



Universiteit
Leiden

The Netherlands

Increased intracranial pressure in patients with traumatic brain injury in the intensive care unit

Dolmans, R.G.F.

Citation

Dolmans, R. G. F. (2024, November 28). *Increased intracranial pressure in patients with traumatic brain injury in the intensive care unit*. Retrieved from <https://hdl.handle.net/1887/4170310>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4170310>

Note: To cite this publication please use the final published version (if applicable).





CHAPTER 4

Deriving Automated Device Metadata from Intracranial Pressure Waveforms: A TRACK-TBI ICU Physiology Cohort Analysis

Sophie E. Ack*, Rianne G.F. Dolmans*, Brandon P. Foreman, Geoffrey T. Manley, Eric S. Rosenthal, Morteza Zabihi and TRACK-TBI Investigators.

* shared first authorship

Critical Care Explorations. 2024 Jul 16;6(7):e1118.



Abstract

Importance: Treatment for intracranial pressure (ICP) has been increasingly informed by machine learning (ML) derived ICP waveform characteristics. There are gaps however in understanding how ICP monitor type may bias waveform characteristics utilized for these predictive tools since differences between external ventricular drain (EVD) and intraparenchymal monitor (IPM) derived waveforms have not been well accounted for.

Objectives: We sought to develop a proof-of-concept ML model differentiating ICP waveforms originating from an EVD or IPM.

Design, Setting and Participants: We examined raw ICP waveform data from the Intensive Care Unit (ICU) Physiology cohort within the prospective Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) multicenter study.

Main Outcomes and Measures: Nested patient-wise 5-fold cross-validation and group analysis with bagged decision trees (BDT) and linear discriminant analysis (LDA) were utilized for feature selection and fair evaluation. Nine patients were kept as unseen hold-outs for further evaluation.

Results: 14,110 hours of ICP waveform data were included from 82 patients (EVD, 47; IPM, 26; both, 9). Mean age, GCS total, and GCS motor score upon admission, as well as the presence and amount of midline shift, were similar between groups. Model mean area under the receiver operating characteristic curve (AU-ROC) exceeded 0.874 across all folds. In additional rigorous cluster-based subgroup analysis, targeted at testing the resilience of models to cross-validation with smaller subsets constructed to develop models in one confounder set and test them in another subset, AU-ROC exceeded 0.811. In a similar analysis using propensity score- rather than cluster-based subgroup analysis, mean AU-ROC exceeded 0.827. 62 of 842 extracted ICP features were invariant within every analysis, representing the most accurate and robust differences between ICP monitor types. For the 9 patient hold-outs, AU-ROC of 0.826 was obtained using BDT, respectively.

Conclusions and Relevance: The developed proof-of-concept ML model identified differences in EVD- and IPM-derived ICP signals, which can provide missing

contextual data for large-scale retrospective datasets, prevent bias in computational models that ingest ICP data indiscriminately, and control for confounding using our model's output as a propensity score by to adjust for the monitoring method that was clinically indicated. Moreover, the invariant features may be leveraged as ICP features for anomaly detection.

Introduction

Intracranial pressure (ICP) monitoring is used for guiding management of acute brain injury (ABI) in patients at risk for elevated ICP or hydrocephalus^{1,2}. While mean ICP has been widely integrated into treatment standards³, ICP waveforms capture continuous, high-resolution data on the trajectory of ICP, providing valuable information about cerebral perfusion pressure (CPP), cerebral compensatory reserve, and regulation of cerebral blood flow and volume^{4,5}. ICP-derived features have provided significant additional information and been increasingly implemented in recent years, such as pulse amplitude index (PAX), the correlation between cerebral CPP, and pulse amplitude (AMP)^{6,7}.

Machine learning (ML) approaches are powerful tools to further analyze and interpret ICP waveform data, which offer the promise to expand our ability to detect and predict intracranial hypertension or clinical deterioration⁸⁻¹². Quantitative characteristics have been utilized to enhance ICP signal quality and recognize non-artifactual ICP pulses⁷, identify clamping of external ventricular drain (EVD)-derived ICP waveforms¹³, and quantify the P1, P2, and P3 peaks within ICP pulse (which reflect the routine cycle of ICP in the brain, containing valuable continuous information on dynamic cerebrospinal pathophysiology, rather than the overall mean value)¹⁴ (Figure 1).

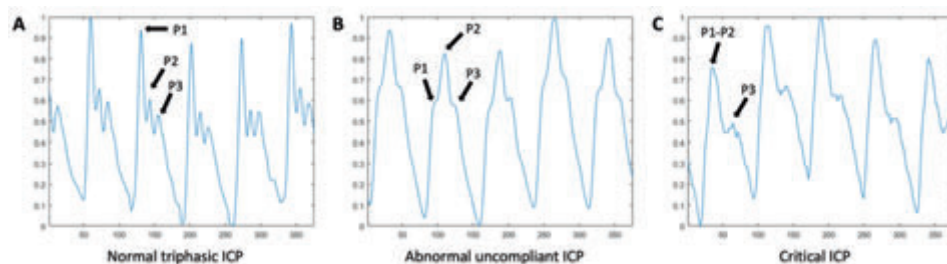


Figure 1: Example ICP waveform data from this cohort. A) Normal triphasic ICP. B) Abnormal, uncompliant ICP. C) Critical ICP.

While ML-derived ICP waveform characteristics have increasingly informed treatment, there are gaps in understanding how ICP monitor type may bias waveform characteristics utilized for these predictive tools. Differences between EVD- and IPM-derived waveforms have not been well accounted for, which may be critical to prevent bias in computational models that ingest ICP data

indiscriminately, particularly as this contextual data is missing in many large-scale retrospective datasets.

The decision to place an EVD versus an IPM may relate to the need for cerebrospinal fluid (CSF) drainage to manage intracranial hypertension^{1,3,15-17}. Accordingly, tools that utilize waveform data may be confounded by clinical indication, predicting outcomes based on patient characteristics associated with EVD placement, rather than signal features that encode the physiologic underpinnings of future risk. For example, EVD waveforms are altered during continuous or intermittent CSF drainage^{1,18,19}, may be dampened by catheter malpositioning or partial occlusion^{1,20}, and may be inaccurate after patient movement until re-leveling^{1,20}. IPMs may also have inaccuracies related to varying degrees of zero drift over time, which influence the ICP waveform^{1,3,16,21}. We, therefore, hypothesized that there are differences in the ICP waveform features between EVDs and IPMs. However, in many clinical datasets, ICP monitor-type data is unavailable; the information is not encoded in physiologic monitor outputs and can be missed or expensive to obtain²². If predictive models developed without this contextual information are to be robust and reliable in guiding clinical management without confounding by monitor indication, learning the provenance of ICP data will likely be necessary to ensure models are accurate and generalizable. For example, this type of device information is essential when seeking FDA approval for software algorithms, given the need to define a "context of use" for algorithms using clinical data. We, therefore, sought to understand the effect of ICP monitor type on ICP waveform features and to develop a tool identifying the source of an ICP waveform that can aid efforts to utilize ICP waveform data for clinical decision support^{8,14,23-25}.

In this proof-of-concept framework, we examined the multicenter Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI) Intensive Care Unit (ICU) Physiology cohort. We aimed to identify the distinctive waveform characteristics associated with each type of ICP monitor and to develop a machine learning framework for classifying ICP waveform data as originating from an EVD or an IPM.

Methods

Population and inclusion criteria

The TRACK-TBI study was approved by the institutional review board (IRB) of each site, and written informed consent was obtained from all subjects (see Supplementary Materials for IRB details). Procedures were followed in accordance

with the ethical standards of IRBs and with the Helsinki Declaration as revised in 2013. Inclusion and exclusion criteria for the TRACK-TBI study (ClinicalTrials.gov #NCT02119182) have been previously described²⁶⁻²⁸. Inclusion criteria for this study were age ≥ 18 years, enrollment from hospitals with the capability to record and extract bedside telemetry, available ICP waveform recordings with ground-truth labeling of ICP source, and undergoing ICP monitoring. Exclusion criteria for ICP epochs were corrupted files, flat signal with no physiologic content, or duration less than ten minutes (precluding appropriate windowing for feature extraction; figure 2). The term "ICP epochs" here refers to the division of ICP recordings into separate files or segments. Each epoch typically corresponds to a distinct time period within the recordings. The Kruskal-Wallis one-way analysis of variance test for non-parametric data was used to investigate if there were clinical differences between groups with EVD, IPM, and both monitor types (Supplemental Table S1).

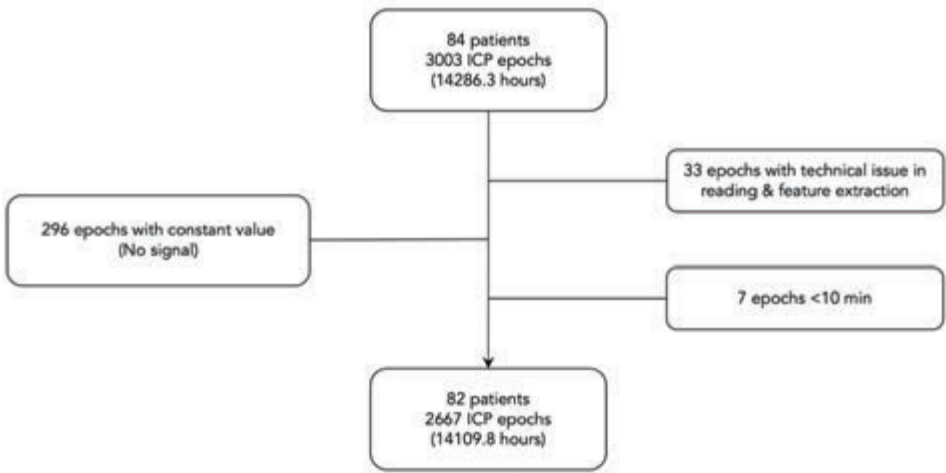


Figure 2: CONSORT diagram of included data.

ICP data preprocessing

To ensure consistency in our analysis, we resampled the ICP waveforms to a common sampling frequency using a finite impulse response (FIR) antialiasing low pass filter on the raw signal²⁹. This low pass filter attenuates the frequency components above a specific cutoff frequency, downsampling the bandwidth of a signal. Given that most waveforms were recorded at a sampling frequency of 125 Hz and that this was the lowest frequency present in our study population, we

selected this as our desired sampling frequency and subsequently down-sampled the remaining waveforms to align with it. Furthermore, a straightforward outlier detection protocol was implemented based on the probability of occurrence of each sample value within each waveform. This protocol identifies outliers as samples with a probability equal to or less than 0.01, which was empirically obtained. These outliers were replaced with the median value of the respective waveform to ensure that artifacts with high absolute amplitude do not impact our features. In post-hoc experiments to evaluate whether the care environment or the innate monitor signal was imprinting information in the waveform, we extended our analysis by removing to ensure that our model performance was representative of true elements of the ICP signals, rather than the influence of environmental factors such as adjustment of EVD height. In the supplementary materials, we have conducted comparative analyses to demonstrate that the added value of such artifacts to our model did not significantly improve its performance.

Feature extraction

The necessity to extract features from various domains (including time, frequency, time-frequency, and nonlinear) in physiological waveform analysis stems from the multi-dimensional nature of such data. Each domain captures distinct aspects of the data, facilitating a comprehensive understanding of the underlying patterns and dynamics. Time series data often contains noise or irrelevant information that can obscure the underlying signal. Transforming the data into different domains through techniques like Fourier or wavelet transforms facilitates the separation of the signal from the noise, enhancing the ability to identify meaningful features and patterns within the data. Thus, in this study, a comprehensive and established set of features from various domains was extracted and can be classified into four distinct categories: (1) General features, capturing the overall characteristics of the entire ICP waveform; (2) Beat-to-beat features, capturing the characteristics of each ICP beat, defined as the period between each P1 (percussion wave); (3) Beat variability features, capturing the fluctuations of time between ICP beats caused by the cardiac pulsation cycle; and (4) Window-based features, capturing the ICP features within rolling windows with a length of five minutes and no overlap. Selection of features was based on previous literature in the field of time series analysis. These included, but were not limited to, autocorrelation, energy in short-time Fourier transform domain, power spectral density features, approximate and Shannon entropy, Hurst exponent, Higuchi fractal dimension, and polynomial curve

fitting, calculated as previously described^{6,30-35}. To account for the varying size of the ICP waveforms and incorporate temporal information, we extracted a secondary set of features that included the maximum, median, standard deviation, 1st, 25th, 75th, and 99th percentiles, as well as the mean and standard deviation of first and second differences (lags) of beat-to-beat, beat variability, and window-based features. Lag features incorporate historical data into a time series analysis or forecasting model and can help capture important patterns and trends in the data. The first difference is created by taking the difference between a value in a time series and a previous value. The second difference is the first difference of the first difference feature. This approach allowed us to create a fixed input feature vector despite the variability in waveform length. In total, we extracted 842 features (Table 1).

Table 1. List of extracted features. STFT, short-time Fourier transform; SD, standard deviation, $diff_1$, first difference, $diff_2$, second difference.

Feature Category	Description	Features	Secondary features
General	Captures the overall characteristics of the entire ICP waveform	Autocorrelation Hurst exponent ³⁶ Kolmogorov Complexity ³⁷ Approximation entropy Higuchi fractal dimension ³⁸ Katz fractal dimension ³⁹ Energy at different frequency bands in the STFT domain Spectral centroid ⁴⁰ Spectral spread Spectral entropy Spectral rolloff ⁴¹ Mean and Median frequency ⁴² Occupied and power bandwidth ⁴³ Energy in different frequency bands in power spectral density Shannon entropy ⁴⁴ Periodogram slope ⁴⁵	None
Beat-to-Beat	Captures the characteristics of each ICP beat, defined as the period between each	Autocorrelation Energy in different frequency bands in power spectral density Polynomial curve fitting over ICP's beat with degree 4	Mean Maximum Minimum Median SD

Feature Category	Description	Features	Secondary features
	P1 (percussion wave)		1 st , 25 th , 75 th , and 99 th percentiles Mean($diff_1$) SD($diff_1$) Mean($diff_2$) SD($diff_2$)
Beat variability	Captures the characteristics of the fluctuations of time between ICP beats	Root mean square of successive differences Probability of intervals greater than 50 ms or smaller than -50 ms Average beats interval in one minute Mean of the first lag of beats' intervals Histogram-based features from beat intervals ⁴⁶	Mean Maximum Minimum Median SD 1 st , 25 th , 75 th , and 99 th percentiles Mean($diff_1$) SD($diff_1$) Mean($diff_2$) SD($diff_2$)
Window-based	Captures the inherent ICP features using a rolling window with a length of 5 minutes and no overlap	Hurst exponent ³⁶ Energy in different frequency bands in power spectral density Petrosian fractal dimension ^{47,48} Shannon entropy Higuchi fractal dimension ³⁸ Spectral centroid ⁴⁰ Spectral spread Spectral entropy Spectral rolloff ⁴¹ Mean and Median frequency ⁴² Occupied and power bandwidth Energy in different frequency bands in spectral power density Shannon entropy ⁴⁴ Periodogram slope ⁴⁵ Eigenvalues of phase space ³⁴ Phase space nullcline ³⁴	Mean Maximum Minimum Median SD 1 st , 25 th , 75 th , and 99 th percentiles Mean($diff_1$) SD($diff_1$) Mean($diff_2$) SD($diff_2$)

Feature selection and classification

For feature selection and classification tasks, we employed a nested five-fold patient-wise cross-validation scheme. The outer loop partitioned the data into five folds, with four of these folds comprising 80% of the data being utilized for feature selection and model training, while the remaining 20% was utilized for evaluating

the model performance (Supplemental Figure S1). This process was repeated until each fold had served as the test set. The inner loop applied another five-fold cross-validation procedure to the training set to select the most significant features and repeated this process 10 times with different random seeds. In each fold of the inner loop, the top features were selected based on the out-of-bag feature importance^{49,50}. We defined a threshold, referred to as $k1$, as the top 600 features. The threshold for selecting top features is a trade-off between choosing only the most significant features and ensuring we have a sufficiently large set for analysis (i.e., identifying jointly selected features) within our nested cross-validation framework. Our choice of threshold is purely data-driven and was determined through various trials and iterations only on the training data to strike the right balance. The inner five-fold cross-validation loop was repeated 10 times, resulting in 50 sets of $k1$ features. These $k1$ features were then ranked by selection frequency, and based on a threshold, referred to as $k2$, at the 75th percentile, the most frequent ones were chosen as the final selected feature set. For each outer loop iteration, our model was trained to differentiate waveform data from an EVD from an IPM, utilizing the features selected in the inner loop by utilizing 150 bootstrap-aggregated decision trees. Subsequently, the model was evaluated on the hold-out test set. To conduct a comparative analysis, we also employed linear discriminant analysis (LDA) for classification in the outer loop. As a more simple classifier, LDA allowed us to evaluate the robustness of the model and ensure performance was not a result of overfitted decision trees^{47,49,50}. Additionally, for a fair evaluation, as an extra step, we retained the 9 patients with both monitoring types as unseen held-out data and utilized them solely for evaluation purposes.

Evaluation metrics

We employed a comprehensive set of seven metrics to evaluate the model's performance. These included the area under the receiver operating characteristic curve (AU-ROC), sensitivity, specificity, precision, false positive rate, accuracy, and F1 score, where EVDs were classified as the positive class and IPMs as the negative class.

Subgroup analysis

To ensure the reliability of our model, we analyzed its robustness and extracted features by assessing them across various patient subgroups. These subgroups were created using 24 potential clinical confounding variables (Supplemental Tables S1 and S2). Two different strategies were utilized: clustering and propensity

score-based confound-isolating cross-validation schemes, as illustrated in Supplemental Figure S2. These methods allowed us to assess the extent to which our model and features remained robust in the presence of patient variability and potential confounding factors.

Unlike conventional cross-validation, which randomly partitions data into training and testing sets, confound-isolating cross-validation employs various techniques to stratify the data into a series of strata or folds^{35,47}. These methods isolate one or a specific combination of confounding variables within each fold, allowing for a more thorough evaluation of the model's robustness. During the training and testing phases, the model is tested on a fold with a distinct confounding distribution than the training folds, decreasing the probability of spurious correlations influencing the evaluation.

First, we adopted a cluster-based approach in which we employed the *k*-medoids clustering algorithm⁵¹ and iteratively partitioned the data into three, five, and eight clusters with unique clinical phenotypes. Testing on a distinct cluster from those utilized for training enabled us to assess the model's performance or features when accounting for the confounding variable.

Additionally, we used a propensity score-based approach to stratify the data based on a calculated propensity score, representing the likelihood of an ICP waveform belonging to the EVD class given the patient's clinical characteristics. We partitioned the data by propensity score quantile, forming four distinct subgroups employed within the confound-isolating cross-validation process (Supplemental Figure S2).

Results

Patient characteristics

Of 84 patients meeting inclusion criteria, 82 patients with 2667 ICP epochs, totaling 14,109.8 hours of signal available, met the criteria for analysis (Figure 2). 47 patients were monitored with an EVD, 26 with an IPM, and 9 with both (Supplemental Table S1). As mentioned in Figure S1, we utilized the random downsampling approach to balance the number of EVD and IPM samples before training the model. Mean age, GCS total, GCS motor score upon admission, presence and amount of midline shift, decompressive hemicraniectomy, and traumatic brain pathology were similar between groups. Monitor type was significantly associated with institutions due to differing local practices regarding monitor choice at the contributing institutions despite otherwise similar clinical presentation amongst their patients. Additionally,

IPM was used more frequently in patients with unilateral sluggish or nonreactive pupils with the right pupil reactivity, yielding a statistically significant difference between the EVD and IPM groups. Other non-significant univariate differences between patients with an EVD or IPM can be found in Supplemental Table S1.

Model performance

In patient-wise 5-fold cross-validation in the unseen test data using BDT and LDA classifiers, sensitivity, specificity, F1, and AU-ROC exceeded 0.81 (BDT: 0.944 ± 0.038 , LDA: 0.883 ± 0.051), 0.704 (BDT: 0.869 ± 0.097 , LDA: 0.817 ± 0.064), 0.828 (BDT: 0.922 ± 0.032 , LDA: 0.876 ± 0.032), and 0.866 (BDT: 0.932 ± 0.043 , LDA: 0.900 ± 0.016) across all folds, respectively (Tables S3-4).

Adding 11 features based solely on artifactual segments did not meaningfully alter model performance, with no changes in any evaluation metric larger than 0.005 for BDT (Table S10) and 0.015 for LDA (Table S11).

Furthermore, utilizing the BDT (our primary classifier), we obtained a sensitivity of 0.848, specificity of 0.747, F1 score of 0.805, and AU-ROC of 0.826, based on nine patients with both monitoring types as unseen held-out data (Table S12).

Subgroup analyses for rigor and robustness

In cluster-based confound isolation cross-validation, we observed an average sensitivity above 0.947 when classifying with BDT (3-cluster, 0.947 ± 0.017 ; 5-cluster, 0.948 ± 0.038 , 8-cluster, 0.963 ± 0.033 ; figure 4A, table S6) and 0.834 when classifying with LDA (3-cluster, 0.834 ± 0.079 ; 5-cluster, 0.890 ± 0.056 ; 8-cluster, 0.917 ± 0.039 ; figure 4B, table S7). Average specificity was above 0.724 with BDT (3-cluster, 0.724 ± 0.084 ; 5-cluster 0.813 ± 0.160 ; 8-cluster 0.844 ± 0.146 ; figure 4A, table S6) and 0.692 with LDA (3-cluster, 0.692 ± 0.047 ; 5-cluster, 0.772 ± 0.010 ; 8-cluster, 0.813 ± 0.128 ; figure 4B, table S7). F1 scores were above 0.864 for BDT (3-cluster, 0.864 ± 0.071 ; 5-cluster, 0.902 ± 0.074 ; 8-cluster 0.922 ± 0.072 ; figure 4A, table S6) and 0.796 for LDA (3-cluster, 0.796 ± 0.091 ; 5-cluster, 0.858 ± 0.066 ; 8-cluster, 0.874 ± 0.048 ; figure 4B, table S7). Lastly, AU-ROC was greater than 0.852 for BDT (3-cluster, 0.852 ± 0.094 ; 5-cluster, 0.901 ± 0.090 ; 8-cluster, 0.929 ± 0.092 ; figure 4A, table S6) and 0.811 for LDA (3-cluster, 0.811 ± 0.024 ; 5-cluster, 0.888 ± 0.049 ; 8-cluster, 0.919 ± 0.058 ; figure 4B, table S7). Across all metrics, performance increased as the number of clusters increased. Detailed reporting and further metrics are available in the Supplementary Materials.

The worst groups in our cluster analyses, defined as those with sensitivity or specificity below 70%, were significantly less severe in pupil reactivity, total GCS score, and midline shift. The 3-cluster analysis contained one worst group (versus remaining patients: L pupil, $p=0.033$; R pupil, $p=0.039$; GCS, $p < 0.001$; midline shift, $p=0.003$; table 2). There were two worst groups in the 5-cluster (versus remaining patients: L pupil, $p=0.001$ and $p=0.063$; R pupil, $p=0.004$ and $p=0.072$; GCS, $p=0.016$ and $p < 0.001$; midline shift, $p=0.004$ and $p=0.011$) and 8-cluster analyses (versus remaining patients: L pupil, $p=0.095$ and $p=0.086$; R pupil, $p=0.111$ and $p=0.097$; GCS, $p=0.800$ and $p < 0.001$; midline shift, $p=0.004$ and $p=0.022$; table 2).

Table 2: Comparison of four potential confounding factors between the worst-group patients against the remaining patients in two subgroup analyses (p-values)

Confounder Subgroup Strategy	Worst Group vs. Rest ^a	Left pupil Reactivity ^b	Right pupil Reactivity ^b	Total GCS Score	Midline Shift
Confounder cluster strata	Cluster 3	0.033	0.039	<0.001	0.003
	Cluster 5 ^c	0.001	0.004	0.016	0.004
	Cluster 5 ^c	0.063	0.072	<0.001	0.011
	Cluster 8 ^c	0.095	0.111	0.800	0.004
	Cluster 8 ^c	0.086	0.097	<0.001	0.022
Propensity score strata	BDT	0.905	0.788	0.593	0.649
	LDA	0.572	0.447	0.384	0.741

^a Worst-group patients were defined as any of those with either sensitivity or specificity below 70%.

^b At admission.

^c Cluster 5 and Cluster 8, both contained the worst-group.

Using propensity score-based cross-validation, we found a mean sensitivity of 0.954 ± 0.048 , mean specificity of 0.788 ± 0.143 , mean F1 score of 0.871 ± 0.090 , and mean AU-ROC of 0.904 ± 0.120 when classifying with BDT (figure 4C). When using LDA, mean sensitivity was 0.915 ± 0.037 , mean specificity was 0.761 ± 0.090 , mean F1 score was 0.887 ± 0.032 , and mean AU-ROC was 0.898 ± 0.043 (figure 4D). A complete list of results for the subgroup analysis can be found in the Supplementary Materials (tables S6-9).

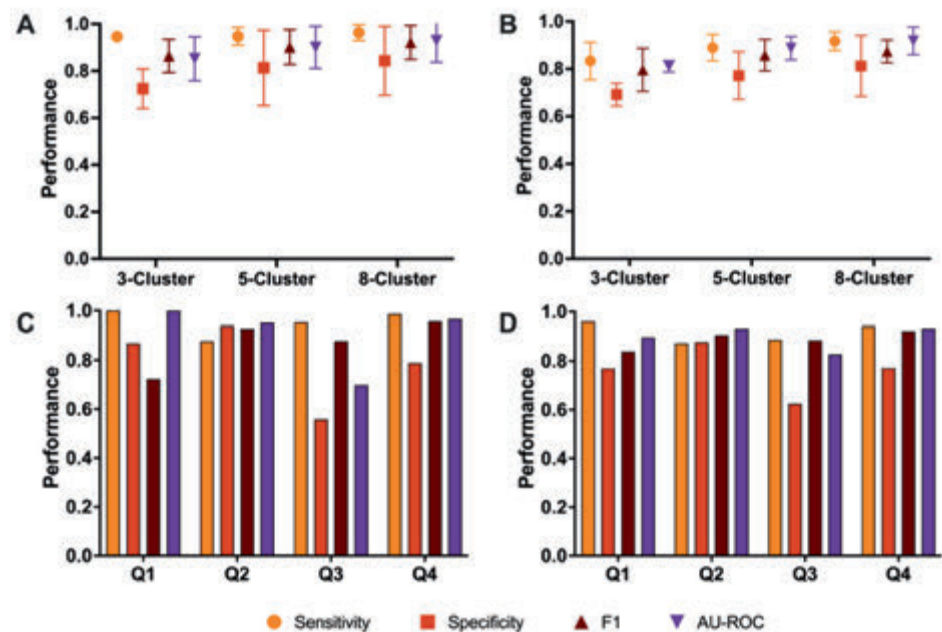


Figure 4: Model performance of confounder-isolating cross-validation for subgroup analysis. A) Evaluation metrics from cluster-based approach to confounder-isolation with bagged decision tree classifier. B) Evaluation metrics from cluster-based approach to confounder-isolation with linear discriminant analysis. C) Evaluation metrics from propensity score-based approach to confounder-isolation with bagged decision tree classifier (Fold 1: 0-0.478, Fold 2: 0.478-0.691, Fold 3: 0.691-0.835, Fold 4: 0.835-1). D) Evaluation metrics from propensity score-based approach to confounder-isolation with linear discriminant analysis (Fold 1: 0-0.228, Fold 2: 0.228-0.582, Fold 3: 0.582-0.927, Fold 4: 0.927-1).

Feature selection

Examining standard, cluster-based, and propensity score-based cross-validation schemes, feature selection converged on 62 features jointly selected by both BDT and LDA methods across all evaluated cross-validation schemes. Supplemental figure S4 articulates observed differences between EVDs and IPMs in boxplots and kernel density plots for a sample of the selected features, which demonstrate greater complexity and volatility in EVDs versus consistency and stability in IPMs as key differentiators. Features selected were Higuchi fractal dimension short-time-frequency energy band (11-16 Hz), spectral centroid, summation of power spectral density in 20-40 Hz, autocorrelation, coefficients of polynomial fit on ICP beat, Hurst exponent, eigenvalues of phase space³⁴ (Supplemental figure S4A-B), coefficients of an autoregressive model with order 4 (Supplemental figure S4C-D), Petrosian

fractal dimension, Shannon entropy, singular values of phase space, spectral entropy, spectral roll-off, mean normalized frequency of the power spectrum, spectral slope, and phase space nullcline³⁴ (Supplemental Figure S4). For a complete list of features, including their definitions, we refer to Supplemental Table S2.

Discussion

This proof-of-concept study developed an ML model to accurately and robustly classify ICP waveform features distinct to ICP monitor type and generate automated metadata designating the ICP monitor as originating from an EVD or IPM. Our model was able to achieve an average F1 score of 0.922 (± 0.032) and 0.876 (± 0.032) using BDT and LDA, respectively, in a patient-wise 5-fold cross-validation. This demonstrates the promise of our approach, which is further reflected in all computed evaluation metrics and confusion matrices. These results indicate that there are meaningful differences in ICP waveforms from EVDs and IPMs and that the data source should be carefully considered and accounted for as a potential source of bias in future ICP waveform analysis. Additionally, where this contextual metadata is unavailable, our model may be trusted to generate it through analysis of ICP waveform characteristics with further validation on broader datasets.

To our knowledge, this is the first effort to detect the source of ICP waveforms, i.e., EVDs versus IPMs, from signatures of their recordings. As we have demonstrated, meaningful differences in ICP waveforms from EVDs and IPMs indicate that the ICP monitor type should be carefully considered and accounted for in future ICP waveform analyses, as it may cause bias and influence clinical decision-making. The developed framework illustrates that features of ICP monitor types are identifiable, and such features can be potentially informative for various analyses of ICP waveforms.

Lessons learned from ICP features

We identified 62 distinct characteristics of ICP that consistently emerged as the most significant features, regardless of the classifier, cross-validation scheme, or subgroup analysis employed. These invariant features exhibit promising potential for other ICP analyses. Notably, out of these 62 features, 44 are window-based. In contrast, only 15 and 4 are chosen from beat-to-beat and general features, respectively. This observation implies that the type of monitor significantly

influences trends within the ICP waveform over time, as opposed to each cycle of the ICP pulse or the overall ICP signal. Additionally, the most frequently selected secondary features (24 out of 62) were lag (first and second differences) features, which supports this conclusion and underscores the efficacy of lag features in ICP analysis. As adding features based on artifacts did not meaningfully alter model performance, we conclude that our data preprocessing did not eliminate important information for accurate classification.

We observed a trend of higher complexity level in EVDs versus IPMs. Differences in the eigenvalues of the covariance matrix of reconstructed phase space of the ICP waveform (Supplemental figure S4A-B) indicate unique underlying dynamics and patterns present in the time series data, especially in such cases of complex and nonlinear systems⁵². We speculate that this increased complexity in EVD signals arises from either the use of a fluid column to generate the signal and resulting debris or transmitted vibrations, lack of discretization at an intermediate ICP monitor breakout box before being relayed to the patient's physiologic monitor, or alternatively the complex milieu of the ventricular system, in which CSF pulsations in the lateral ventricles have phasic coupling between adjacent compartments such as fourth ventricle or circle of Willis arterial signals⁵³.

EVDs exhibit a higher standard deviation of autoregressive parameters (Supplemental figure S4C-D), which capture the temporal patterns and dependencies within the time series⁵⁴, i.e., how consistent or uniform the signal is. This suggests that EVDs have more significant fluctuations and vary widely from one window to another than IPMs, which demonstrate more stability and consistency in the model parameters. This indicates greater complexity or non-linearity in the underlying process generating EVDs compared to IPMs. Similarly, the distances among the intersected nullclines in the phase space (Supplemental figure S4E-F), representing three clusters exhibiting distinct characteristics within each ICP monitor type³⁴, are greater in EVDs compared to IPMs, suggesting more volatility. This may arise from various factors, including abrupt shifts in underlying ICP patterns, outliers, or alterations in the system's dynamics, all of which could be attributed to clamping and drainage of the EVD. Unfortunately, due to the historical nature of this dataset, details on EVD probes and institutional clamping protocols were not available to include in this analysis.

Lessons learned from subgroup analysis

To ensure the reliability of our model, we have utilized subgroup analysis using clustering and propensity scores. In our cluster analysis, we found that increasing the number of clusters led to improved performance across all evaluation metrics. This improvement can be attributed to providing more clusters for training, allowing the model to learn more efficiently. This observation was held for BDT and LDA models (7.7% and 10.8% improvement in AU-ROC of BDT and LDA, respectively; Tables S6 and S7). However, we also found that certain patients consistently performed relatively poorly (the worst-group, defined as those with either sensitivity or specificity below 70%) in the unseen test fold, regardless of the chosen model or number of clusters. Upon closer inspection, we discovered that these patients had significantly less severe total GCS scores, pupil reactivity upon admission, and amount of midline shift in most clustering configurations (Table 2). This may suggest that our identifiable ICP features are more apparent when ICP is abnormal and that our model may be limited in application to less severe patients, although no significant changes were identified between the potential confounders of the worst-group and other patients in the propensity score-based analysis (Table 2). Importantly, our model performance was strongest for the most critical populations compared to the marginally weaker performance in clusters of less severe patients, allowing for reasonable applicability of our findings in the clinical areas of greatest need. However, future studies should examine this finding in a larger cohort, including other diagnoses of ABI beyond TBI, to increase generalizability for this proof-of-concept study.

Sample size to the number of features ratio

In this study, the number of extracted features is relatively larger than that of the sample (patients). In classical statistics, this implies overfitting. However, we would like to provide the following perspectives: (1) *Tree-based ensembles are highly adaptive to "large p , small n " problems*: The robustness of tree-based ensembles (our primary choice for classifier) has been proved empirically when the number of samples is lower than the features⁵⁵. Such scenarios are common in studies focused on genomic data analysis, where the number of features (genes) often exceeds the number of samples (patients). (2) *Proper evaluation matters*: To ensure our findings' validity and avoid overfitting, we not only employed nested cross-validation but also utilized confound-isolating cross-validation to detect potential biases, albeit at the expense of reducing our performance. (3) *A naive classifier shows the true*

discriminant power of features: In addition to using bagged decision trees, we repeated all the analyses by utilizing a simple classifier, i.e., LDA. This choice was made to underscore the discriminative capabilities of features rather than relying solely on advanced machine learning techniques. Thus, our proof-of-concept framework provides a robust performance despite the lower number of features.

Limitations

Our analysis confirms that decision tree-based models have demonstrated their considerable power and superiority in various applications^{56,57}. We also employed a simple LDA classifier to showcase the discriminating ability of our features. Although LDA exhibited relatively lower performance than BDT, it still yielded promising results on average (0.9 AU-ROC and F1 score 0.876) and showed smaller decreases in performance in subgroup analyses (5.6% reduction in specificity and 0.2% decrease in AU-ROC versus 8.1% and 2.8% for BDT). Further, in the worst-performing group, LDA significantly outperformed BDT in terms of specificity and AU-ROC (figure 3A-B). These results demonstrate that BDT is prone to overfitting despite being more sophisticated than LDA, suggesting that bias may contribute to the model performance and highlighting the necessity of conducting further subgroup analyses with external datasets to robustly evaluate the model's performance. In addition, the developed ML model in this study is based solely on patients with TBI. Further research is necessary to investigate if the developed framework can also be used in patients with diagnoses beyond TBI.

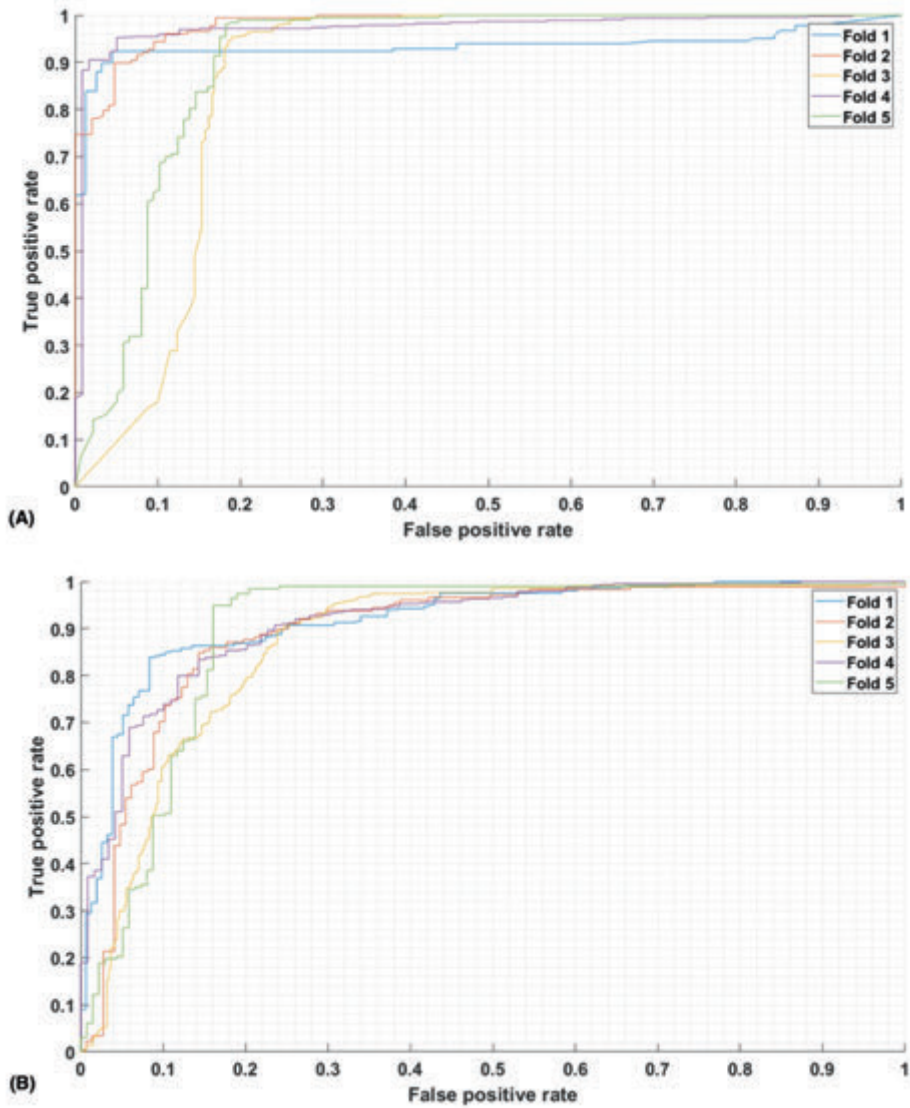


Figure 3: Model performance of ICP waveform-derived features. A) AU-ROC curve with bootstrap-aggregated decision trees used for classification. B) AU-ROC curve with the linear discriminant analysis (LDA) used for classification. Note that the LDA method demonstrated better consistency of the AU-ROC across all folds (better worst-fold performance).

Conclusion

The developed proof-of-concept ML framework accurately and robustly identified features of ICP derived from an EVD or IPM and was able to generate automated metadata designating the provenance of an ICP monitor based on monitor type. Accuracy was evident across procedures designed to ensure robustness through confound-isolating cross-validation. The developed framework can impute metadata from retrospective ICP datasets lacking data for monitor type, enabling future waveform analysis controlling for the bias introduced by the monitor type. This is especially critical for multicenter collaboration across datasets in which labels have not been harmonized or remain unknown due to the lack of metadata during data collection. Our model can also consequently summarize confounding by indication for placing one monitor type versus another, which may itself account for differences in treatment response or outcomes. The invariant ICP features may additionally be leveraged in future strategies implementing ML for classification and prediction.

References

1. Berlin T, Murray-Krezan C, Yonas H. Comparison of parenchymal and ventricular intracranial pressure readings utilizing a novel multi-parameter intracranial access system. *SpringerPlus* 2015;4(1):1-8.
2. Hagel S, Bruns T, Pletz M, Engel C, Kalff R, Ewald C. External ventricular drain infections: risk factors and outcome. *Interdisciplinary perspectives on infectious diseases* 2014;2014.
3. Carney N, Totten AM, O'Reilly C, et al. Guidelines for the Management of Severe Traumatic Brain Injury, Fourth Edition. *Neurosurgery* 2017;80(1):6-15. DOI: 10.1227/neu.0000000000001432.
4. Kirkness CJ, Mitchell PH, Burr RL, March KS, Newell DW. Intracranial pressure waveform analysis: clinical and research implications. *J Neurosci Nurs* 2000;32(5):271-7. (In eng). DOI: 10.1097/01376517-200010000-00007.
5. Czosnyka M, Pickard JD. Monitoring and interpretation of intracranial pressure. *J Neurol Neurosurg Psychiatry* 2004;75(6):813-21. (In eng). DOI: 10.1136/jnnp.2003.033126.
6. Dai H, Jia X, Pahren L, Lee J, Foreman B. Intracranial pressure monitoring signals after traumatic brain injury: a narrative overview and conceptual data science framework. *Frontiers in neurology* 2020;11:959.
7. Megjhani M, Alkhachroum A, Terilli K, et al. An active learning framework for enhancing identification of non-artifactual intracranial pressure waveforms. *Physiological measurement* 2019;40(1):015002.
8. Güiza F, Depreitere B, Piper I, Van den Berghe G, Meyfroidt G. Novel methods to predict increased intracranial pressure during intensive care and long-term neurologic outcome after traumatic brain injury: development and validation in a multicenter dataset. *Critical care medicine* 2013;41(2):554-564.
9. Hüser M, Kündig A, Karlen W, De Luca V, Jaggi M. Forecasting intracranial hypertension using multi-scale waveform metrics. *Physiological measurement* 2020;41(1):014001.
10. Quachtran B, Hamilton R, Scalzo F. Detection of intracranial hypertension using deep learning. 2016 23rd international conference on pattern recognition (ICPR): IEEE; 2016:2491-2496.
11. Schweingruber N, Mader MMD, Wiehe A, et al. A recurrent machine learning model predicts intracranial hypertension in neurointensive care patients. *Brain* 2022;145(8):2910-2919.
12. Ye G, Balasubramanian V, Li JK, Kaya M. Machine learning-based continuous intracranial pressure prediction for traumatic injury patients. *IEEE Journal of Translational Engineering in Health and Medicine* 2022;10:1-8.

13. Megjhani M, Terilli K, Kwon SB, et al. Automatic identification of intracranial pressure waveform during external ventricular drainage clamping: segmentation via wavelet analysis. *Physiological Measurement* 2023.
14. Hu X, Xu P, Scalzo F, Vespa P, Bergsneider M. Morphological clustering and analysis of continuous intracranial pressure. *IEEE Transactions on Biomedical Engineering* 2008;56(3):696-705.
15. Brean A, Eide PK, Stubhaug A. Comparison of intracranial pressure measured simultaneously within the brain parenchyma and cerebral ventricles. *J Clin Monit Comput* 2006;20(6):411-4. (In eng). DOI: 10.1007/s10877-006-9047-7.
16. Harary M, Dolmans RGF, Gormley WB. Intracranial Pressure Monitoring-Review and Avenues for Development. *Sensors (Basel)* 2018;18(2) (In eng). DOI: 10.3390/s18020465.
17. Schimpf MM. Diagnosing increased intracranial pressure. *J Trauma Nurs* 2012;19(3):160-7. (In eng). DOI: 10.1097/JTN.0b013e318261cfb4.
18. Muralidharan R. External ventricular drains: Management and complications. *Surg Neurol Int* 2015;6(Suppl 6):S271-4. (In eng). DOI: 10.4103/2152-7806.157620.
19. Slazinski T, Anderson T, Cattell E, et al. Nursing management of the patient undergoing intracranial pressure monitoring, external ventricular drainage, or lumbar drainage. *Journal of Neuroscience Nursing* 2011:233+.
20. Zhong J, Dujovny M, Park HK, Perez E, Perlin AR, Diaz FG. Advances in ICP monitoring techniques. *Neurol Res* 2003;25(4):339-50. (In eng). DOI: 10.1179/016164103101201661.
21. Lescot T, Reina V, Le Manach Y, et al. In vivo accuracy of two intraparenchymal intracranial pressure monitors. *Intensive Care Med* 2011;37(5):875-9. (In eng). DOI: 10.1007/s00134-011-2182-8.
22. Foreman B, Lissak IA, Kamireddi N, Moberg D, Rosenthal ES. Challenges and Opportunities in Multimodal Monitoring and Data Analytics in Traumatic Brain Injury. *Curr Neurol Neurosci Rep* 2021;21(3):6. (In eng). DOI: 10.1007/s11910-021-01098-y.
23. Scalzo F, Hamilton R, Asgari S, Kim S, Hu X. Intracranial hypertension prediction using extremely randomized decision trees. *Medical engineering & physics* 2012;34(8):1058-1065.
24. Scalzo F, Liebeskind D, Hu X. Reducing false intracranial pressure alarms using morphological waveform features. *IEEE Transactions on biomedical engineering* 2012;60(1):235-239.
25. Wijayatunga P, Koskinen L-OD, Sundström N. Probabilistic prediction of increased intracranial pressure in patients with severe traumatic brain injury. *Scientific Reports* 2022;12(1):9600.
26. Krishnamoorthy V, Temkin N, Barber J, et al. Association of Early Multiple Organ Dysfunction With Clinical and Functional Outcomes Over the Year Following

- Traumatic Brain Injury: A Transforming Research and Clinical Knowledge in Traumatic Brain Injury Study. *Critical Care Medicine* 2021;49(10):1769-1778. DOI: 10.1097/ccm.0000000000005055.
27. McCrea MA, Giacino JT, Barber J, et al. Functional Outcomes Over the First Year After Moderate to Severe Traumatic Brain Injury in the Prospective, Longitudinal TRACK-TBI Study. *JAMA Neurology* 2021;78(8):982-992. DOI: 10.1001/jamaneurol.2021.2043.
 28. Yue JK, Vassar MJ, Lingsma HF, et al. Transforming research and clinical knowledge in traumatic brain injury pilot: multicenter implementation of the common data elements for traumatic brain injury. *J Neurotrauma* 2013;30(22):1831-44. (In eng). DOI: 10.1089/neu.2013.2970.
 29. Harris FJ. Multirate signal processing for communication systems: CRC Press, 2022.
 30. Kesić S, Spasić SZ. Application of Higuchi's fractal dimension from basic to clinical neurophysiology: A review. *Computer methods and programs in biomedicine* 2016;133:55-70.
 31. Kulkarni N, Bairagi V. Role of Different Features in Diagnosis of Alzheimer Disease. EEG-Based Diagnosis of Alzheimer Disease: A Review and Novel Approaches for Feature Extraction and Classification Techniques. : Elsevier Science; 2018:37-46.
 32. Tobore I, Li J, Kandwal A, Yuhang L, Nie Z, Wang L. Statistical and spectral analysis of ECG signal towards achieving non-invasive blood glucose monitoring. *BMC Medical Informatics and Decision Making* 2019;19(6):266. DOI: 10.1186/s12911-019-0959-9.
 33. Torres-García A, Mendoza-Montoya O, Molinas M, Antelis J, Moctezuma L, Hernández-Del-Toro T. Pre-processing and feature extraction. *Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications* 2022:59-91.
 34. Zabihi M, Kiranyaz S, Jantti V, Lipping T, Gabbouj M. Patient-Specific Seizure Detection Using Nonlinear Dynamics and Nullclines. *IEEE J Biomed Health Inform* 2020;24(2):543-555. (In eng). DOI: 10.1109/jbhi.2019.2906400.
 35. Zabihi M, Rubin DB, Ack SE, et al. Resting-State Electroencephalography for Continuous, Passive Prediction of Coma Recovery After Acute Brain Injury. *bioRxiv* 2022:2022.09. 30.510334.
 36. Torres-García AA, Mendoza-Montoya O, Molinas M, Antelis JM, Moctezuma LA, Hernández-Del-Toro T. Pre-processing and feature extraction. In: *Biosignal Processing and Classification Using Computational Learning and Intelligence: Principles, Algorithms, and Applications* 2022:Pages 59-91.
 37. Kaspar F, Schuster HG. Easily calculable measure for the complexity of spatiotemporal patterns. *Phys Rev A Gen Phys* 1987;36(2):842-848. (In eng). DOI: 10.1103/physreva.36.842.

38. Kesić S, Spasić SZ. Application of Higuchi's fractal dimension from basic to clinical neurophysiology: A review. *Comput Methods Programs Biomed* 2016;133:55-70. (In eng). DOI: 10.1016/j.cmpb.2016.05.014.
39. Wijayanto I, Hartanto R, Nugroho HA. Higuchi and Katz Fractal Dimension for Detecting Interictal and Ictal State in Electroencephalogram Signal. 11th International Conference on Information Technology and Electrical Engineering (ICITEE), Pattaya, Thailand 2019:pp. 1-6. DOI: 10.1109/ICITEED.2019.8929940.
40. Kulkarni N, Bairagi V. Role of Different Features in Diagnosis of Alzheimer Disease. In: *EEG-Based Diagnosis of Alzheimer Disease: A Review and Novel Approaches for Feature Extraction and Classification Techniques*. 2018:Pages 37-46.
41. Tobore I, Li J, Kandwal A, Yuhang L, Nie Z, Wang L. Statistical and spectral analysis of ECG signal towards achieving non-invasive blood glucose monitoring. *BMC Med Inform Decis Mak* 2019;19(Suppl 6):266. (In eng). DOI: 10.1186/s12911-019-0959-9.
42. Phinyomark A, Thongpanja S, Hu H, Phukpattaranont P, Limsakul C. The Usefulness of Mean and Median Frequencies in Electromyography Analysis. In: *Computational Intelligence in Electromyography Analysis* 2012.
43. Mert A. ECG feature extraction based on the bandwidth properties of variational mode decomposition. *Physiol Meas* 2016;37(4):530-43. (In eng). DOI: 10.1088/0967-3334/37/4/530.
44. Liang Z, Wang Y, Sun X, et al. EEG entropy measures in anesthesia. *Front Comput Neurosci* 2015;9:16. (In eng). DOI: 10.3389/fncom.2015.00016.
45. Lendner JD, Helfrich RF, Mander BA, et al. An electrophysiological marker of arousal level in humans. *Elife* 2020;9 (In eng). DOI: 10.7554/eLife.55092.
46. Oster J, Behar J, Colloca R, Li Q, Li Q, Clifford GD. Open source Java-based ECG analysis software and Android app for Atrial Fibrillation screening. *Computing in Cardiology* 2013:pp. 731-734.
47. Chyzyk D, Varoquaux G, Milham M, Thirion B. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience* 2022;11.
48. Zabihi M, Rubin DB, Ack SE, et al. Resting-State Electroencephalography for Continuous, Passive Prediction of Coma Recovery After Acute Brain Injury. *bioRxiv* 2022:202209 30510334.
49. Loh W-Y. Regression tress with unbiased variable selection and interaction detection. *Statistica sinica* 2002:361-386.
50. Nguyen T-T, Huang JZ, Nguyen TT. Unbiased feature selection in learning random forests for high-dimensional data. *The Scientific World Journal* 2015;2015.
51. Kaufman L, Rousseeuw PJ. *Finding groups in data: an introduction to cluster analysis*: John Wiley & Sons, 2009.

52. Takens F. Detecting strange attractors in turbulence Lecture Notes in Mathematics. Dynamical Systems of Turbulence 1981:366-381. (<https://cir.nii.ac.jp/crid/1571135649040545152>).
53. Butler WE, Agarwalla PK, Codd P. CSF in the ventricles of the brain behaves as a relay medium for arteriovenous pulse wave phase coupling. PLoS One 2017;12(11):e0181025. (In eng). DOI: 10.1371/journal.pone.0181025.
54. Kay S. Spectral estimation. Advanced topics in signal processing 1988:58-122.
55. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics 2012;99(6):323-9. (In eng). DOI: 10.1016/j.ygeno.2012.04.003.
56. Kotu V, Deshpande B. Data science: concepts and practice: Morgan Kaufmann, 2018.
57. Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Information Fusion 2022;81:84-90. DOI: <https://doi.org/10.1016/j.inffus.2021.11.011>.

Supplementary Materials

Table of Contents

Methods

Population and inclusion criteria

Table S1

Feature selection and classification

Figure S1.

Subgroup analysis

Table S2

Figure S2

Management of artifacts

Results

Conventional cross-validation

Table S3

Table S4

Table S5

Misclassification samples

Figure S3

Subgroup analysis (cluster analysis)

Table S6

Table S7

Subgroup Analysis by Propensity Score

Table S8

Table S9

Subgroup Analysis by Potential Confounders

Table S10

Table S11.

Feature selection

Figure S4

Held-out evaluation

Methods

Population and inclusion criteria

Table S1: Baseline characteristics

	EVD monitor	IPM monitor	Both monitors	P-value (EVD vs. IPM)
Patients, n	47	26	9	
ICP hours total, mean (SD)	10,423.2 (57.2)	2,740.3 (15.0)	5,046.0 (27.7)	<0.001
ICP hours per patient, mean (SD)	7.80 (23.02)	4.61 (0.87)	5.02 (0.58)	0.478
Age, mean (SD)	45.34 (16.80)	46.69 (17.78)	50.22 (13.35)	0.748
Female sex, n (%)	8 (17.0)	7 (26.9)	0 (0.0)	0.484
BMI, n (%)				0.415
Normal (18.5-24.9)	8 (21.1)	8 (33.3)	2 (22.2)	
Underweight (<18.5)	1 (2.6)	0 (0.0)	0 (0.0)	
Overweight (24.9-29.9)	21 (55.3)	9 (37.5)	6 (66.7)	
Obese (>29.9)	8 (21.1)	7 (29.2)	1 (11.1)	
Race, n (%)				0.610
White	39 (84.8)	24 (92.3)	5 (55.6)	
Black	4 (8.7)	2 (7.7)	3 (33.3)	
Asian	2 (4.3)	0 (0.0)	0 (0.0)	
Native Hawaiian/Pacific Islander	0 (0.0)	0 (0.0)	1 (11.1)	
Mixed Race	1 (2.2)	0 (0.0)	0 (0.0)	
Ethnicity, n (%)				0.480
Hispanic	14 (29.8)	5 (19.2)	0 (0.0)	
Non-Hispanic	33 (70.2)	21 (80.8)	8 (100.0)	
Institution, n (%)				<0.001
Hospital A	14 (29.8)	0 (0.0)	0 (0.0)	
Hospital B	1 (2.1)	0 (0.0)	0 (0.0)	
Hospital C	22 (46.8)	5 (19.2)	7 (77.8)	
Hospital D	0 (0.0)	13 (50.0)	2 (22.2)	
Hospital E	10 (21.3)	8 (30.8)	0 (0.0)	
ICP monitor source, n (%)				<0.001
Natus Camino	0 (0.0)	8 (30.8)	0 (0.0)	
Philips Intellivue	47 (100.0)	18 (69.2)	9 (100.0)	
GCS score, median [IQR]				
Eye	1.00 [1.00, 3.00]	1.00 [1.00, 1.00]	1.00 [1.00, 3.00]	0.890
Verbal	1.00 [1.00, 2.00]	1.00 [1.00, 2.00]	2.00 [1.00, 2.25]	0.190
Motor	1.00 [1.00, 5.00]	1.00 [1.00, 4.75]	3.50 [2.50, 5.00]	0.780

	EVD monitor	IPM monitor	Both monitors	P-value (EVD vs. IPM)
Total	3.50 [3.00, 10.0]	3.00 [3.00, 7.00]	7.50 [5.25, 9.25]	0.539
Left pupil reactivity, n (%)				0.074
Brisk	23 (53.5)	6 (25.0)	5 (83.3)	
Sluggish	9 (20.9)	9 (37.5)	1 (16.7)	
Non-reactive	11 (25.6)	9 (37.5)	0 (0.0)	
Right pupil reactivity, n (%)				0.022
Brisk	24 (57.1)	5 (21.7)	5 (14.4)	
Sluggish	8 (19.0)	9 (39.1)	1 (14.3)	
Non-reactive	10 (23.8)	9 (39.1)	1 (14.3)	
Hypoxia upon ED arrival, n (%)	3 (8.1)	5 (22.7)	1 (14.3)	0.233
Hypotension upon ED arrival, n (%)	4 (11.4)	1 (4.3)	0 (0.0)	0.644
Decompressive hemicraniectomy, n (%)	22 (46.8)	10 (38.5)	3 (33.3)	0.658
Marshall CT score, n (%)				0.939
Category II	14 (31.8)	8 (32.0)	4 (44.4)	
Category III	3 (6.8)	3 (12.0)	1 (11.1)	
Category IV	2 (4.5)	1 (4.0)	0 (0.0)	
Category V	24 (54.5)	12 (48.0)	4 (44.4)	
Category VI	1 (2.3)	1 (4.0)	0 (0.0)	
SDH, n (%)	35 (79.5)	22 (88.0)	7 (77.8)	0.575
EDH, n (%)	13 (29.5)	6 (24.0)	2 (22.2)	0.830
SAH, n (%)	40 (90.9)	24 (96.0)	8 (88.9)	0.763
Midline shift, n (%)	22 (50.0)	14 (56.0)	0 (0.0)	0.819
Midline shift, mm (median [IQR])	1.50 [0.00, 5.25]	4.00 [0.00, 9.00]	0.00 [0.00, 0.00]	0.389
LOC, n (%)				0.364
No	0 (0.0)	1 (4.3)	0 (0.0)	
Yes	39 (92.9)	21 (91.3)	8 (88.9)	
Suspected	3 (7.1)	1 (4.3)	1 (11.1)	
Acceleration-deceleration event, n (%)	20 (42.6)	9 (34.6)	5 (55.6)	0.679

EVD: external ventricular drain; IPM: intraparenchymal monitor; ICP: intracranial pressure; BMI: body mass index; SD: standard deviation; GCS: Glasgow coma scale; IQR: interquartile range; ED: emergency department; CT: computed tomography; EDH: epidural hemorrhage; LOC: loss of consciousness; SAH: subarachnoid hemorrhage; SDH: subdural hemorrhage.

Feature selection and classification

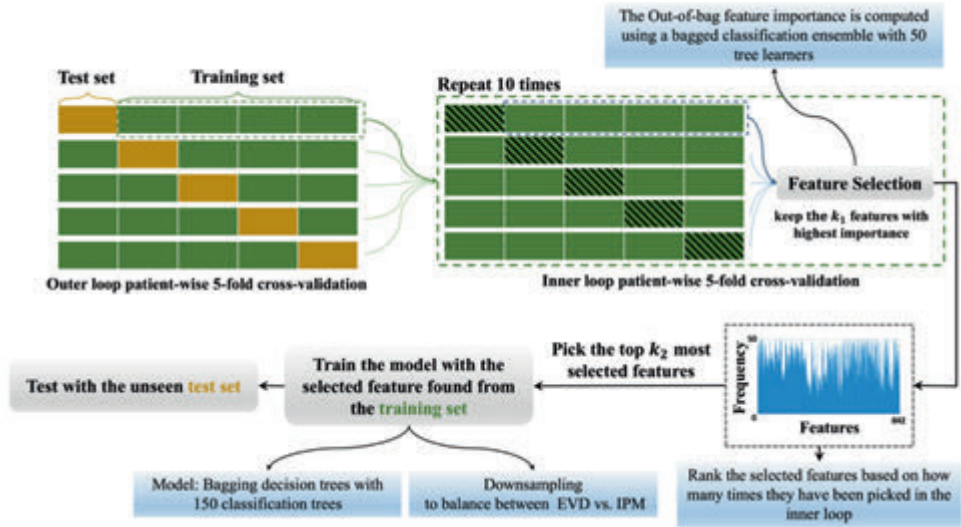


Figure S1: Schematic of model development. The outer loop partitioned the data into five folds, with four of these folds comprising 80% of the data being utilized for feature selection and model training, while the remaining 20% was utilized for evaluating the model performance. This process was repeated until each fold had served as the test set. The inner loop applied another five-fold cross-validation procedure to the training set to select the most significant features. In each fold of the inner loop, the top 600 features (referred to as k_1) were selected based on the out-of-bag feature importance. The inner five-fold cross-validation loop was repeated 10 times, resulting in 50 sets of k_1 features. These k_1 features were then ranked by selection frequency, and based on a 75th percentile threshold, referred to as k_2 , chosen as the final selected feature set. For each outer loop iteration, our model was trained to differentiate waveform data from an EVD from an IPM, using the features selected in the inner loop by utilizing 150 bootstrap-aggregated decision trees. Subsequently, the model was evaluated on the hold-out test set. To conduct a comparative analysis, we also employed linear discriminant analysis (LDA) for classification in the outer loop.

*Subgroup analysis***Table S2.** List of potential confounders and their dichotomization.

Potential Confounder	Dichotomized
Age	2 categories; ≤ 45 years and > 45 years
Sex	2 categories; male and female
Race	7 categories; White, Black, Asian, native Hawaiian/Pacific Islander, Mixed Race, Indian, Alaska Native/Inuit
Ethnicity	2 categories; Hispanic and non-Hispanic
BMI	4 categories; underweight (< 18.5), normal ($18.5-24.9$), overweight ($24.9-29.9$), obese (> 29.9)
Acceleration-deceleration injury	2 categories; yes & no
Loss of consciousness	3 categories; yes, no & suspected
Admitting hospital	5 categories; A, B, C, D & E
Presence of SDH on CT	2 categories; yes & no
Presence of SAH on CT	2 categories; yes & no
Presence of midline shift	2 categories; yes & no
Amount of midline shift	3 categories; 0, 1-8, > 8
Marshall CT score	6 categories; category I through category VI
Hypoxia upon admission	2 categories; yes & no
Hypotension upon admission	2 categories; yes & no
Left pupil reactivity upon admission	3 categories; brisk, sluggish & non-reactive
Right pupil reactivity upon admission	3 categories; brisk, sluggish & non-reactive
GCS total score upon admission	3 categories; severe (3-8), moderate (9-12), mild (13-15)
GCS motor score upon admission	6 categories; 1 through 6
GCS verbal score upon admission	5 categories; 1 through 5
GCS eye score upon admission	4 categories; 1 through 4
Computer type	2 categories; Philips Intellivue & Natus Camino
ICP signal duration	4 categories (quartiles); Q1 (0 - 105.07h), Q1-Q2 (105.07h-184.53h), Q2-Q3 (184.53h - 405.7h), $> Q3$ ($> 405.7h$)

BMI: body mass index; CT; computed tomography; GCS: Glasgow coma scale; ICP: intracranial pressure; SAH: subarachnoid hemorrhage; SDH: subdural hemorrhage.

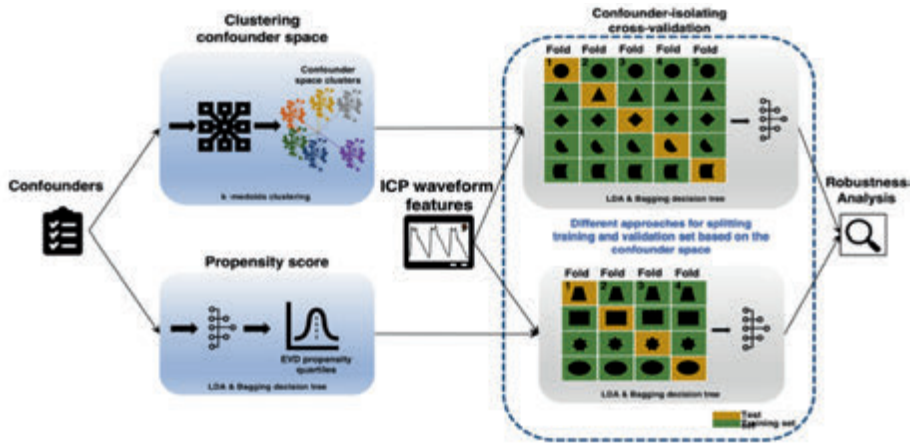


Figure S2: Procedures to ensure rigor and robustness included cluster- and propensity score-based confound-isolating cross-validation schemes. The cluster-based approach (top) employed the k -medoids clustering algorithm and iteratively partitioned the data into three, five, and eight clusters with unique clinical phenotypes. Testing on a clinically distinct cluster from the cluster utilized for training enabled us to assess the model's performance or features when accounting for confounding variables. In the propensity score-based approach (bottom), the data was stratified based on the likelihood, or propensity, of an ICP waveform belonging to the EVD class given the patient's clinical characteristics. The data was partitioned by the highest to lowest quartile of propensity for one monitor type, forming four distinct subgroups employed within the confound-isolating cross-validation process.

Management of artifacts

The primary focus of this study was to analyze the impact of the ICP monitor on the ICP waveforms rather than the artifacts it produces. However, removing these artifacts from our analysis could potentially affect our ability to accurately classify the type of ICP monitor being used (e.g., EVD or IPMs). This is because these artifacts may contain useful information that has not yet been captured by the 842 features that have been extracted so far. To test this hypothesis, we have extracted 11 features based solely on artifacts. These features were extracted from segments of signals detected as artifacts, hereafter referred to as "artifactual segments," based on our developed artifact detection algorithm (as described in the manuscript). The features include the numbers, maximum, and minimum values of artifactual segments, as well as the 25th, 50th, 75th, and 97th percentiles of Shannon entropy values and the lengths of artifactual segments.

We utilized a combination of artifact-based data and our extensive list of 842 features in our discussed feature selection and classification approach. This aimed to evaluate the contribution of artifact segments in distinguishing between EVD and IPM.

Results

Conventional cross-validation

Table S3. Patient-wise 5-fold cross-validation results over the unseen test data using bagged decision trees.

Performance Metric ^a	Fold					Mean±SD
	1	2	3	4	5	
Sensitivity	0.902	0.916	0.995	0.922	0.985	0.944±0.038
Specificity	0.955	0.925	0.706	0.950	0.810	0.869±0.097
Precision	0.979	0.937	0.757	0.986	0.882	0.908±0.084
False positive rate	0.045	0.075	0.294	0.050	0.190	0.131±0.097
Accuracy	0.918	0.92	0.845	0.928	0.913	0.905±0.030
F1	0.939	0.926	0.860	0.953	0.930	0.922±0.032
AU-ROC	0.935	0.981	0.866	0.974	0.907	0.932±0.043

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Table S4. Patient-wise 5-fold cross-validation results over the unseen test data using linear discriminant analysis.

Performance Metric ^a	Fold					Mean±SD
	1	2	3	4	5	
Sensitivity	0.852	0.871	0.933	0.810	0.949	0.883±0.051
Specificity	0.891	0.803	0.704	0.857	0.832	0.817±0.064
Precision	0.948	0.842	0.744	0.957	0.890	0.876±0.078
False positive rate	0.109	0.197	0.296	0.143	0.168	0.183±0.064
Accuracy	0.864	0.840	0.814	0.820	0.901	0.848±0.032
F1	0.898	0.856	0.828	0.877	0.919	0.876±0.032
AU-ROC	0.920	0.894	0.874	0.912	0.897	0.900±0.016

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Table S5. Average confusion matrices of 5-fold cross-validation over the unseen test data (n, epoch number) for the two methods

Actual Class	Predicted Class	
	EVD	IPM
Bagged Decision Trees		
EVD	309	18
IPM	37	168
Linear Discriminant Analysis		
EVD	286	40
IPM	45	160

EVD: extraventricular drain (ventriculostomy); IPM: intraparenchymal monitor.

Misclassification samples

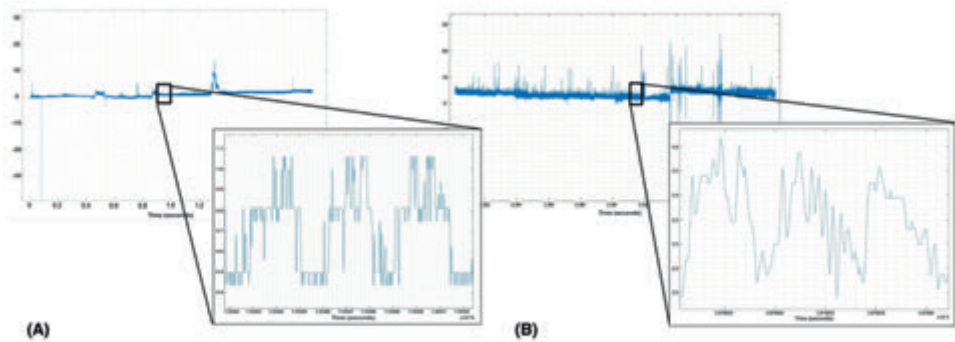


Figure S3: Examples of noisy ICP signals that were misclassified. (A) ICP signal from IPM detected as EVD. (B) ICP signal from EVD detected as IPM. As can be seen from the figure, the ICP signals were highly noisy and, thus, morphologically did not look like an ICP signal, causing them to be misclassified.

Subgroup analysis (cluster analysis)

Table S6. Cluster-based confounder-isolating patient-wise cross-validation results over the unseen test data using bagged decision trees. FPR and SD indicate the false positive rate and standard deviation, respectively. (Positive class: EVD, Negative class: IPM)

Performance Metric ^a	3-cluster	5-cluster	8-cluster
Sensitivity, mean (SD)	0.947 (0.017)	0.948 (0.038)	0.963 (0.033)
Specificity, mean (SD)	0.724 (0.084)	0.813 (0.160)	0.844 (0.146)
Precision, mean (SD)	0.804 (0.119)	0.874 (0.135)	0.891 (0.114)
False positive rate, mean (SD)	0.276 (0.084)	0.187 (0.160)	0.157 (0.146)
Accuracy, mean (SD)	0.852 (0.037)	0.890 (0.066)	0.918 (0.069)
F1 score, mean (SD)	0.864 (0.071)	0.902 (0.074)	0.922 (0.072)
AUROC, mean (SD)	0.852 (0.094)	0.901 (0.090)	0.929 (0.092)

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Table S7. Cluster-based confounder-isolating patient-wise cross-validation results over the unseen test data using linear discriminant analysis.

Performance Metric ^a	3-cluster	5-cluster	8-cluster
Sensitivity, mean (SD)	0.834 (0.079)	0.890 (0.056)	0.917 (0.039)
Specificity, mean (SD)	0.692 (0.047)	0.772 (0.100)	0.813 (0.128)
Precision, mean (SD)	0.768 (0.119)	0.834 (0.098)	0.846 (0.106)
False positive rate, mean (SD)	0.308 (0.047)	0.228 (0.100)	0.187 (0.128)
Accuracy, mean (SD)	0.786 (0.045)	0.844 (0.050)	0.871 (0.043)
F1 score, mean (SD)	0.796 (0.091)	0.858 (0.066)	0.874 (0.048)
AUROC, mean (SD)	0.811 (0.024)	0.888 (0.049)	0.919 (0.058)

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Subgroup Analysis by Propensity Score

Table S8. Propensity-based 4-fold patient-wise cross-validation results over the unseen test data using bagged decision trees.

Performance Metric ^a	Fold (Propensity Score)				Mean±SD
	1 (0-0.478)	2 (0.478-0.691)	3 (0.691-0.835)	4 (0.835-1)	
Sensitivity	1	0.877	0.953	0.986	0.954±0.048
Specificity	0.868	0.939	0.558	0.788	0.788±0.143
Precision	0.566	0.986	0.813	0.928	0.823±0.161
False positive rate	0.132	0.061	0.442	0.212	0.212±0.143
Accuracy	0.887	0.887	0.822	0.934	0.883±0.040
F1	0.723	0.928	0.877	0.956	0.871±0.090
AU-ROC	1	0.951	0.698	0.966	0.904±0.120

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Table S9. Propensity-based 4-fold patient-wise cross-validation results over the unseen test data using linear discriminant analysis.

Performance Metric ^a	Fold (Propensity Score)				Mean±SD
	1 (0-0.228)	2 (0.228-0.582)	3 (0.582-0.927)	4 (0.927-1)	
Sensitivity	0.96	0.872	0.886	0.942	0.915±0.037
Specificity	0.769	0.877	0.625	0.771	0.761±0.090
Precision	0.744	0.942	0.881	0.903	0.867±0.075
False positive rate	0.231	0.123	0.375	0.229	0.239±0.090
Accuracy	0.848	0.873	0.823	0.890	0.859±0.025
F1	0.838	0.906	0.884	0.922	0.887±0.032
AU-ROC	0.897	0.933	0.827	0.934	0.898±0.043

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Subgroup Analysis by Potential Confounders

Utilizing a bagged decision tree (**Table S10**) improves the sensitivity, specificity, F1-score, and AU-ROC by an average of 0.004, 0.001, 0.003, and 0.004, respectively, in a 5-fold cross-validation. On the other hand (**Table S11**), when using LDA, these values decrease by an average of 0.014, 0.005, and 0.009 for sensitivity, specificity, and F1-score, respectively, with no change in AU-ROC. This discrepancy may be due to the different classifiers' ability to handle features. We conclude that the list of 842 features already encompasses information captured by artifactual segments. Further research could explore non-ICP features to automatically generate metadata in a more efficient manner.

Table S10. Patient-wise 5-fold cross-validation results over the unseen test data using bagged decision trees (artifact-based features + 842 features).

Performance Metric ^a	Fold					Mean±SD
	1	2	3	4	5	
Sensitivity	0.896	0.938	0.991	0.933	0.980	0.948±0.034
Specificity	0.962	0.925	0.713	0.933	0.818	0.870±0.093
Precision	0.982	0.938	0.761	0.982	0.885	0.910±0.083
False positive rate	0.038	0.075	0.287	0.067	0.182	0.130±0.093
Accuracy	0.916	0.932	0.846	0.933	0.913	0.908±0.032
F1	0.937	0.938	0.861	0.957	0.930	0.925±0.033
AU-ROC	0.937	0.987	0.863	0.970	0.926	0.936±0.043

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Table S11. Patient-wise 5-fold cross-validation results over the unseen test data using LDA (artifact-based features + 842 features).

Performance Metric ^a	Fold					Mean±SD
	1	2	3	4	5	
Sensitivity	0.877	0.843	0.947	0.774	0.904	0.869±0.058
Specificity	0.846	0.789	0.736	0.874	0.818	0.813±0.048
Precision	0.930	0.829	0.768	0.960	0.877	0.873±0.069
False positive rate	0.154	0.211	0.264	0.126	0.182	0.188±0.048
Accuracy	0.868	0.818	0.837	0.794	0.868	0.837±0.029
F1	0.903	0.836	0.848	0.857	0.890	0.866±0.026
AU-ROC	0.924	0.889	0.898	0.911	0.892	0.903±0.013

^a Positive class, extraventricular drain (ventriculostomy); negative class, intraparenchymal monitor.

Feature selection

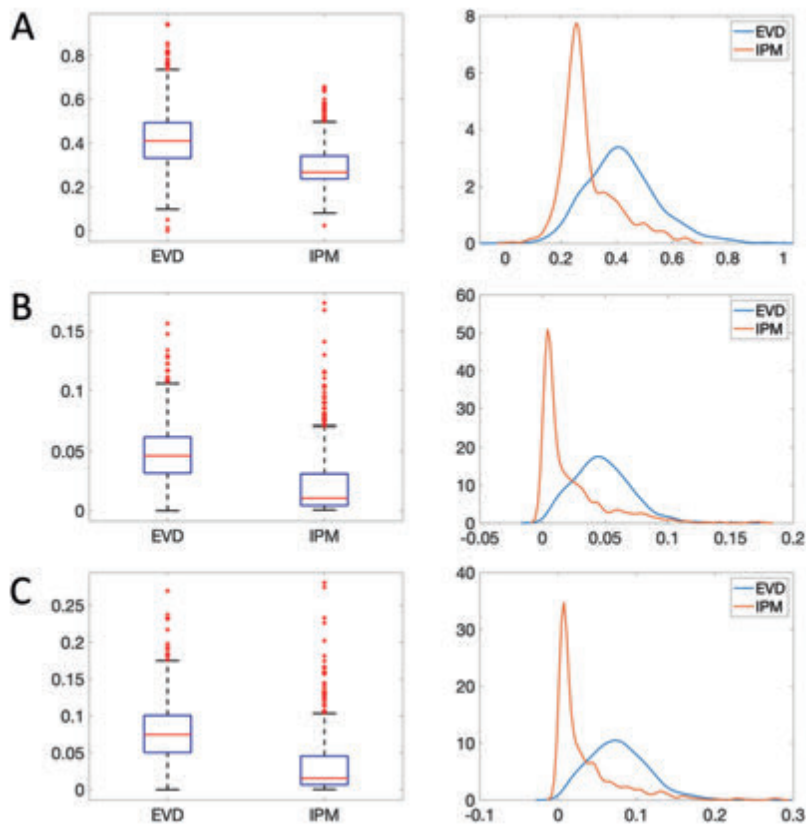


Figure S4: Example box plots and kernel density plots for a sample of the 62 selected features. A) Mean eigenvalues in phase space between rolling windows. B) Standard deviation of coefficients of an autoregressive model between rolling windows. C) Standard deviation of the distance between phase space nullclines (or clusters) between rolling windows.

Held-out evaluation

We kept the subgroup with both monitoring types of unseen and utilized them for evaluation. This means that we train the model using the rest of the patients and test it on the subgroup with both (9 patients). The results of bagged decision tree (BDT) and linear discriminant analysis (LDA) classifiers are as follows:

Table S12. Performance of Bagged Decision Tree (BDT) and Linear Discriminant Analysis (LDA) classifiers on held-out subgroup with both monitoring types.

	Sensitivity	Specificity	Precision	FPR	Accuracy	F1	AU-ROC
BDT	0.848	0.747	0.766	0.253	0.797	0.805	0.826
LDA	0.823	0.626	0.683	0.373	0.724	0.747	0.762

As evident from the results of the 9 unseen patients, the achieved sensitivity and specificity are promising. However, the performance is diminished compared to the cross-validation results. Additional validation data from multiple centres are necessary for a fair evaluation.