Predicting the trajectory of intracranial pressure in patients with traumatic brain injury: evaluation of a foundation model for time series

Florian D. van Leeuwen[1][0009−0009−7092−2848], Shubhayu Bhattacharyay[2][0000−0001−7428−5588], Alex Carriero[3][0009−0007−4499−8043], Ethan Jacob Moyer[4], and Richard Moberg[4]

[1] Department of Methods and Statistics, Faculty of Social Science, Utrecht University, Utrecht, The Netherlands
`f.d.vanleeuwen@uu.nl`
[2] Harvard Medical School, 25 Shattuck St, Boston, MA 02115, USA
[3] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[4] Moberg Analytics, Inc, Philadelphia, PA, USA

Abstract. Patients with traumatic brain injury (TBI) often experience pathological increases in intracranial pressure (ICP), leading to intracranial hypertension—a common and serious complication. Early warning of an impending rise in ICP could potentially improve patient outcomes by enabling preemptive clinical intervention. However, the limited availability of patient data poses a challenge in developing reliable prediction models. In this study, we aim to determine whether foundation models, which leverage transfer learning, may offer a promising solution. We compare a foundation model for time series (MOMENT) with established time-series architectures, including long short-term memory (LSTM) and exponential smoothing (ES), for forecasting ICP. We found that for forecasting ICP with a 30-minute horizon, the MOMENT and LSTM models yielded similar results and both outperformed the ES model. However, all models performed poorly in predicting volatile periods, and there was substantial variability in model performance between patients.

Keywords: Foundation model · Intracranial pressure · Time series prediction · Transfer learning · Traumatic brain injury

1 Introduction

Traumatic brain injury (TBI) afflicts millions of people annually. In Europe, it is estimated that each year 1 million individuals are admitted to the hospital, and 75,000 die as a result of TBI [1]. TBI is a highly complex condition due to its heterogeneous nature: "TBI is considered the most complex disease in our most complex organ. It is characterized by great heterogeneity in terms of etiology, mechanisms, pathology, severity, and treatment, with widely varying outcomes" [1, p. 68].

For moderate-to-severe TBI patients,[5] A common complication is traumatic intracranial hypertension (tIH), which negatively affects patient outcomes [4]. tIH occurs when the volume within the skull rapidly increases beyond the skull's capacity [5], causing a spike in intracranial pressure (ICP). By monitoring ICP, potential tIH episodes can be detected. Clinicians may then decide whether to intervene, potentially exposing the patient to riskier treatments, or to tolerate the rise in ICP. While detecting a tIH event might prompt ad-hoc intervention, it remains unclear whether such interventions are beneficial after a damaging tIH episode [6,7]. To provide the best care (or at least improved care), an early warning of a pending tIH episode would be beneficial for clinicians, potentially allowing for timely prevention of the episode [5].

Such a warning can potentially be achieved through clinical prediction modeling. Many different models for ICP forecasting (e.g., [8,9,10]) and tIH prediction (e.g., [11,12]) have been proposed. However, the process is not straightforward. In developing these models, decisions must be made regarding the features (variables used as inputs in the model), the length of the input sequence (how long to monitor a patient before making a prediction), the forecast horizon (how far into the future the forecast extends), and, in the case of tIH prediction, how to define a tIH event. Furthermore, tough choices regarding the model development (e.g., hyper parameters) need to be made with limited data, as the sample size is usually low. There is significant variation in these choices; for a comprehensive overview, refer to [5]. Although predicting a tIH event may be easier than forecasting the ICP signal, it might not provide enough clinical information to guide decision-making.

In this study, we focus on forecasting the ICP signal. The raw ICP signal is very noisy and requires preprocessing before it can be used in a model. Often, parts of the signal are not recorded (typically at the start or the end), there may be frequent interruptions or artifacts due to clinical management, or unrealistic values are measured. Approaches such as denoising and smoothing have been applied to ICP signals before the data is fed into a model for making predictions [8].

There is an abundance of methodologies used to model time series, which are applied in many domains [13]. Widely adopted algorithms, such as Exponential Smoothing (ES) [14] and ARIMA [15], have been used for decades. There are also newer neural network-based sequence models, such as RNNs [16,17] and transformers [18]. It is not always clear which model performs best in the healthcare domain [19].

More recently, a foundation model for time series was published [16,17]. A foundation model is a model trained on a broad dataset (usually through self-supervised learning) that can be fine-tuned for a wide range of tasks [20]. These models contain prior information about the task and/or domain for which they

---

[5] The initial clinical severity of TBI is often assessed using the Glasgow Coma Scale (GCS), which is commonly trichotomized into three classes of severity: mild (GCS 13-15), moderate (GCS 9-12), and severe (GCS $\leq$ 8) [2], although the usefulness of this trichotomy is a point of discussion [3].

are designed. Foundation models have pre-trained weights that are adapted to new tasks using transfer learning [21]. While transfer learning enables foundation models, it is the scale of these models (i.e., the number of parameters) that enables their true power [20]. The proposed MOMENT model is a 385-million-parameter model trained on 13 million unique time series from different domains (e.g., medical, energy, nature)[6] using self-supervised learning and tested on several different tasks.[7] In healthcare, where data is often limited, a foundation model may help improve predictions [22]. In this study, we specifically investigate whether foundation models, and thus transfer learning, are helpful for ICP forecasting.

The main contributions of this study are as follows:

– Outlining the preprocessing procedure for ICP signals,
– Comparing the performance of a foundation model (MOMENT) to established time-series model architectures that do not employ transfer learning (ES and LSTM) for ICP forecasting,
– Inspecting the performance of the models per patient and per forecast made for each patient.

This paper is structured as follows: Section 2 outlines the data and methodology used in the analysis. Section 3 presents the results, and Section 4 provides the discussion.

2   Methods

2.1   Data

The training and internal validation data were from the high-resolution multimodal dataset from TRACK-TBI [8]. TRACK-TBI was a prospective, multi-center observational study conducted at 18 Level 1 trauma centers in the US, enrolling patients with TBI between February 26, 2014, and August 8, 2018. The study was approved by the institutional review board of each TRACK-TBI site. Participants, or their legally authorized representatives, provided written informed consent to participate. The data were accessed through the Moberg AI Ecosystem (Moberg Analytics, Inc, Philadelphia, PA, USA). The data was obtained data through the FITBIR (Federal Interagency Traumatic Brain Injury Research) Informatics System under a data use agreement.

We have ICP data for multiple patients, collected from three different sites. Individual patients can have multiple recordings. We only used recordings with a duration of two hours or more, to ensure sufficient data for training and evaluation. We started with a total of 39 patients and 94 recordings. After removing

---

[6] The datasets can be found here: https://huggingface.co/datasets/AutonLab/Timeseries-PILE

[7] Long-horizon forecasting, short-horizon forecasting, classification, anomaly detection, imputation.

[8] More information can be found here: https://tracktbi.ucsf.edu

unrealistic signals (see Appendix B for details), we ended up with 32 patients and a total of 83 recordings, comprising 5,142 hours of data. The distribution of recording times is shown in Figure 9 (Appendix A). We only used data from patients whose ICP was measured with an intraparenchymal fiberoptic monitor (within the brain tissue). Data from patients with an external ventricular drain were not used, as their measurements are affected by interventional draining of cerebrospinal fluid.

For external validation purposes, we use the CHARIS database [23], which is publicly available[9] on PhysioNet [24]. The ICP data were monitored using either a subarachnoid bolt or ventriculostomy in 13 patients diagnosed with TBI. All patients were from the same hospital and had only one recording session. The data were preprocessed as described in Appendix B. To match the data used in [8], parts of the recordings were removed based on the figures of the signals shown in [8]. For 10 of the 13 signals, the last part (10-120 hours) was removed. After preprocessing, there were a total of 1,122 hours of recordings.

ICP is measured in mm Hg; normal values for a person in a supine position range between 0.9 and 16.3 mm Hg [25]. The Brain Trauma Foundation guidelines set the threshold for a tIH event at 22 mm Hg, though this is contested, and other values (e.g., 20 mm Hg) are sometimes used [4,5].

2.2   Preprocessing

ICP data are not "clean" (ready for model development); there is a significant amount of uncertainty (noise) in the signal. Some of this uncertainty arises from measurement error and missing observations [26]. We want the algorithms to learn the noiseless ICP signal. To assist in this, we use the preprocessing step explained in Figure 10 (Appendix B). The output of this preprocessing is a segment where each data point represents the ICP (mm Hg) value for a second. Figure 1 provides an example of the transformation before and after preprocessing. The preprocessing and down-sampling appear to be effective.

Many of the signals have an unrealistic ending (Figure 13). It seems that the measuring device was not functioning correctly during the last part of some recordings, possibly due to the premature disconnection of the sensor. There are also signals for which the preprocessing did not eliminate the noise well (Figure 12); these were removed (11 out of 94).

After preprocessing, each minute has one ICP observation. We believe that a forecasting horizon of 30 minutes (similar to [9,27,11]) provides clinicians with enough information to assist in the decision-making process. The input length is set to 60 minutes, meaning we use the previous 60 minutes of ICP data to forecast the next 30 minutes. To train and validate the model, we segment the data by cutting the signal into 90-minute sections (60 for X and 30 for Y). This process is further explained in Figure 11 (Appendix B).

Both datasets were scaled by subtracting the mean and dividing by the standard deviation of all signals in the training set.

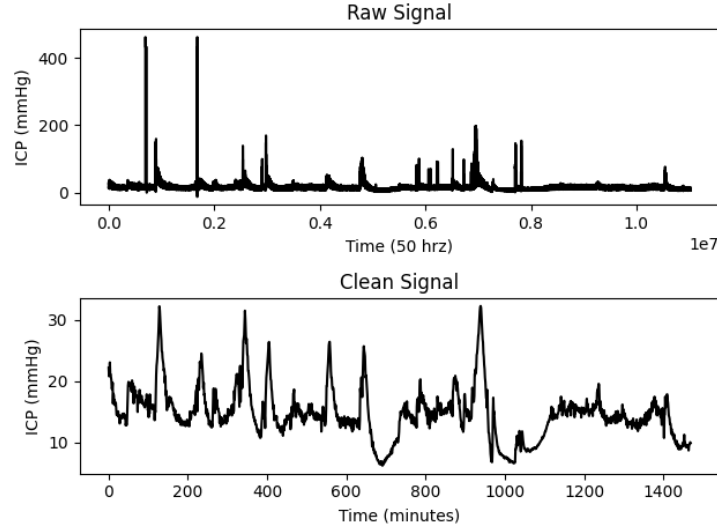[9] The data are available here: https://physionet.org/content/charisdb/1.0.0/

Fig. 1. The raw versus pre-processed data for a random recording.

2.3   Models

In this study, we compare three models: simple ES, a Recurrent Neural Network (LSTM), and the MOMENT model. The main characteristics of the models can be seen in Table 1, and are explained further here. For all models, the only input variable was ICP.

Table 1. The characteristics of the models used.

Model	Requires extensive training	Input-length	Output-length	Only Univariate
MOMENT	Yes (Fine-tuning)	512	Variable	Partly [a]
LSTM	Yes	Variable	Variable	No
ES	No	Variable	Variable	Yes

[a] It can independently model univariate time series.

The ES model is a classical statistical model that makes predictions without requiring a lengthy tuning procedure. It is a univariate model which can theoretically handle any input or forecast length. The model uses weighted averages of past observations, where more recent observations are given higher weights.

The Recurrent Neural Network (RNN) is a more flexible machine learning algorithm typically chosen for time series prediction tasks in healthcare [28]. It can handle any input or output length and can handle multiple input features. In our case the RNN uses an LSTM cell with a size of 512. The architecture

includes an encoder and a decoder, and teacher forcing (i.e., using the ground truth instead of the prediction) is applied 50% of the time.

The MOMENT model[10] is a transformer with an encoder-decoder architecture. The MOMENT model is trained on various time series datasets by masking small parts of a time series and then reconstructing the masked parts. The model input is a univariate time series of length 512. The MOMENT model can handle multiple input features independently. The model can be adapted to any forecasting horizon by replacing the reconstruction head with a forecasting head. This relatively small prediction head maps the output of the encoder to the desired output length (65536 x output length). The prediction head needs to be fine-tuned for the task at hand while the parameters of the encoder are fixed. Dropout was included in the prediction head, set at 0.1 (default setting). See [29] for a complete explanation of the MOMENT architecture.

We use a mean squared error (MSE) loss function with an Adam optimizer [30] and a learning rate of 1e-5 for both the LSTM and the MOMENT models. A batch size of 64 was used for both the training and validation sets. Gradient clipping was applied with a maximum value of 5. *PyTorch (2.4.0)* was used to implement the architectures for the MOMENT and LSTM models. Both were trained for 10 epochs. For the ES model, we used the *tsa.holtwinters.ExponentialSmoothing* function from the *statsmodels* package (0.14.2). All analyses were performed in *Python* (3.10.12).

2.4   Performance assessment

The performance of the different models is assessed using MSE and mean absolute error (MAE), both commonly used for time series forecasting [31]. There is a nested structure in the data, as each patient has multiple segments where the metrics are calculated, and there are multiple patients. For further explanation, see Appendix C.

We create a training and validation set by randomly sampling patients. We use 80% of the patients for training and 20% for validation. Some patients have multiple recordings. Due to the heterogeneous nature of TBI and the limited number of patients, splitting the training and validation data by patient ID can lead to significant variations in the performance metrics, depending on how patients are randomly assigned to each group. To ensure that we capture the performance of the models accurately, we perform k-fold cross-validation (CV) with k = 5. This will give us an indication of the internal validity of the model [32]. The number of recording samples per training and validation set were not uniform across CV partitions, due to the varying recording lengths and varying number of recordings per patient. The results presented for the internal validation will be the mean and the standard deviation (SD) of the performance measures over the 5-folds.

---

[10] In the paper, three models were trained: small, base, and large. Their respective sizes are 40, 125, and 385 million parameters. Only the large model is publicly available and can be found here: https://huggingface.co/AutonLab/MOMENT-1-large. We use the large model in this study.

To assess the performance of the models in a new setting, we perform an external validation using the CHARIS database. Because CV does not yield one final model, we retrain the MOMENT and LSTM models on all available data from TRACK-TBI. The performance of all models is then assessed on the CHARIS dataset. All model parameters are held constant, based on the values used in the CV.

3 Results

3.1 Internal validation

Table 2 shows the results for the internal validation for each model. We see that the MOMENT model performs, based on the MSE and MAE, a little better than the LSTM, and both perform much better than the ES model in forecasting the next 30 ICP values (minutes) based on the previous 60 values. The lowest MAE was 1.78 (MOMENT). Although the MOMENT models outperformed LSTM and ES, the error is still considerably high given the normal range (without an iTH event) is between 0.9 and 16.3. For certain (more difficult) segments, all three models had equally poor performance. In the case of the MOMENT model, 10% of the segments have a MAE value higher than 3.85 and 1% higher than 9.45. The SD of the metrics is high for all models, which indicates heterogeneity in the data. The results for the training set can be found in Appendix D, Table 4. In the training set, the LSTM and MOMENT model perform similarly and again better than the ES model. The SD of the metrics is a lot lower in the training set.

Table 2. Average internal validation performance over 5 CV folds, SD is in brackets.

Metric	MOMENT	LSTM	ES
MSE	9.06 (3.70)	10.19 (3.50)	22.56 (8.31)
MAE	1.78 (0.40)	1.86 (0.32)	3.04 (0.61)
90th percentile MAE	3.85 (0.78)	3.91 (0.72)	6.43 (1.31)
99th percentile MAE	9.45 (1.98)	11.13 (3.53)	14.48 (1.57)

Individual forecasts: In Figures 2 and 3 we can see the (appended) 30-minute predictions with the observed time series (black) for two patients. The patient in Figures 2 has a more stable signal then the patient in Figures 3. We see that the predictions of the MOMENT (red) and LSTM (blue) models resemble the observed signal more closely than the ES (green) model. The models work quite well for the patient in Figure 2, but are substantially worse for the patient in Figure 3.

We will continue discussing the results of the MOMENT and LSTM models, as these performances are relatively close and much better than the ES model.
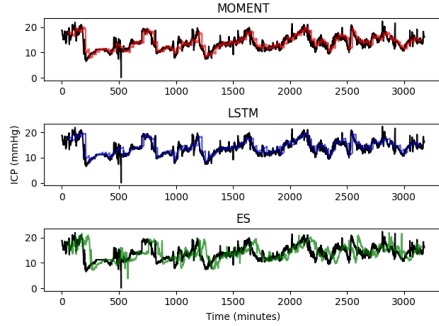
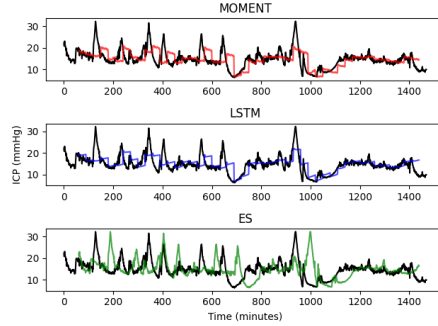**Fig. 2.** ICP signal with "good" forecast.   **Fig. 3.** ICP signal with "bad" forecast.

To illustrate how the predictions change over time, we zoom in on a specific (volatile) segment of 130 minutes. Here we make a new 30-minute prediction every 10 minutes, based on the past 60 minutes (Figure 4 and 5). The black line indicates the observed signal and the colored lines indicate separate 30 minute predictions. For both the MOMENT and LSTM models, the the level of the predictions changes over time, while the shape of the line stays similar. The MOMENT model produces a more variable forecast (squiggly lines) than the LSTM model (straight lines). It is evident that both models lack the ability to predict high-magnitude, low-frequency changes in the observed signal.

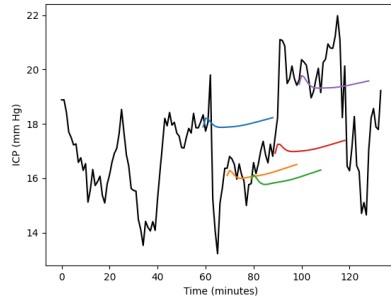**Fig. 4.** 30-minute forecast every 10 min (MOMENT).   **Fig. 5.** 30-minute forecast every 10 min (LSTM).

Performance by segment: It seems that the models are not effective at forecasting changes in the signal, focusing instead on the general trend. We can demonstrate this by examining the relationship between performance (MAE) of the forecast in a segment (Equation 2) and the variance of the ICP itself within

that observed segment. The variance in a segment reflects stability; if the variance is low, the signal remains relatively stable over time. Figure 6 illustrates this relationship for the MOMENT model, showing all segments in the validation sets. The color of the points indicates the density of the segments, revealing that most points correspond to segments with low variance and low MAE. The red line represents a linear model depicting the relationship between variance and MAE. It appears that the more variable a signal is, the higher the MAE. Note that there are some points where the variance is low and the MAE is high, which can be explained by a sudden change in the signal followed by a stable period that is not predicted well by the model. The patterns observed for the MOMENT model are mirrored in the LSTM results (Figure 17, Appendix D).
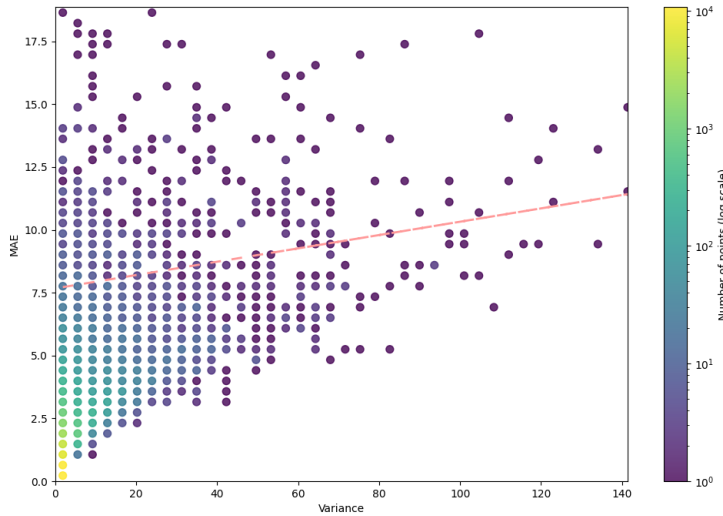

Fig. 6. MAE vs variance for each segment (MOMENT).

Performance by patient: In Figure 7, we observe the MAE for different validation patients using the MOMENT model. The size of the points indicates the average variance in the segments of a patient; a larger size corresponds to a greater average variance. Patients with multiple recordings are represented by multiple points with the same color. From the figure, we can make the following observations: there is substantial variance between patients, and recordings of a patient tend to cluster around the same MAE value. Additionally, larger points correspond to higher MAE, consistent with previous observations. A similar pattern is observed for the LSTM model, as shown in Figure 18 (Appendix D).
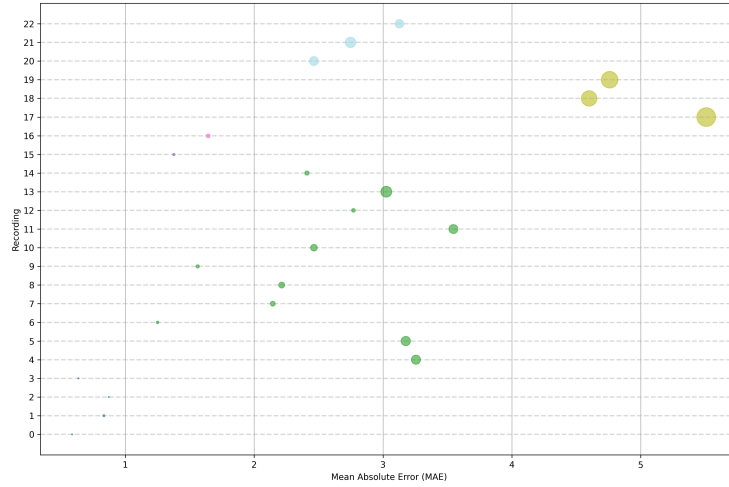
Fig. 7. The MAE of the internal validation recordings from one CV run (MOMENT), grouped by patients (color of the points). The size of the points indicates the variance in the recording.

3.2   External validation

The results from the external validation are shown in Table 3. The results show that the MOMENT and the LSTM models perform very similarly and outperform the ES model. The scale of the performance is comparable with that of the internal validation (Table 2). The training results are shown in Table 5 (Appendix D).

Table 3. External validation performance.

Metric	MOMENT	LSTM	ES
MSE	9.64	9.56	24.77
MAE	1.95	1.92	3.43
90th percentile MAE	4.13	4.02	7.03
99th percentile MAE	9.25	9.22	14.13

The MAE per segment versus variance per segment has a similar pattern as before (Figures 19 and 20, Appendix D). The upper bound for the variance in the segments is higher, indicating that there are more "difficult" segments in the CHARIS dataset compared to the TRACK-TBI dataset.

In Figure 8, we show the MAE for each patient in the external validation set for the MOMENT model. In this dataset, each patient only has one recording. There is again a large spread in the observed MAE values. The larger points, with more variance in the signal, seem to have higher MAE then the smaller

points. The LSTM model's performance characteristics closely resemble those of the MOMENT model (Figure 21, Appendix D).
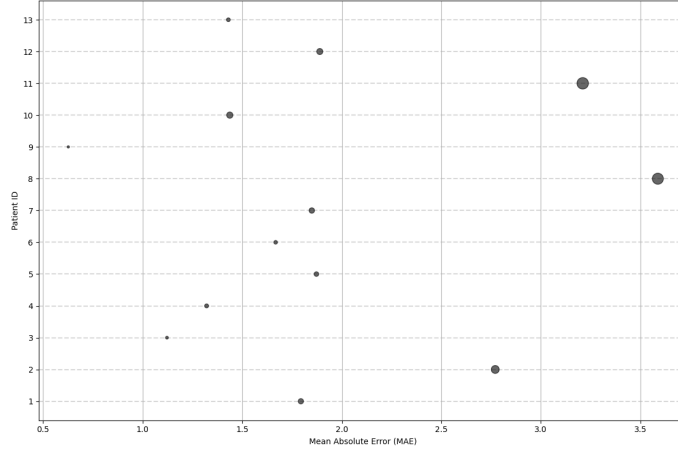

Fig. 8. The MAE of the external validation patients (MOMENT). The size of the points indicates the variance in the recording.

4   Discussion

We developed the first fine-tuned foundation model for forecasting 30 minutes of ICP signal (MOMENT) and evaluated its performance against broadly used time series models (LSTM and ES). We measured their performance using internal validation (TRACK-TBI) and external validation (CHARIS). We have shown that when forecasting the ICP signal, based on the past 60 minutes of the signal, the models do not perform well. Variable segments cannot be predicted well, and there is a large variation in the performance between patients.

The results are worse than those reported in [8], where an RMSE of 2.18 (MSE = 4.75) was achieved. We believe this discrepancy can be partly explained. Firstly, the choice was made to forecast 10 minutes ahead, while we forecast 30 minutes ahead. This highlights the principle that the further into the future we forecast, the more challenging it becomes to make accurate predictions. Secondly, with only 13 patients available, they opted to use only a training set in the development and assessment of their model. Additionally, the data they used come from a single site, which might have helped the model learn more effectively. In contrast, our training comes from three different sites, which may introduce more variability into the data.

The models used in this study differ in their ease of use. We observed the following when using the different models. The ES model is very easy to use

and converges quickly. The LSTM model trains quickly and can also be easily adapted to take in multiple features or to adjust the length of the input/forecast. The MOMENT model was most difficult to implement. The input window for MOMENT is fixed at a size of 512. This means that in many cases, zero padding needs to be used, which is computationally wasteful as the "real" input length is then shorter than the one the model uses to produce predictions. The training (fine-tuning) of the MOMENT model is slow compared to that of the LSTM; in this study, training the MOMENT model took about 10 times longer than the LSTM model.

A limitation of our approach is that we do not account for the uncertainty around our predictions, as a clinician might want to know how certain we are about a particular forecast before using it in practice. For the MOMENT model, the architecture is fixed, preventing us from using a Bayesian neural network [33], which would also be challenging due to computational constraints. Other options that could be used for obtaining uncertainty estimates include Monte Carlo dropout [34] and the conformal prediction framework for time series [35].

Before a clinical decision support model could actually be implemented, we need to be sure it will work in different settings. The performance observed in this study is likely an underestimate of how well the model would work in an external setting, as evidenced by the discrepancy between the training and external validation performance. Therefore, before considering implementation, the model should be externally validated (multiple times) in similar settings to where they will be applied [36]. If external validation occurs in many settings then one could create a prediction interval for how well the model will perform in a new setting [37]. This prediction interval could serve as a basis for a check to ensure there is not too much uncertainty to implement the model. If we see that the prediction interval is very wide, then it might be a good idea to fine-tune the model with some local data before implementation.

We conclude this work by noting that based on our research, ICP forecasting based solely on the signal does not achieve sufficient performance for practical implementation. The concept of incorporating prior information into a model has a strong theoretical advantage, which was only partially realized in this study. We speculate that the ICP signal may differ too significantly from the signals used in the training of the MOMENT model, or alternatively, that the unexplained variance in the ICP signal is simply too large.

References

1. Joukje van der Naalt, Alex Maas, David K Menon, E.W. Steyerberg, Giuseppe Citerio, F. Lecky, G.T. Manley, S Hill, Victor Legrand, A. Sorgner, and CENTER-TBI investigators. Collaborative European Neuro Trauma Effectiveness Research in traumatic brain injury (CENTER-TBI): a prospective longitudinal observational study. *Neurosurgery*, 76(1):67–80, January 2015.
2. ACRM Mild Traumatic Brain Injury Committee. Definition of mild traumatic brain injury. *J Head Trauma Rehabil*, 8(3):86–7, 1993.
3. Olli Tenovuo, Ramon Diaz-Arrastia, Lee E. Goldstein, David J. Sharp, Joukje Van Der Naalt, and Nathan D. Zasler. Assessing the severity of traumatic brain injury—time for a change? *Journal of clinical medicine*, 10(1):148, 2021. Publisher: MDPI.
4. Nino Stocchetti and Andrew I.R. Maas. Traumatic Intracranial Hypertension. *New England Journal of Medicine*, 370(22):2121–2130, May 2014.
5. Robert McNamara, Shiv Meka, James Anstey, Daniel Fatovich, Luke Haseler, Toby Jeffcote, Andrew Udy, Rinaldo Bellomo, and Melinda Fitzgerald. Development of Traumatic Brain Injury Associated Intracranial Hypertension Prediction Algorithms: A Narrative Review. *Journal of Neurotrauma*, 40(5-6):416–434, March 2023.
6. D. James Cooper, Jeffrey V. Rosenfeld, Lynnette Murray, Yaseen M. Arabi, Andrew R. Davies, Paul D'Urso, Thomas Kossmann, Jennie Ponsford, Ian Seppelt, Peter Reilly, and Rory Wolfe. Decompressive Craniectomy in Diffuse Traumatic Brain Injury. *New England Journal of Medicine*, 364(16):1493–1502, April 2011.
7. Peter J. Hutchinson, Angelos G. Kolias, Ivan S. Timofeev, Elizabeth A. Corteen, Marek Czosnyka, Jake Timothy, Ian Anderson, Diederik O. Bulters, Antonio Belli, C. Andrew Eynon, John Wadley, A. David Mendelow, Patrick M. Mitchell, Mark H. Wilson, Giles Critchley, Juan Sahuquillo, Andreas Unterberg, Franco Servadei, Graham M. Teasdale, John D. Pickard, David K. Menon, Gordon D. Murray, and Peter J. Kirkpatrick. Trial of Decompressive Craniectomy for Traumatic Intracranial Hypertension. *New England Journal of Medicine*, 375(12):1119–1130, September 2016.
8. Guochang Ye, Vignesh Balasubramanian, John K-J. Li, and Mehmet Kaya. Machine Learning-Based Continuous Intracranial Pressure Prediction for Traumatic Injury Patients. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–8, 2022. Conference Name: IEEE Journal of Translational Engineering in Health and Medicine.
9. Akram Farhadi, Joshua J. Chern, Daniel Hirsh, Tod Davis, Mingyoung Jo, Frederick Maier, and Khaled Rasheed. Intracranial Pressure Forecasting in Children Using Dynamic Averaging of Time Series Data. *Forecasting*, 1(1):47–58, December 2019. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
10. Bin Han, Michael Muma, Mengling Feng, and Abdelhak M. Zoubir. An online approach for intracranial pressure forecasting based on signal decomposition and robust statistics. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6239–6243, May 2013. ISSN: 2379-190X.
11. Giorgia Carra, Fabian Güiza, Bart Depreitere, Geert Meyfroidt, Audny Anke, Ronny Beer, Bo-Michael Bellander, Erta Beqiri, Andras Buki, Manuel Cabeleira, Marco Carbonara, Arturo Chieregato, Giuseppe Citerio, Hans Clusmann, Endre Czeiter, Marek Czosnyka, Bart Depreitere, Ari Ercole, Shirin Frisvold, Raimund Helbok, Stefan Jankowski, Danile Kondziella, Lars-Owe Koskinen, Ana Kowark,

David K. Menon, Geert Meyfroidt, Kirsten Moeller, David Nelson, Anna Piippo-Karjalainen, Andreea Radoi, Arminas Ragauskas, Rahul Raj, Jonathan Rhodes, Saulius Rocka, Rolf Rossaint, Juan Sahuquillo, Oliver Sakowitz, Peter Smielewski, Nino Stocchetti, Nina Sundstro¨m, Riikka Takala, Olli Tenovuo, Peter Vajkoczy, Alessia Vargiolu, Rimantas Vilcinis, Stefan Wolf, Alexander Younsi, Frederick A. Zeiler, and CENTER-TBI High-Resolution ICU (HR ICU) Sub-Study Participants and Investigators. Prediction model for intracranial hypertension demonstrates robust performance during external validation on the CENTER-TBI dataset. *Intensive Care Medicine*, 47(1):124–126, January 2021.

12. Nils Schweingruber, Marius Marc Daniel Mader, Anton Wiehe, Frank Röder, Jennifer Göttsche, Stefan Kluge, Manfred Westphal, Patrick Czorlich, and Christian Gerloff. A recurrent machine learning model predicts intracranial hypertension in neurointensive care patients. *Brain*, 145(8):2910–2919, February 2022.

13. Jan G. De Gooijer and Rob J. Hyndman. 25 years of time series forecasting. *International Journal of Forecasting*, 22(3):443–473, January 2006.

14. Charles C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*, 20(1):5–10, 1957.

15. George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 1970. Google-Books-ID: rNt5CgAAQBAJ.

16. Jeffrey L. Elman. Finding Structure in Time. *Cognitive Science*, 14(2):179–211, 1990. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1.

17. Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997. Conference Name: Neural Computation.

18. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

19. Shruti Kaushik, Abhinav Choudhury, Pankaj Kumar Sheron, Nataraj Dasgupta, Sayee Natarajan, Larry A. Pickett, and Varun Dutt. AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. *Frontiers in Big Data*, 3, March 2020. Publisher: Frontiers.

20. Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance,

Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, July 2022. arXiv:2108.07258 [cs].

21. Sebastian Thrun and Tom M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1):25–46, July 1995.

22. Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023.

23. Nam Kim, Alex Krasner, Colin Kosinski, Michael Wininger, Maria Qadri, Zachary Kappus, Shabbar Danish, and William Craelius. Trending autoregulatory indices during treatment for traumatic brain injury. *Journal of Clinical Monitoring and Computing*, 30(6):821–831, December 2016.

24. Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23), June 2000.

25. Venessa L. Pinto, Prasanna Tadi, and Adebayo Adeyinka. Increased Intracranial Pressure. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024.

26. Cornelia Gruber, Patrick Oliver Schenk, Malte Schierholz, Frauke Kreuter, and Göran Kauermann. Sources of Uncertainty in Machine Learning – A Statisticians' View, May 2023. arXiv:2305.16703 [cs, stat].

27. Fabian Güiza, Bart Depreitere, Ian Piper, Giuseppe Citerio, Philippe G. Jorens, Andrew Maas, Martin U. Schuhmann, Tsz-Yan Milly Lo, Rob Donald, Patricia Jones, Gottlieb Maier, Greet Van den Berghe, and Geert Meyfroidt. Early detection of increased intracranial pressure episodes in traumatic brain injury: external validation in an adult and in a pediatric cohort. *Critical care medicine*, 45(3):e316–e320, 2016.

28. Mohammad Amin Morid, Olivia R. Liu Sheng, and Joseph Dunbar. Time Series Prediction Using Deep Learning Methods in Healthcare. *ACM Transactions on Management Information Systems*, 14(1):1–29, March 2023.

29. Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A Family of Open Time-series Foundation Models, May 2024. arXiv:2402.03885 [cs].

30. Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. arXiv:1412.6980 [cs].

31. Ratnadip Adhikari and R. K. Agrawal. An Introductory Study on Time Series Modeling and Forecasting, February 2013. arXiv:1302.6613 [cs, stat].

32. Gary S Collins, Paula Dhiman, Jie Ma, Michael M Schlussel, Lucinda Archer, Ben Van Calster, Frank E Harrell, Glen P Martin, Karel G M Moons, Maarten Van Smeden, Matthew Sperrin, Garrett S Bullock, and Richard D Riley. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*, page e074819, January 2024.

33. Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? In *Pro-

ceedings of the 38th International Conference on Machine Learning, pages 4629–4640. PMLR, July 2021. ISSN: 2640-3498.

34. Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. June 2016.

35. Kamile Stankeviciute, Ahmed M. Alaa, and Mihaela van der Schaar. Conformal Time-series Forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pages 6216–6228. Curran Associates, Inc., 2021.

36. Ben Van Calster, Ewout W. Steyerberg, Laure Wynants, and Maarten Van Smeden. There is no such thing as a validated prediction model. *BMC Medicine*, 21(1):70, February 2023.

37. Florian D. van Leeuwen, Ewout W. Steyerberg, David van Klaveren, Ben Wessler, David M. Kent, and Erik W. van Zwet. Empirical Evidence That There Is No Such Thing As A Validated Prediction Model, June 2024. arXiv:2406.08628 [stat].

38. Nicolas Hernandez Norager, Markus Harboe Olsen, Sarah Hornshoej Pedersen, Casper Schwartz Riedel, Marek Czosnyka, and Marianne Juhler. Reference values for intracranial pressure and lumbar cerebrospinal fluid pressure: a systematic review. *Fluids and Barriers of the CNS*, 18:19, April 2021.

Appendix A

We see in Figure 9 that most recording sessions have a duration of less than 100 hours (approximately 4 days). There are a few recordings with a much longer recording time.
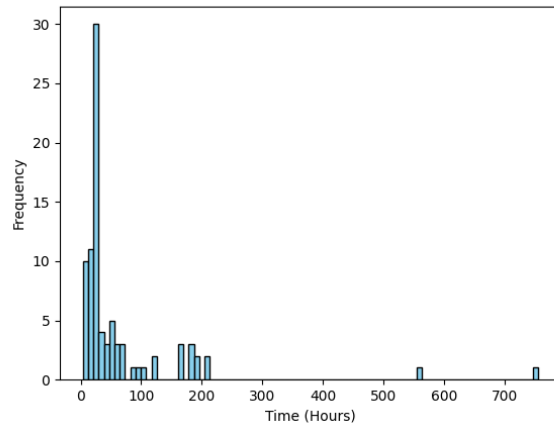

Fig. 9. Frequency of the recording times of the ICP recordings from TRACK-TBI.

Appendix B

In Figure 10 the preprocessing algorithm is outlined. The goal of the algorithm is to remove artifacts, smooth the time series, and downsample the data to make training computationally feasible. The input is a signal with 50 measurements every second (50 Hz). In the TRACK-TBI data not all recordings were measured at 50 Hz, so we look the mean for every 1/50 seconds. Some functions used in the algorithm are explained here:

– Forward_fill(): If a value is missing, the next available value is used to fill it in.
– Down_sampled(): The time series is downsampled to 1/ds of the original length by taking the mean.

For the inputs in Figure 10 we use the following values: $w = 60000$ (20 minutes), $st = 3000$, $ds = 3000$, $ICP_{max} = 50$, $ICP_{min} = -5$ [11].

Algorithm 1 Preprocess the ICP input signal by the method adapted from Ye et al. 2022.

Input: x: the input signal
Input: w: size of the sliding window
Input: st: steps of the sliding window
Input: ds: the number of points in each segment to take the mean from in downsampling
Input: ICP_{max} the maximun value of ICP considered non an artifact
Input: ICP_{min} the minimum value of ICP considered non an artifact
Output: The preprocessed downsampled signal

```
1  i ← 1
2  while i + w ≤ len(x) do
3      segment ← x[i + w]
4      for j = 1 to w do
5          if segment[j] > ICP_max or segment[j] < ICP_min then
6              segment[j] ← NA
7          end if
8      end for
9      segment ← Forward_fill(segment)
10     mean_s ← Mean(segment), sd_s ← SD(segment)
11     for k = 1 to w do
12         if segment[k] > mean_s + 3 × sd_s or segment[k] < mean_s − 3 × sd_s then
13             segment[k] ← mean_s
14         end if
15     end for
16     x[i + w] ← segment
17     i ← i + st
18 end while
19 y ← Down_sampled(x, ds)
```

Fig. 10. Preprocessing algorithm

Similarly, in Figure 11, the segmentation process is outlined. The goal is to transform a time series into smaller parts that can be used for training the model. We start by taking the first 'in_len' data points from the time series as the first 'X', and the adjacent 'out_len' data points as the 'Y'. We then move forward in the time series by a step size of 'str_len', and perform the same procedure. This results in multiple 'X' and 'Y' segments from a single time series, with many segments containing overlapping data points. The 'Y' segments always follow the corresponding 'X' segments. The Zeros function creates a vector of length (512 - 'in_len') to append to the segment. This is necessary for the MOMENT

---

[11] The ICP values depend on the position of the patient, negative values can be observed when a patient is in an upright position [38]

model, as the input vector has a fixed length of 512. For the inputs in Figure 11, we use the following values: in_len = 60, out_len = 30, str_len = 5.

Algorithm 2 Segmentation of the timeseries.

Input: x: the input signal
Input: in_len: the length of the input window
Input: out_len: the length of the output window
Input: str_len: the stride length
Input: *moment*: a boolean flag for zero padding
Output: X_{full}, Y_{full}: The preprocessed sequences for training the models
1 $X_{full} \leftarrow []$
2 $Y_{full} \leftarrow []$
3 **for** $i = 0$ **to** $(\text{len}(x) - \text{in\_len} - \text{out\_len})$ **step** str_len **do**
4   **if** *moment* **is True then**
5          $X_{full}.\text{append}(\text{append}(signal[i : (i + \text{in\_len}], \text{Zeros}(512 - \text{in\_len})))$
6   **else**
7          $X_{full}.\text{append}(signal[i : (i + \text{in\_len})])$
8   **end if**
9      $Y_{full}.\text{append}(signal[(i + \text{in\_len}) : (\text{in\_len} + \text{out\_len})])$
10 **end for**

Fig. 11. Segmentation algorithm

After the preprocessing, there were some signals that we considered unrealistic. An example can be seen in Figure 12. The straight section in the middle of the time series is caused by an excessive number of missing or unrealistic values, which were set to missing. We do not want the models to learn such behavior.

For some of the segments, we observed that the end of the signal appeared distorted. This could be due to disconnection during measurement. An example can be seen in Figure 13. For these signals (6 out of 83), we removed the last part of the signal.

In Figure 14, we see all the signals from TRACK-TBI used for training and evaluating the model.
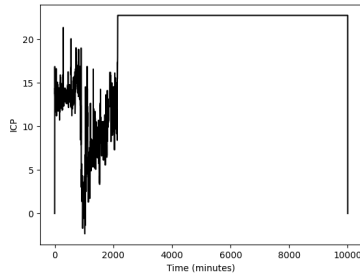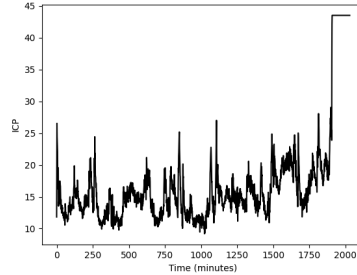

Fig. 12. Unrealistic signal

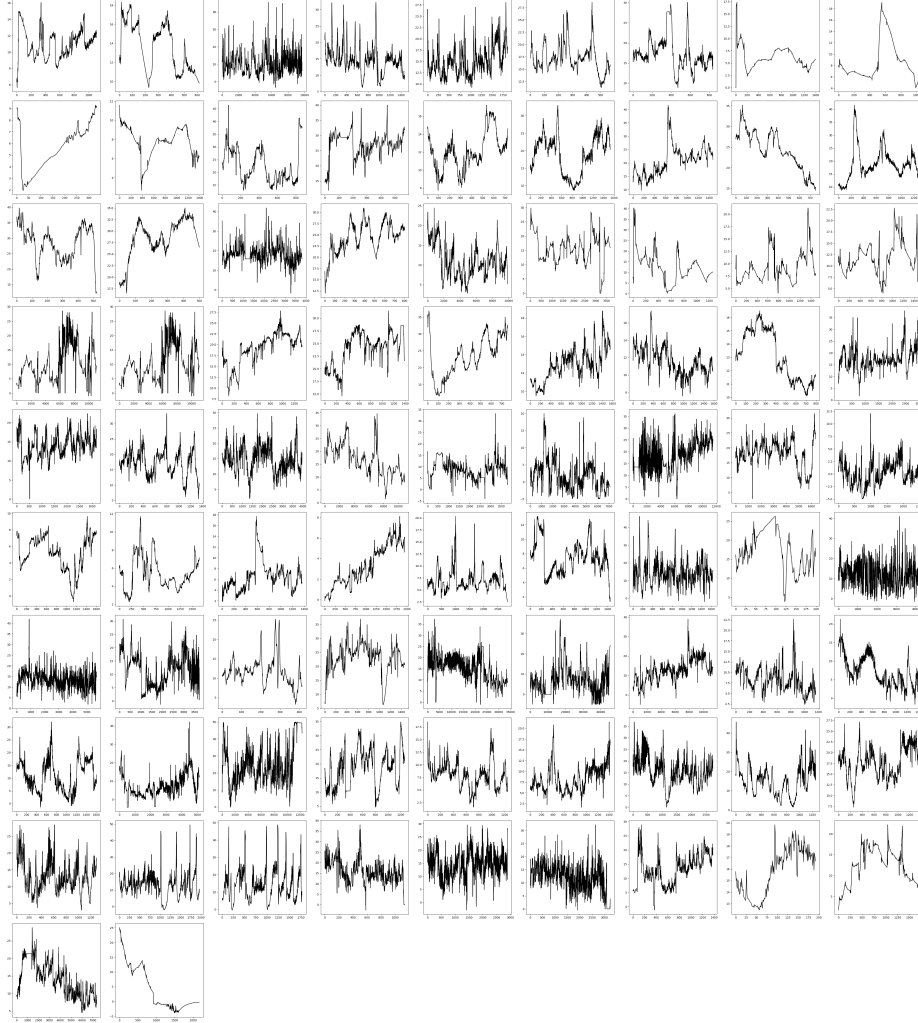**Fig. 13.** Unrealistic ending of signal

Fig. 14. All the preprocessed signals from TRACK-TBI.

Appendix C

The performance measurements can be denoted as follows. The MSE of a segment is defined as:

$$MSE_{segment} = \frac{1}{s} \sum_{i=1}^{s} (y_i - \hat{y}_i)^2 \tag{1}$$

where i represents the number of data points in a segment, y_i the observed ICP value, $\hat{y}_i$ the predicted ICP value and s is the total number of data points in the segment. Similarly for the MAE:

$$MAE_{segment} = \frac{1}{s} \sum_{i=1}^{s} |y_i - \hat{y}_i| \tag{2}$$

To obtain the complete MAE of a patient, we need to take the average MAE over all segments:

$$MAE_{patient} = \frac{1}{k} \sum_{j=1}^{k} MAE_{segment_j} \tag{3}$$

with k being the number of segments. The overall MAE is then the average over all patients:

$$MAE_{model} = \frac{1}{N} \sum_{n=1}^{N} MAE_{patient_n} \tag{4}$$

With N being the number of patients.

Appendix D

For the internal validation, the training/validation loss per epoch can be seen in Figure 15 and 16 for the MOMENT and LSTM model respectively. The bright lines are the mean losses over the CV runs, and the shaded lines are the CV runs themselves. There are two validation runs with a very high loss; after inspection of the signals, it seems that in both validation sets there was one patient with multiple recordings with a very volatile signal. The MOMENT model seems to converge quicker and more smoothly than the LSTM.

Fig. 15. Loss vs epoch for MOMENT.   **Fig. 16.** Loss vs epoch for LSTM.

The training performance is a bit better than the validation performance (Table 4). Interestingly enough, the LSTM (slightly) outperforms the MOMENT model in the training set. It thus seems like the LSTM overfits on the training set, as the internal validation performance is better for the MOMENT model Table 2.

Table 4. Average training performance over 5 CV folds, SD is in brackets.

Metric	MOMENT	LSTM	ES
MSE	7.77 (0.87)	7.92 (0.89)	20.22 (1.61)
MAE	1.62 (0.11)	1.61 (0.14)	2.83 (0.13)
90th percentile MAE	3.66 (0.23)	3.60 (0.23)	6.11 (0.29)
99th percentile MAE	8.88 (0.43)	9.03 (0.44)	14.68 (0.32)

The MAE versus the variance per segment in the LSTM model is shown in Figure 17. It shows the same relationship as was seen in Figure 6; most segments have a low MAE and a low variance, and it seems that as the variance increases, the MAE increases.
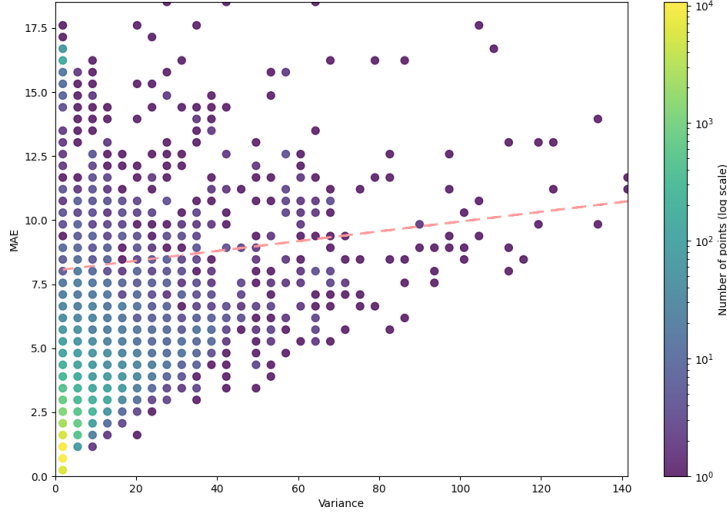
Fig. 17. MAE vs variance for each segment (LSTM model).

The MAE per recording, grouped by patients, is shown in Figure 18. This is
for the same CV run, but this time for the LSTM model. We again see a very
similar pattern to Figure 7.

The training performance when the model is trained using all available data
is shown in Table 5. The performance is a bit better than the average training
performance of the 5-fold CV (Table 4). The MOMENT and LSTM models still
perform very similarly, and a lot better than the ES model.

Table 5. Training performance using all TRACK-TBI data.

Metric	MOMENT	LSTM	ES
MSE	7.57	7.70	20.07
MAE	1.60	1.61	2.82
90th percentile MAE	3.61	3.58	6.08
99th percentile MAE	8.71	8.82	14.66

For the external validation dataset, we can also inspect the relationship be-
tween the variance in the segments and the MAE for the MOMENT and LSTM
models (Figures 19 and 20). There seem to be segments with higher variance in
this dataset compared to the TRACK-TBI data. The relationship between the
MAE and variance is similar for the MOMENT and LSTM models and seems
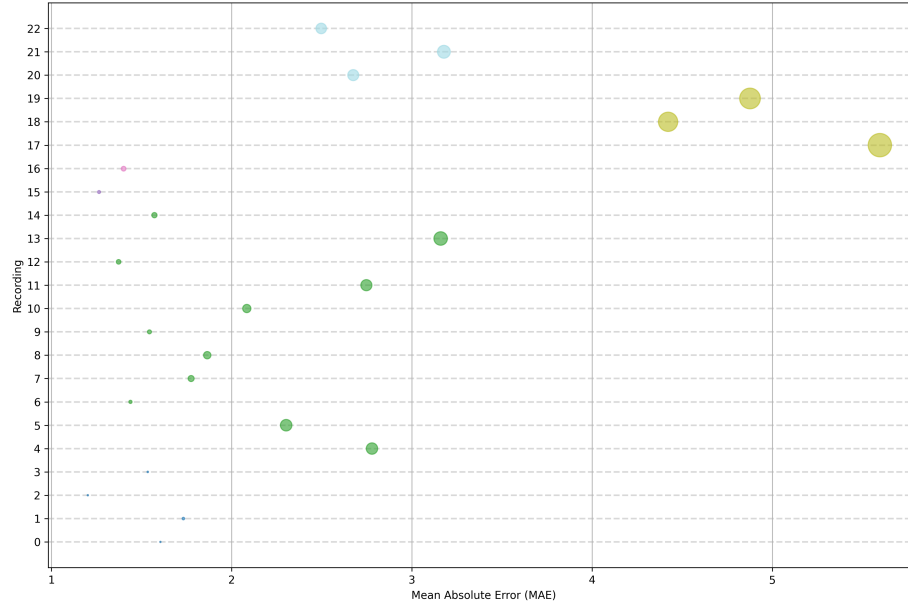more extreme than in the internal validation (Figure 6 and17).

Fig. 18. The MAE of the validation recording from one CV run (LSTM), grouped by patients (color of the points). The size of the points indicates the variance in the recording.
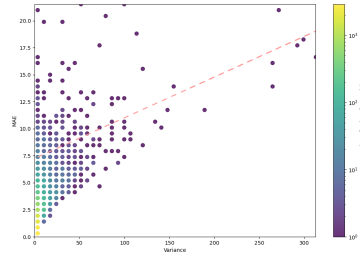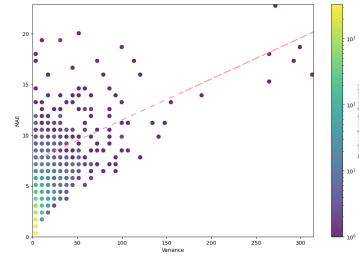

Fig. 19. MAE vs variance (MOMENT).

Fig. 20. MAE vs variance (LSTM).

The MAE of the LSTM versus the variance of in the signal of the external validations patients is shown in 21.
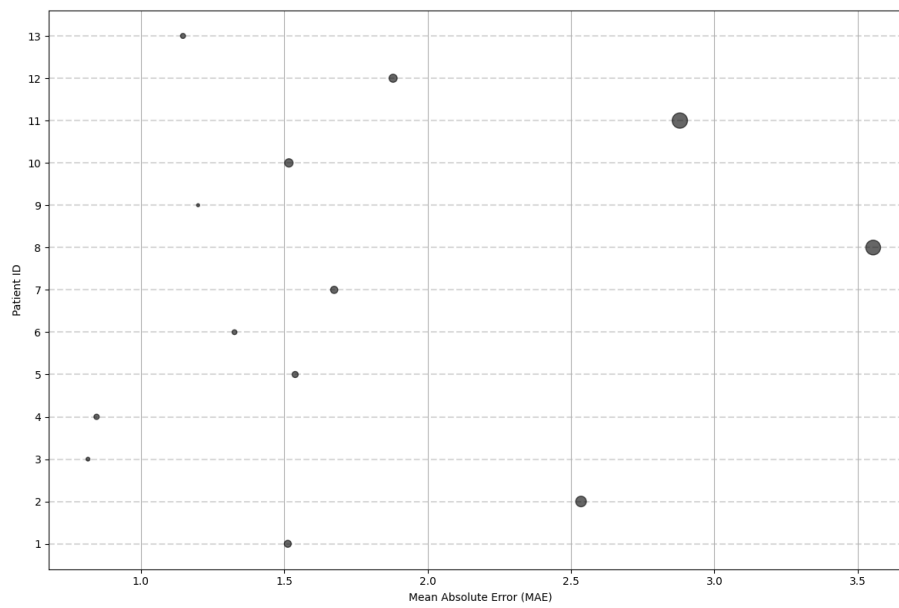
Fig. 21. The MAE of the external validation patients (LSTM). The size of the points indicates the variance in the recording.