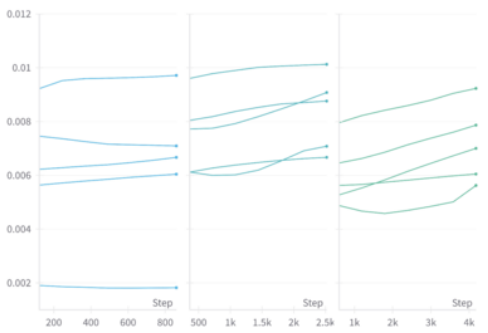


Fig 8: Sepsis k-fold trials across 20% (left), 60% (middle), and 100% (right) of the dataset [author's]

5. Conclusions

- The **validation framework** I developed shows **promising** results for further use
- The main **bottleneck** seems to be **data quality** and quantity, the former of which could be improved by adding time-series diagnoses to clinical datasets
- Performance estimation must be target-specific, as different clinical predictions showed **distinct scaling behaviors and generalization patterns**

6. Future work

- Developing more **refined temporal validation** techniques
- Adding **time-series diagnoses to clinical datasets** to remove the need of constructing less reliable targets after the fact
- Implementing **automatic data collection** in hospitals to reduce the human error factor in data quality

7. References

References and detailed acknowledgements are available in my thesis, which is reachable via the QR code on the right:

The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant №CEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organised under the Erasmus + Programme of the European Union.