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

Supplementary Materials

Table of Contents

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

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 (1.4)	
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 category, 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)	
Subdural hemorrhage, n (%)	35 (79.5)	22 (88.0)	7 (77.8)	0.575
Epidural hemorrhage, n (%)	13 (29.5)	6 (24.0)	2 (22.2)	0.830
Subarachnoid hemorrhage, n (%)	40 (90.9)	24 (96.0)	8 (88.9)	0.763
Presence of 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
Acceleration-deceleration event, n (%)	20 (42.6)	9 (34.6)	5 (55.6)	0.679
Any loss of consciousness, 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)	

EVD: external ventricular drain; IPM: intraparenchymal monitor; ICP: intracranial pressure; BMI: body mass index; SD: standard deviation; GCS: Glasgow coma score; IQR: interquartile range; ED: emergency department; CT: computed tomography.

Feature selection and classification

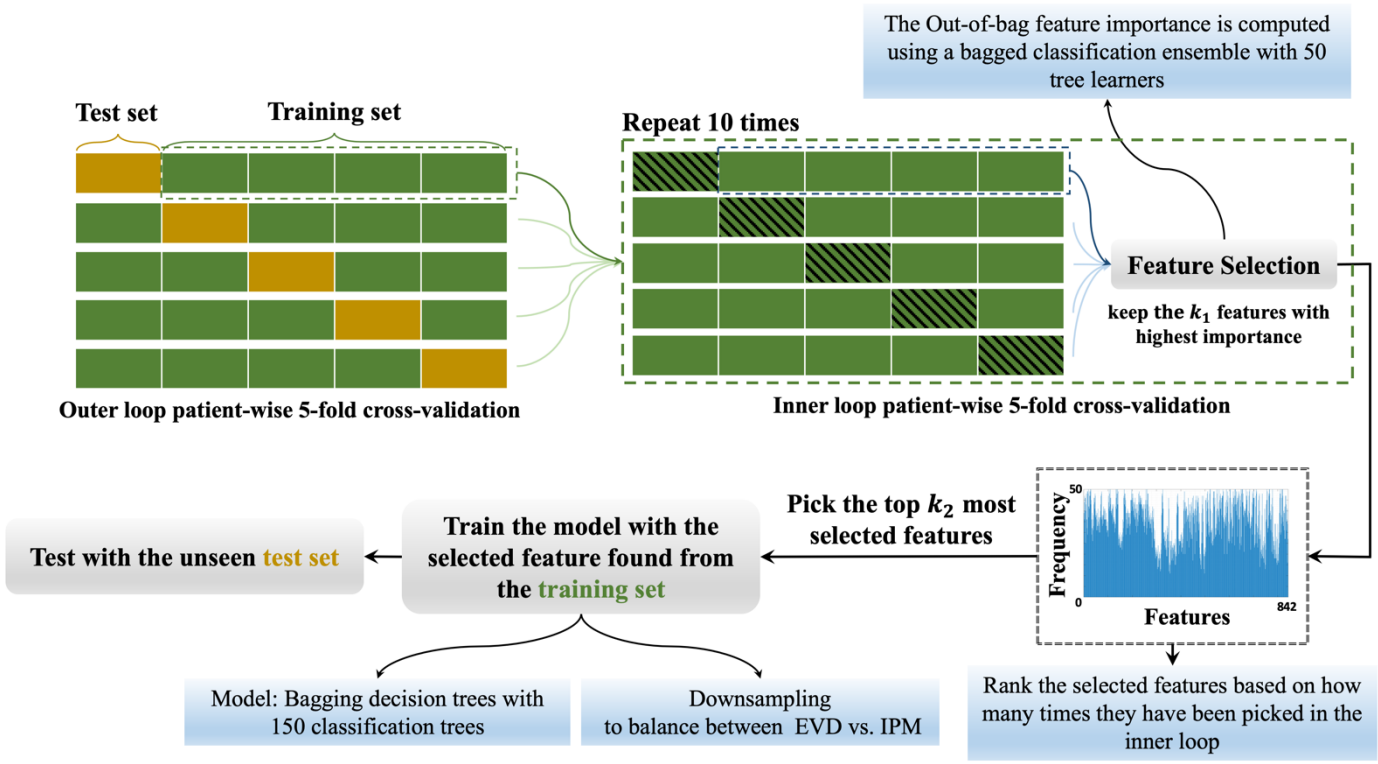


Figure S1: Schematic of model development. The outer loop partitioned the data into five folds, with four of these folds comprising 80% of the data being utilized for feature selection and model training, while the remaining 20% was utilized for evaluating the model performance. This process was repeated 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 $k1$) 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 $k1$ features. These $k1$ features were then ranked by selection frequency, and based on a 75^{th} percentile threshold, referred to as $k2$, 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 subdural hemorrhage on CT	2 categories; yes & no
Presence of subarachnoid hemorrhage 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)*

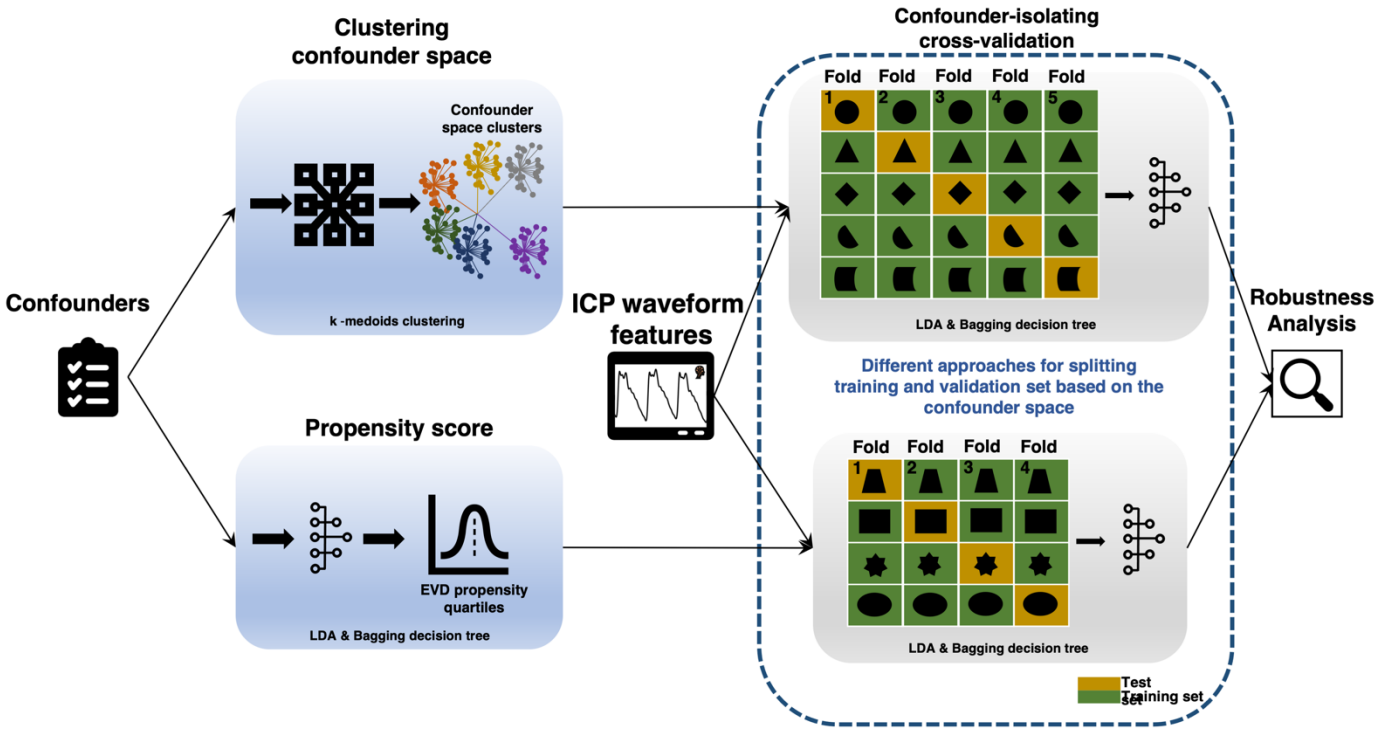


Figure S2: Procedures to ensure rigor and robustness included cluster- and propensity score-based confound-isolating cross-validation schemes. The cluster-based approach (top) employed the *k*-medoids clustering algorithm and iteratively partitioned the data into three, five, and eight clusters with 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 25^{th}, 50^{th}, 75^{th}, and 97^{th} 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

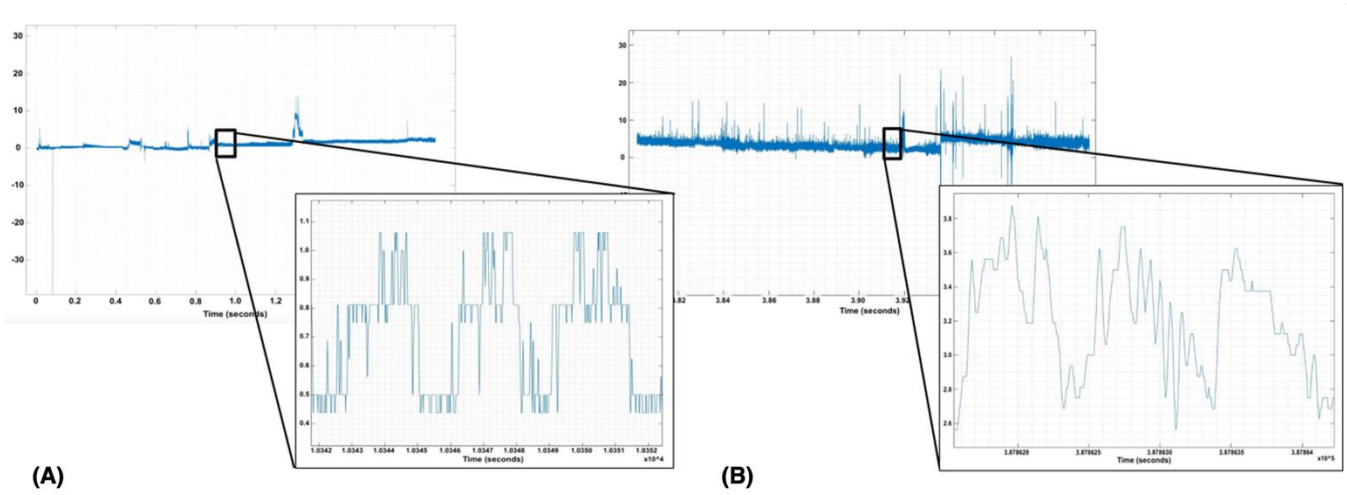
Table S5. Average confusion matrices of 5-fold cross-validation over the unseen test data (n, epoch number) for the two methods

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

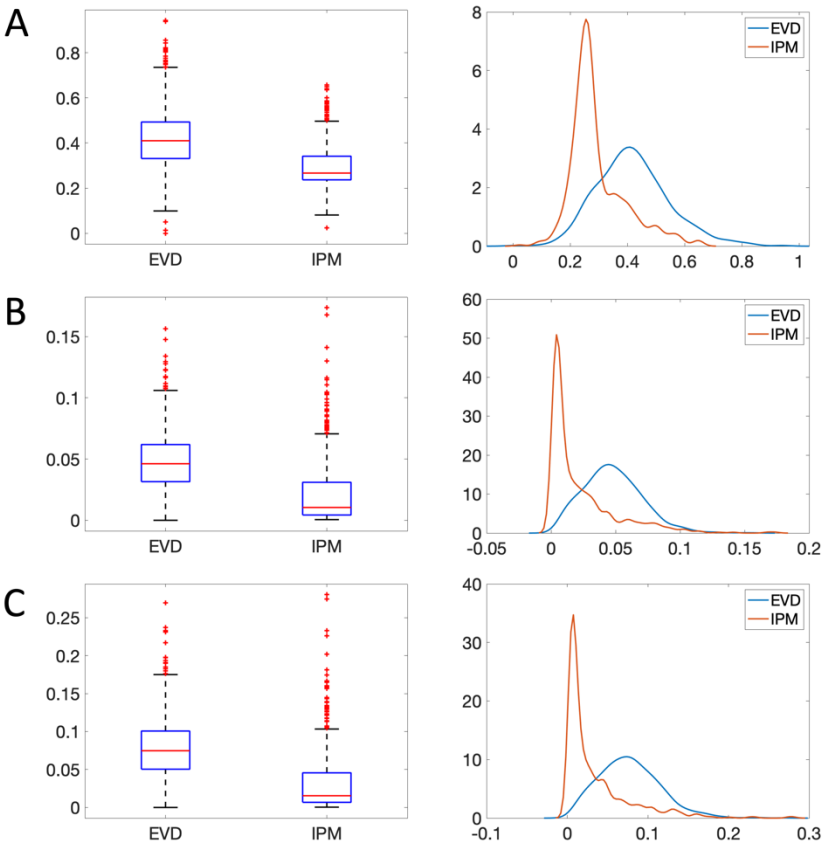


Figure S4: Example box plots and kernel density plots for a sample of the 62 selected features. A) Mean eigenvalues in phase space between rolling windows. B) Standard deviation of coefficients of an autoregressive model between rolling windows. C) Standard deviation of the distance between phase space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

IRB site info

Baylor, Protocol Number: H-31230, 6/15/2012. Institutional Review Board for Baylor College of Medicine and Affiliated Hospitals. TRANSFORMING RESEARCH AND CLINICAL KNOWLEDGE IN TRAUMATIC BRAIN INJURY II (TRACK-TBI II).

Utah, IRB # 00100289, 11/16/2017. Institutional Review Board the University of Utah. - Transforming Research and Clinical Knowledge in Traumatic Brain Injury (TRACK-TBI).

UCSF, IRB # 12-09465, 12/23/2013. UCSF Human research protection program committee on human research. Transforming Research and Clinical Knowledge in Traumatic Brain Injury U01.

University of Cincinnati, Study ID 2013-8278, 4/22/2014, Institutional Review Board- Federalwide Assurance #00003152 University of Cincinnati. Transforming Research and Clinical Knowledge in Traumatic Brain Injury: TRACK-TBI Clinical Protocol.

University of Pittsburgh medical center, PRO08080168, 2/14/2014, University of Pittsburgh Institutional Review Board. Transforming Traumatic Brain Injury Research and Clinical Care.