Machine Learning to Predict In-Hospital Morbidity and Mortality after Traumatic Brain Injury

Matsuo, Kazuya ; Aihara, Hideo ; Nakai, Tomoaki ; Morishita, Akitsugu ; Tohma, Yoshiki ; Kohmura, Eiji

(Citation)
Journal of Neurotrauma,37(1):202-210

(Issue Date)
2019-12-11

(Resource Type)
journal article

(Version)
Accepted Manuscript

Full title

Machine Learning to Predict In-hospital Morbidity and Mortality after Traumatic Brain Injury

Full names and institutions of authors

Kazuya Matsuo, M.D., Hideo Aihara, Ph.D., Tomoaki Nakai, Ph.D., Akitsugu Morishita, Ph.D., Yoshiki Tohma, Ph.D., Eiji Kohmura, Ph.D.

Institutional affiliation of Kazuya Matsuo, Tomoaki Nakai, and Eiji Kohmura: Department of Neurosurgery, Kobe University Graduate School of Medicine.

Institutional affiliation of Hideo Aihara and Akitsugu Morishita: Department of Neurosurgery, Hyogo Prefectural Kakogawa Medical Center.

Institutional affiliation of Yoshiki Tohma: Department of Emergency and Critical Care Medicine, Hyogo Prefectural Kakogawa Medical Center.

Running title

Machine Learning Prediction for Brain Trauma

Table of Contents title

Machine Learning to Predict Morbidity and Mortality after Brain Trauma

*Correspondence to: Kazuya Matsuo, M.D., Department of Neurosurgery, Kobe University Graduate School of Medicine, 7-5-1 Kusunoki-cho, Chuo-ku, Kobe, Hyogo 650-0017, Japan

Tel: +81-78-382-5966

Fax: +81-78-382-5979
E-mail: kkmatsuo@outlook.jp

Contact information for each author

Kazuya Matsuo
Address: Department of Neurosurgery, Kobe University Graduate School of Medicine. 7-5-1 Kusunoki-cho, Chuo-ku, Kobe, Hyogo 650-0017, Japan.
Tel: +81-78-382-5966 Fax: +81-78-382-5979 E-mail: kkmatsuo@outlook.jp
Hideo Aihara
Address: Department of Neurosurgery, Hyogo Prefectural Kakogawa Medical Center. 203 Kanno, Kanno-cho, Kakogawa, Hyogo 675-8555, Japan.
Tel: +81- 79-497-7000 Fax: +81-79-438-8800 E-mail: aihara@hp.pref.hyogo.jp
Tomoaki Nakai
Address: Department of Neurosurgery, Kobe University Graduate School of Medicine. 7-5-1 Kusunoki-cho, Chuo-ku, Kobe, Hyogo 650-0017, Japan.
Tel: +81-78-382-5966 Fax: +81-78-382-5979 E-mail: tomoakin@med.kobe-u.ac.jp
Akitsugu Morishita
Address: Department of Neurosurgery, Hyogo Prefectural Kakogawa Medical Center. 203 Kanno, Kanno-cho, Kakogawa, Hyogo 675-8555, Japan.
Tel: +81- 79-497-7000 Fax: +81-79-438-8800 E-mail: morishita@hp.pref.hyogo.jp
Yoshiki Tohma
Address: Department of Emergency and Critical Care Medicine, Hyogo Prefectural Kakogawa Medical Center. 203 Kanno, Kanno-cho, Kakogawa, Hyogo 675-8555, Japan.
Tel: +81- 79-497-7000 Fax: +81-79-438-8800 E-mail: yoshiki_tohma_01@mac.com
Eiji Kohmura

Address: Department of Neurosurgery, Kobe University Graduate School of Medicine. 7-5-1
Kusunoki-cho, Chuo-ku, Kobe, Hyogo 650-0017, Japan.
Tel: +81-78-382-5966 Fax: +81-78-382-5979 E-mail: ekohmura@med.kobe-u.ac.jp

Abstract

Recently, successful predictions using machine learning (ML) algorithms have been reported in various fields. However, in traumatic brain injury (TBI) cohorts, few studies have examined modern ML algorithms. To develop a simple ML model for TBI outcome prediction, we conducted a performance comparison of nine algorithms: ridge regression, LASSO regression, random forest, gradient boosting, extra trees, decision tree, Gaussian naïve Bayes, multinomial naïve Bayes, and support vector machine. Fourteen feasible parameters were introduced in the ML models, including age, Glasgow coma scale, systolic blood pressure, abnormal pupillary response, major extracranial injury, computed tomography findings, and routinely collected laboratory values (glucose, C-reactive protein, and fibrin/fibrinogen degradation products). Data from 232 TBI patients were randomly divided into a training sample (80%) for hyperparameter tuning and validation sample (20%). The bootstrap method was used for validation. Random forest demonstrated the best performance for in-hospital poor outcome prediction and ridge regression for in-hospital mortality prediction: the mean statistical measures were 100% sensitivity, 72.3% specificity, 91.7% accuracy, and 0.895 area under the receiver operating characteristic curve (AUC); and 88.4% sensitivity, 88.2% specificity, 88.6% accuracy, and 0.875 AUC, respectively. Based on the feature selection method using the tree-based ensemble algorithm, age, Glasgow coma scale, fibrin/fibrinogen degradation products, and glucose were identified as the most important prognostic factors for poor outcome and mortality. Our results indicated the relatively good predictive performance of modern ML for TBI outcome. Further external validation is required for more heterogeneous samples to confirm our results.

Keywords

Artificial intelligence, Machine learning, Traumatic brain injury, Outcome predictor.

Introduction

Can artificial intelligence (AI) accurately predict the outcome of traumatic brain injury (TBI)? After TBI, a reliable prediction of outcome is crucial for determining the optimal treatment strategy and supporting anxious caregivers so that they can manage the situation and make decisions. However, only 37% of clinicians agree that they currently assess prognosis accurately. [1] Although several prognostic parameters and models have been proposed to achieve the best prediction of both morbidity and mortality, [2-5] there is still no effective and comprehensive model to predict TBI outcome based on routinely available variables. [2] To be useful, prognostic models need to be applicable across all severities of injury and capable of being expressed in a manner that indicates the likelihood of an individual patient achieving different outcomes at some future time. This depends on combining information on the different individual prognostic parameters. [3] On this point, modern AI should have the capability to combine information and may obtain a good prediction.

Recently, AI has been used widely in the medical and healthcare fields because of tremendous advances and feasibility regarding using various machine learning (ML) algorithms for successful prediction and diagnosis. [6] In fact, the U.S. Food and Drug Administration has cleared some healthcare companies to market their deep learning technology, one of the latest ML techniques, to medical professionals. [7] However, for TBI cohorts, few studies have examined relatively new ML algorithms, and many studies have applied former-generation algorithms, such as a simple decision tree model. [8] Advanced research using an artificial neural network (ANN), a modern ML algorithm, has demonstrated high accuracy in the prediction of in-hospital survival after TBI; [9] however, an ANN, such as a multilayer perceptron method, is still not easy to apply for general use because it requires tuning many hyperparameters and a large amount of data for training.

The aim of the present study is to develop a simple and accurate ML model for the prediction of morbidity and mortality after TBI using only a small set of parameters that are rapidly and easily applicable in routine emergent practice. To achieve this aim, we used publicly available ML algorithms in Python 3.6 (Python Software Foundation) with only general clinical factors, including systolic blood pressure (SBP), abnormal pupillary response, Glasgow coma scale (GCS), computed tomography (CT) findings, and routine laboratory values. Additionally, important parameters for outcome prediction were determined using the ML technique.

Materials and Methods

Study Population

We screened consecutive TBI patients admitted between October 2013 and September 2016 at Hyogo Prefectural Kakogawa Medical Center, which is a tertiary emergency center in Japan. All 268 patients with non-penetrating TBI that required emergency hospitalization because of abnormal findings for head CT were then reviewed. Patients were excluded if they experienced cardiopulmonary arrest (CPA) on arrival, were a child under 10 years old, were pregnant, or had insufficient admission laboratory or clinical data.

Treatments

All patients admitted to the emergency room received the same initial standardized evaluation and treatment protocol based on the advanced trauma life support concept. [10,11] Patients were evaluated by CT scan as soon after stabilization as possible. When the CT scan showed severe TBI with midline shift greater than 5 mm or compressed basal cisterns, burr-hole craniostomy with partial drainage of the subdural collection was soon performed in the emergency room after the CT scan. Even when the CT scan showed midline shift less than 5 mm, burr-hole craniostomy was performed in the emergency room after the CT scan on patients who had anisocoria or abnormal pupillary responses, or had GCS of eight or less. Simultaneously, an intraparenchymal intracranial pressure (ICP) sensor was placed for ICP monitoring and a subdural catheter for ICP management. Subsequently, a craniotomy was performed in the operation room to evacuate the massive subdural hematoma or traumatic intracerebral hemorrhage. A large unilateral decompressive craniectomy was also performed when significant brain swelling was evident intraoperatively. Despite the patient meeting the above treatment criteria, emergency brain surgery was not performed for the following: patients who progressed to brain death, patients who had unstable vital signs because of major extracranial injuries, or cases in which the patient's family requested conservative treatment. Details of management in the intensive care unit are provided in the Appendix.

Predictive parameters

A total of 14 feasible parameters were introduced in the ML algorithm based on their known or expected influence on the outcome. These parameters included age, [2-4] GCS score, [2-4,8] abnormal pupillary response, [2-4,8] SBP, major extracranial injury, CT findings, [2-4,12] and routinely collected laboratory values (glucose, [2,4,5,8] C-reactive protein (CRP), and fibrin/fibrinogen degradation products (FDP) [5,13]). All laboratory and clinical data were recorded at admission. CT findings were derived from both the admission CT scan and the second scan, which was obtained within 3 hours after the initial CT scan or within the next day. The CT findings consisted of cerebral contusion, acute subdural hematoma (ASDH), traumatic subarachnoid hemorrhage (TSAH), epidural hematoma, and skull fracture. In addition, Marshall CT classification [12] was obtained and used as the parameter (Appendix Table 1). Cerebral contusion included traumatic cerebral hemorrhage, such as gliding contusion. Acute subdural hematoma was defined as a collection of blood that is a crescent-shaped hyperdense lesion on CT and located between the dura mater and subarachnoid membrane. Epidural hematoma was defined as a collection of blood that is typically a biconvex-shaped hyperdense signal on CT and located between the inner surface of the skull and dura. Every hematoma was diagnosed in the presence of any amount of blood. Major extracranial injury was defined as an injury with an abbreviated injury scale score ≥3 in the thorax, abdomen/pelvis, and extremities. Among clinical parameters, investigations on the relation between CRP and TBI are scarce. However, CRP was used as the prognostic parameter because it is known to reflect the impact of trauma on the body and is associated with tissue damage. [14]

Machine learning model development

We conducted a performance comparison of the following nine ML algorithms considered to be useful for various datasets: [15] ridge regression, least absolute shrinkage and selection operator (LASSO) regression, random forest, gradient boosting, extra trees, decision tree, Gaussian naïve Bayes, multinomial naïve Bayes, and support vector machine (SVM) (its kernel consisted of linear, radial basis function (RBF), polynomial (poly), and sigmoid). All these supervised algorithms were implemented using Scikit-learn, [16] which is a free ML library for Python. The patient data were randomly divided into a training sample (80%), which was used for hyperparameter tuning to generate a plausible model, and a validation sample (20%), which was used to test the performance of each model that was generated in the training sample. All the cases with missing elements were not included in either the training or test data. To adjust and identify the best set of hyperparameters for each ML algorithm, we performed a stratified five-fold cross-validation procedure on the training sample. Briefly, the *k*-fold cross-validation algorithm partitions the dataset into *k* sets and uses *k*-1 sets for training and the remaining set for testing. This is performed *k* times and the results of the different test sets are averaged, which guarantees the independence of the results from the actual dataset subdivision. [17]

Ridge regression and LASSO regression models are the most fundamental regularization techniques among modern regression algorithms, and work well in cases of high dimensionality and multicollinearity among the variables in the data. [18] Random forest, gradient boosting, and extra trees are common ensemble methods that combine multiple simple tree models and have proved to produce reliable predictions. Tree-based ensemble models have been defined as the most accurate model on various datasets. [15] Naïve Bayes is the simplest form of Bayesian network, in which all attributes are independent given the value of the class variable. It is an efficient and effective inductive learning algorithm for classification and has been found to perform well, despite its simplicity. [19] An SVM has high sensitivity and generalization ability as a result of its kernel function, for example, polynomial or RBF, which constructs a decision surface in a very high-dimensional feature space to perform binary classification. [20,21]

Model comparisons

After the optimal hyperparameters were determined for each ML algorithm on the training sample, each model was ranked according to the sensitivity, specificity, prediction accuracy, and area under the receiver operating characteristic curve (AUC). The models with the highest score for each statistical measure were then selected for further analysis in the test sample as external validation. In this procedure, we used the standard bootstrap method as an additional internal validation technique. The utility of this method has been demonstrated in previous studies. [22,23] The bootstrap method is an iterative resampling technique used to estimate summary statistics, such as the mean or standard deviation, by sampling a dataset with replacement. In ML, it is common to use a sample size that is the same as that of the original dataset. As a result, some samples are represented multiple times in the bootstrap sample, whereas others are not represented. In this study, the number of resampling repetitions, which should be as large as possible to ensure the stability of the estimates, was set to 1,000 repetitions. [22,23] The bootstrap averages of sensitivity, specificity, prediction accuracy, and AUC were calculated, and 95% confidence intervals were estimated.

Additionally, the importance values of each predictive parameter were measured based on the feature selection method using the tree-based ensemble algorithm. All analyses were conducted using Python version 3.6.2 and Scikit-learn version 0.19.1.

Outcomes

Outcome at discharge based on the Glasgow outcome score (GOS) was determined for all patients. [24] GOS 1–3 (death, persistent vegetative state, and severe disability) was regarded as a poor outcome. We used death and poor outcome as outcome measures.

Results

Study participants

Among 268 TBI patients, 36 (13.4%) were excluded for the following reasons: CPA at arrival (n = 15), younger than 10 years old (n = 2), pregnant (n = 2), lack of pupillary findings (n = 2), and lack of measurement of FDP (n = 15). Thus, a total of 232 patients with TBI were included in the data and separated into training data and test data. The study diagram is shown in Figure 1, and the baseline characteristics of the patients, including CT findings and laboratory data, are shown in Table 1. The mean age at injury was 59.4 years (SD 21.9; median 66.5; range 11–92), 169 patients were male (72.8%), and the mean GCS was 9.1 (SD 4.4; median 10; range 3–15). Approximately half of the patients had severe TBI (GCS score of 3–8). The most frequent CT findings were skull fracture (53.9%) and cerebral contusion (49.1%). The mean length of stay was 28.7 days (SD 25.7; median 24.5; range 1–134) and the in-hospital mortality rate was 26.3%. At discharge, 7.8% had good recovery (GOS score 5), 14.7% had moderate disability (GOS score 4), and 77.6% had a poor outcome (GOS score 1–3).

Prediction of poor outcome

To evaluate the predictive performance of each ML model for poor outcome, first, five-fold cross-validation was performed on the training sample (Table 2). It showed that the highest sensitivity was 97.2%, which was achieved by random forest. The highest specificity was 82.8%, which was achieved by Gaussian naïve Bayes. The highest prediction accuracy was 87.5%, which was achieved by gradient boosting. The highest AUC was 0.894, which was achieved by the SVM (kernel: RBF). The ROC curves of each algorithm are plotted in Figure 2. Next, to confirm the predictive performance of each ML model, which expresses the highest score for each statistical measurement in the training sample, these models were examined on the bootstrapped test data (Table 3). This showed that random forest outperformed the other models in terms of accuracy and AUC. The tuned hyperparameters of these models are listed in Appendix Table 2.

Prediction of death

To evaluate the predictive performance of each ML model for mortality, first, five-fold cross-validation was performed on the training sample (Table 4). It showed that the highest sensitivity was 85.1%, which was achieved by ridge regression. The highest specificity was 99.3%, which was achieved by random forest. The highest prediction accuracy was 89.8%, which was achieved by the SVM (kernel: linear). The highest AUC was 0.960, which was achieved by random forest. The ROC curves of each algorithm are plotted in Figure 3. Next, to confirm the predictive performance of each ML model, which expresses the highest score for each statistical measurement in the training sample, these models were examined on the bootstrapped test data (Table 5). This showed that ridge regression outperformed the other models in terms of sensitivity and AUC. The tuned hyperparameters of these models are listed in Appendix Table 3.

Importance values of the predictors

To detect the importance values of each predictive parameter for both poor outcome and mortality, the feature selection method using the random forest algorithm was applied. As a result, age, GCS, and FDP were ranked as the three most important parameters associated with poor outcome (Figure 4A). In the same manner, FDP, GCS, and abnormal pupillary response were ranked as the three most important parameters associated with mortality (Figure 4B). Interestingly, CRP was identified as the sixth-ranked associated factor for poor outcome, which was more important than any CT findings. Among the CT findings, Marshal CT classification had the strongest association with both poor outcome and mortality.

Discussion

Our results indicated the relatively good predictive performance of modern ML for TBI outcome. Among the readily available ML algorithms, random forest demonstrated the best performance for poor outcome prediction and ridge regression for mortality prediction. For the prediction of poor outcome based on random forest with 14 clinical parameters, the mean statistical measures were 100% sensitivity, 72.3% specificity, 91.7% accuracy, and 0.895 AUC for the bootstrap test. Similarly, ridge regression with the same 14 parameters demonstrated a good prediction of mortality, with 88.4% sensitivity, 88.2% specificity, 88.6% accuracy, and 0.875 AUC. Additionally, based on the feature selection method using the tree-based ensemble algorithm, age, GCS, FDP, and glucose appear to be the most important prognostic factors for both poor outcome and mortality. In particular, SBP tended to be associated with morbidity, and abnormal pupillary response tended to be associated with mortality.

ML and TBI

A few TBI studies have been conducted using modern ML. Although a few studies have focused on ANNs, which is a modern ML algorithm, for TBI outcome prediction, other modern ML algorithms, such as extra trees and gradient boosting, which have demonstrated good prediction performance in other fields, [6,15] have not been studied for TBI cohorts. Rughani et al. applied an ANN for TBI patients to predict in-hospital survival using 11 parameters, including age, sex, first SBP, total GCS score, and individual components of the GCS score both at the scene of injury and in the emergency department. [9] The algorithm was trained on a sample of 7,769 patients and evaluated on an independent sample of 100 patients. The prediction of in-hospital mortality was comparable with our results, which were 87.8% accuracy and 0.86 AUC. However, it should be emphasized that several parameters crucial to outcome prediction were absent, such as CT findings and laboratory values. Eftekhar et al. applied an ANN for 1,271 TBI patients to predict mortality using the following seven parameters: GCS, tracheal intubation status, age, SBP, respiratory rate, pulse rate, and injury severity score (ISS). [25] The prediction of mortality was better than our results, which were 95.1% accuracy and 0.965 AUC. However, the interpretation of the results from that study is limited. First, only 7.5% of the patients had severe TBI (GCS < 8). Generally, moderate or mild TBI patients are unlikely to die, which might make prediction easier and increase the prediction accuracy. Second, it is not clear at which time point mortality was measured. Third, parameters such as GCS and ISS were all converted to binary values using a method that was not defined. Shi et al. applied an ANN for 16,956 TBI patients who had undergone surgery to predict in-hospital mortality using the following six parameters; sex, age, Charlson comorbidity index, hospital volume, surgeon volume, and length of stay. [26] The prediction of in-hospital mortality was comparable with our results, which were 95.2% accuracy and 0.896 AUC. However, it is difficult to make a comparison with our results because Shi et al.'s population consisted of only surgically treated patients. Furthermore, the indication for surgical management of TBI was based on the professional judgment of the individual surgeon. This may be the cause of the good predictive performance of the model that applied surgeon and hospital volume as the parameters. Another potential limitation is that chronic subdural hematoma was included in the sample, and the length of stay was included as a parameter, which cannot be used in emergency settings.

Because our ML model is based on CT findings, including the scan on the day following admission, it is difficult to obtain prediction results on the first day. Therefore, although the presented model can be useful to help the family and clinician make decisions on the choice of treatment by providing an estimation of morbidity and mortality with a specific value, the future model should be improved by using initial CT findings alone, which will enable us to use the prediction results more quickly to optimize the treatment strategy.

Importance values of the predictors

Based on the feature selection method using the tree-based ensemble algorithm, age, GCS, FDP, and glucose were identified as the most important prognostic parameters for both poor outcome and mortality. Age and GCS are already known as reliable predictors. [2-4] Our results also confirm this finding.

In the present study, age was more likely to be associated with poor outcome than mortality. A poor outcome for the elderly could be explained by several factors, such as decreased regeneration or plasticity of the brain, and preexisting comorbidities, which delay the rehabilitation process. Premorbid activities of daily living (ADL) impairment should also influence poor outcome after TBI; however, premorbid ADL was not included as a prognostic parameter in the present study because it depends on conjecture and is therefore difficult to prove. FDP was determined as the most important predictor of mortality and the third-most important predictor of poor outcome following age and GCS based on the feature selection method of ML. TBI-associated coagulopathy is a relatively common pathology, of which the incidence was found to be 35.2% in recent meta-analysis. [13] It is considered that the development of coagulopathy after TBI is significantly associated with increased mortality and higher incidence of delayed injury and disability at discharge. [5,13] Although it is not yet clear which parameters should be used in the assessment of coagulopathy in TBI, [13] it has been suggested that D-dimer and FDP are more useful than platelet count, prothrombin time (PT), and activated partial thromboplastin time in the prediction of mortality. [27] TBI creates a hypercoagulable state, in part because cerebral tissue is a rich source of potent platelet activating and procoagulant molecules, but their contribution to the development of TBI-associated coagulopathy remains largely unknown. [28]

Blood glucose was the most important prognostic laboratory variable in the IMPACT study, which used prospectively collected data from several thousand patients. [5] By contrast, in our study, it was less important than the FDP level. This may be mainly because the investigated laboratory variable in the IMPACT study did not include FDP, but included glucose, sodium, pH, hemoglobin, platelet count, and PT. Stress response is supposed to induce hyperglycemia with an increase in the levels of catecholamine that cause a decrease in insulin secretion. [29]

To the best of our knowledge, there is only limited research regarding the negative relation between CRP and TBI outcome. [30] CRP is known to be released in relation to the extent of tissue damage. Similar to patients undergoing elective surgery, multiple-trauma patients may show an increase of CRP that indicates inflammation during the early post-traumatic period independent of infection. [31] For the same reason, the relatively high importance of CRP in our study might be attributed to the severity of extracranial injury.

For these laboratory variables, the question of causality is relevant when attempts are made to correct abnormal values in the expectation that this will improve outcome. Further studies are required to establish whether the correction of abnormal values is beneficial. Finally, we note a remarkable feature of AI: a property to seek and develop new factors that are out of our scope. Therefore, it might be more meaningful to include some potential predictors that are not well elucidated yet into prognostic parameters for further ML analyses.

Limitations

One of the main limitations of our study is the small sample size. Because big data makes ML models more accurate in general, a larger sample size will be required for more accurate predictions. However, a previous report suggested that 80–560 samples were required for supervised machine learning, except for the deep learning method, and the required sample size depended on the dataset and sampling method. [32] Thus, the sample size in our study might be sufficient for building the prediction model because it provided a certain level of predictive performance even though it was relatively limited. Second, several potentially important parameters, such as hypoxia or anemia, [2] were not considered in our study. Third, the outcome was assessed at discharge. Thus, it should be noted that the proposed models are not applicable to predict long-term outcome. However, many physicians consider that the most important outcome to predict is in-hospital mortality. [1] Finally, almost half of our study population consisted of severe TBI patients. Because of this observed tendency, the prediction accuracy of our models may be possibly decreased when they are applied to mild TBI patients. Although the outcome prediction may be more accurate when the number of mild TBI patients increases in the training sample, it may be difficult to implement this in practice because most of such patients do not require laboratory tests, which was essential to develop our ML models. Additionally, because the samples were obtained retrospectively from a single institution, they may have been biased and the proposed ML models may not be applicable to other institutions where different treatment strategies or patient demographics might exist. Although internal validation was applied with cross-validation and the bootstrap method, further external validation is necessary in another setting that differs in time or place to validate the performance of our prognostic models.

Conclusion

Our results indicated relatively good predictive performance of modern ML for TBI outcome. Random forest demonstrated the best performance for poor outcome prediction and ridge regression for mortality prediction; both of which achieved nearly 90% accuracy. Additionally, based on the feature selection method using the tree-based ensemble algorithm, age, GCS, FDP, and glucose appear to be the most important prognostic factors for both poor outcome and mortality. These results represent a milestone in the comprehensive development of ML in the prediction of outcome after TBI. Looking forward, rapid advances in AI technology will continue to improve the accuracy and reliability of prediction and diagnosis. We believe that the use of such a powerful tool will help to provide more effective and convenient treatment for patients; however, an accurate prediction of patient outcome does not tell us what to do if we want to change that outcome. [6] Even if AI with more accurate prediction performance predicts poor prognosis, we have to intensify our efforts to overcome the predicted poor prognosis with more advanced and specialized treatment.

Author Disclosure Statement

No competing financial interests exist for all authors.

References

1. Perel, P., Wasserberg, J., Ravi, R.R., Shakur, H., Edwards, P., and Roberts, I. (2007). Prognosis following head injury: a survey of doctors from developing and developed countries. J Eval Clin Pract. 13, 464-465.

2. Lingsma, H.F., Roozenbeek, B., Steyerberg, E.W., Murray, G.D., and Maas, A.I. (2010). Early prognosis in traumatic brain injury: from prophecies to predictions. Lancet Neurol. 9, 543-554.

3. Chesnut, R.M., Ghajar, J., Maas, A.I., Marion, D.W., Servadei, F., Teasdale, G.M., Unterberg, A., Von Holst, H., and Walters, B.C. (2000). Part 2: Early indicators of prognosis in severe traumatic brain injury. J Neurotrauma 17, 555-627.

4. Murray, G.D., Butcher, I., McHugh, G.S., Lu, J., Mushkudiani, N.A., Maas, A.I., Marmarou, A., and Steyerberg, E.W. (2007). Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. J Neurotrauma 24, 329-337.

5. Van Beek, J.G., Mushkudiani, N.A., Steyerberg, E.W., Butcher, I., McHugh, G.S., Lu, J., Marmarou, A., Murray, G.D., and Maas, A.I. (2007). Prognostic value of admission laboratory parameters in traumatic brain injury: results from the IMPACT study. J Neurotrauma. 24, 315-328.

6. Chen, J.H. and Asch, S.M. (2017). Machine learning and prediction in medicine - beyond the peak of inflated expectations. N Engl J Med. 376, 2507-2509.

7. Koch, M. (2018). Artificial intelligence is becoming natural. Cell 173, 531-533.

8. Rovlias, A. and Kotsou, S. (2004). Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables. J Neurotrauma 21, 886-893.

9. Rughani, A.I., Dumont, T.M., Lu, Z., Bongard, J., Horgan, M.A., Penar, P.L., and Tranmer, B.I. (2010). Use of an artificial neural network to predict head injury outcome. J Neurosurg. 113, 585-590.

10. ATLS Subcommittee; American College of Surgeons' Committee on Trauma; International ATLS working group. (2013). Advanced trauma life support (ATLS®): the ninth edition. J Trauma Acute Care Surg. 74, 1363-1366.

11. Shigemori, M., Abe, T., Aruga, T., Ogawa, T., Okudera, H., Ono, J., Onuma, T., Katayama, Y., Kawai,

N., Kawamata, T., Kohmura, E., Sakaki, T., Sakamoto, T., Sasaki, T., Sato, A., Shiogai, T., Shima, K., Sugiura, K., Takasato, Y., Tokutomi, T., Tomita, H., Toyoda, I., Nagao, S., Nakamura, H., Park, Y.S., Matsumae, M., Miki, T., Miyake, Y., Murai, H., Murakami, S., Yamaura, A., Yamaki, T., Yamada, K., and Yoshimine, T.; Guidelines Committee on the Management of Severe Head Injury, Japan Society of Neurotraumatology. (2012). Guidelines for the management of severe head injury, 2nd edition guidelines from the guidelines committee on the management of severe head injury, the Japan Society of Neurotraumatology. Neurol Med Chir (Tokyo) 52, 1-30.

12. Marshall, L.F., Marshall, S.B., Klauber, M.R., Van Berkum Clark, M., Eisenberg, H., Jane, J.A., Luerssen, T.G., Marmarou, A., Foulkes, M.A. (1992). The diagnosis of head injury requires a classification based on computed axial tomography. J Neurotrauma Suppl 1, S287-292.

13. Epstein, D.S., Mitra, B., O'Reilly, G., Rosenfeld, J.V., and Cameron, P.A. (2014). Acute traumatic coagulopathy in the setting of isolated traumatic brain injury: a systematic review and meta-analysis. Injury 45, 819-824.

14. Gebhard, F., Pfetsch, H., Steinbach, G., Strecker, W., Kinzl, L., and Brückner, U.B. (2000). Is interleukin 6 an early marker of injury severity following major trauma in humans? Arch Surg. 135, 291-295.

15. Olson, R.S., Cava, W., Mustahsan, Z., Varik, A., and Moore, J.H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. Pac Symp Biocomput. 23, 192-203.

16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,M., and Duchesnay, E. (2011). Scikit-learn: machine learning in Python. J Mach Learn Res. 12, 2825–2830.

17. Bernau, C., Riester, M., Boulesteix, A.L., Parmigiani, G., Huttenhower, C., Waldron, L., and Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. Bioinformatics 30, i105-112.

18. Musoro, J.Z., Zwinderman, A.H., Puhan, M.A., ter Riet, G., and Geskus, R.B. (2014). Validation of prediction models based on lasso regression with multiply imputed data. BMC Med Res Methodol. 14, 116.

19. Cui, S., Zhao, L., Wang, Y., Dong, Q., Ma, J., Wang, Y., Zhao, W., and Ma, X. (2018). Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation. Injury 49, 1865-1870.

20. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning 20, 273-297.

21. Kuo, P.J., Wu, S.C., Chien, P.C., Rau, C.S., Chen, Y.C., Hsieh, H.Y., and Hsieh, C.H. (2018). Derivation and validation of different machine-learning models in mortality prediction of trauma in motorcycle riders: a cross-sectional retrospective study in southern Taiwan. BMJ Open 8, e018252.

22. Bland, J.M., and Altman, D.G. (2015). Statistics notes: Bootstrap resampling methods. BMJ 350, h2622.

23. Brunelli, A., and Rocco, G. (2006). Internal validation of risk models in lung resection surgery: bootstrap versus training-and-test sampling. J Thorac Cardiovasc Surg. 131, 1243-1247.

24. Jennett, B., and Bond, M. (1975). Assessment of outcome after severe brain damage. Lancet 1, 480-484.

25. Eftekhar, B., Mohammad, K., Ardebili, H.E., Ghodsi, M., and Ketabchi, E. (2005). Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data. BMC Med Inform Decis Mak. 5, 3.

26. Shi, H.Y., Hwang, S.L., Lee, K.T., and Lin, C.L. (2013). In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. J Neurosurg. 118, 746-752.

27. Saggar, V., Mittal, R.S., and Vyas, M.C. (2009). Hemostatic abnormalities in patients with closed head injuries and their role in predicting early mortality. J Neurotrauma 26, 1665-1668.

28. Zhang, J., Jiang, R., Liu, L., Watkins, T., Zhang, F., and Dong, J.F. (2012). Traumatic brain injury-associated coagulopathy. J Neurotrauma 29, 2597-2605.

29. Weissman, C. (1990). The metabolic response to stress: an overview and update. Anesthesiology 73, 308-327.

30. Hergenroeder, G., Redell, J.B., Moore, A.N., Dubinsky, W.P., Funk, R.T., Crommett, J., Clifton, G.L., Levine, R., Valadk,a A., and Dash, P.K. (2008). Identification of serum biomarkers in brain-injured adults: potential for predicting elevated intracranial pressure. J Neurotrauma 25, 79-93.

31. Meisner, M., Adina, H., Schmidt, J. (2006). Correlation of procalcitonin and C-reactive protein to inflammation, complications, and outcome during the intensive care unit course of multiple-trauma patients. Crit Care 10, R1.

32. Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., Ngo, L.H. (2012). Predicting sample size required for classification performance. BMC Med Inform Decis Mak. 12, 8. doi: 10.1186/1472-6947-12-8.

Table 1. Demographics and injury characteristics of the study sample

Variable	**Value***
Number of patients	232
Age (years)	59.4 (SD 21.9)
Male sex	169 (72.8%)
Systolic BP (mmHg)	143 (SD 38.2)
Abnormal pupillary response	91 (39.2%)
Median GCS (range)	10 (3–15)
GCS 3–8	106 (45.7%)
Isolated traumatic brain injury	88 (37.9%)
Major extracranial injury	104 (44.8%)
CT findings	
Skull fracture	125 (53.9%)
Cerebral contusion	114 (49.1%)
TSAH	90 (38.8%)
ASDH	85 (36.6%)
AEDH	29 (12.5%)
Marshal CT classification	

1	0 (0%)
2	104 (44.8%)
3	46 (19.8%)
4	6 (2.6%)
5	32 (13.8%)
6	44 (19.0%)
Laboratory values	
Glucose (mg/dl)	181 (SD 70.6)
FDP (μg/mL)	280 (SD 576)
CRP (mg/dl)	0.37 (SD 1.57)
Outcome	
In-hospital poor outcome	180 (77.6%)
In-hospital mortality	61 (26.3%)
Length of stay (days)	28.7 (SD 25.7)

* Values are presented as the number of patients with percent or mean values with SD, unless otherwise noted.

SD: standard deviation; BP: blood pressure; GCS: Glasgow coma scale; CT: computed tomography; TSAH: traumatic subarachnoid hemorrhage; ASDH: acute subdural hematoma; AEDH: acute epidural hematoma; FDP: fibrin/fibrinogen degradation products; CRP: C-reactive protein.

Table 2. In-hospital morbidity prediction performance of ML models for the training sample assessed using five-fold cross-validation (sorted by the AUC value)

ML algorithm	Sensitivity	Specificity	Prediction accuracy	AUC
SVM "rbf"	0.945	0.586	0.865	0.894
SVM "sigmoid"	0.938	0.536	0.849	0.882
Extra trees	0.958	0.536	0.859	0.881
Ridge regression	0.882	0.706	0.843	0.879
SVM "poly"	0.931	0.631	0.865	0.8776
SVM "linear"	0.917	0.536	0.832	0.873
Gradient boosting	0.937	0.628	0.875	0.869
LASSO regression	0.945	0.483	0.843	0.863
Random forest	0.972	0.492	0.860	0.857
Gaussian NB	0.687	0.828	0.718	0.842
Decision tree	0.875	0.581	0.811	0.754
Multinomial NB	0.832	0.411	0.739	0.690

ML: machine learning; AUC: area under the receiver operating characteristic curve; SVM: support vector machine; LASSO: least absolute shrinkage and selection operator; NB: naïve Bayes.

Table 3. In-hospital morbidity prediction performance of ML models for the test sample assessed using the bootstrap technique

ML algorithm	**Sensitivity**	**Specificity**	**Prediction accuracy**	**AUC**
Gaussian NB	0.605 (0.595-0.615)	1 (1-1)	0.717 (0.710-0.724)	0.803 (0.798-0.808)
Gradient boosting	0.972 (0.968-0.976)	0.712 (0.700-0.727)	0.902 (0.896-0.908)	0.842 (0.834-0.850)
Random forest	1 (1-1)	0.723 (0.708-0.737)	0.917 (0.911-0.922)	0.895 (0.889-0.902)
SVM “rbf”	0.948 (0.944-0.953)	0.344 (0.328-0.359)	0.782 (0.775-0.790)	0.646 (0.638-0.654)

Statistical measurements presented with mean (95% confidence interval).

ML: machine learning; AUC: area under the receiver operating characteristic curve; NB: naïve Bayes; SVM: support vector machine.

Table 4. In-hospital mortality prediction performance of ML models on the training sample assessed using five-fold cross-validation (sorted by the AUC value)

ML algorithm	**Sensitivity**	**Specificity**	**Prediction accuracy**	**AUC**
Random forest	0.644	0.993	0.892	0.960
Gradient boosting	0.738	0.932	0.876	0.951
Extra trees	0.680	0.978	0.882	0.949
SVM “sigmoid”	0.702	0.970	0.893	0.942
Ridge regression	0.851	0.849	0.849	0.939
LASSO regression	0.776	0.917	0.876	0.913
SVM “rbf”	0.756	0.948	0.893	0.911
SVM “poly”	0.776	0.940	0.893	0.908
SVM “linear”	0.776	0.948	0.898	0.907
Gaussian NB	0.716	0.894	0.844	0.890
Multinomial NB	0.664	0.925	0.849	0.871
Decision tree	0.685	0.863	0.811	0.813

ML: machine learning; AUC: area under the receiver operating characteristic curve; SVM: support vector machine; LASSO: least absolute shrinkage and selection operator; NB: naïve Bayes.

Table 5. In-hospital mortality prediction performance of ML models on the test sample assessed using the bootstrap technique

ML algorithm	Sensitivity	Specificity	Prediction accuracy	AUC
Random forest	0.636 (0.617-0.655)	1.0 (1.0-1.0)	0.955 (0.952-0.959)	0.818 (0.808-0.827)
Ridge regression	0.884 (0.872-0.896)	0.882 (0.876-0.889)	0.886 (0.880-0.892)	0.875 (0.869-0.882)
SVM “linear”	0.744 (0.726-0.761)	0.901 (0.895-0.906)	0.885 (0.880-0.891)	0.814 (0.806-0.823)

Statistical measurements presented with mean (95% confidence interval).

ML: machine learning; AUC: area under the receiver operating characteristic curve; SVM: support vector machine.

Figure 1

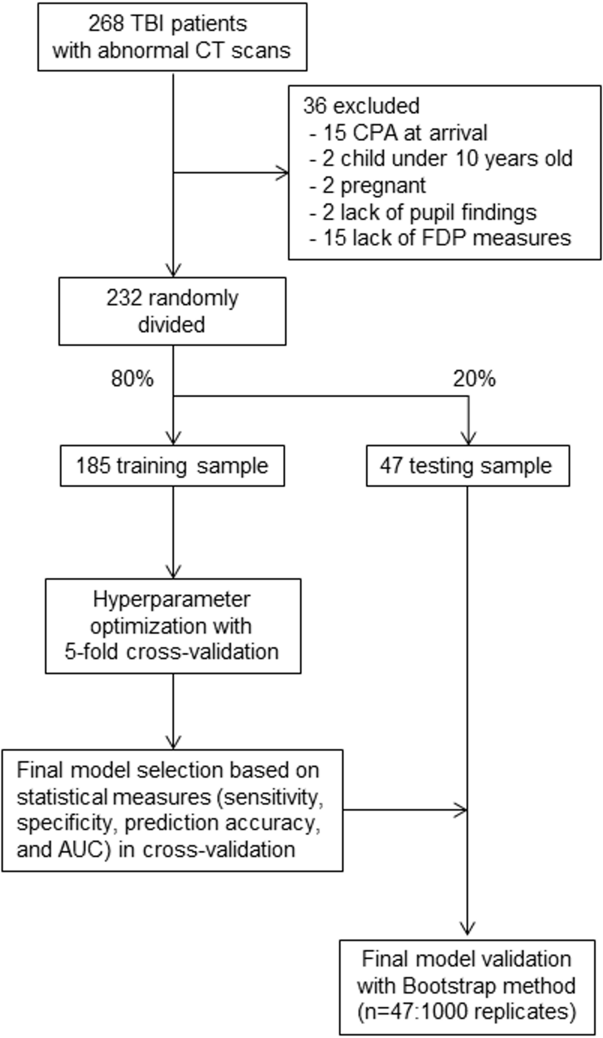

268 TBI patients with abnormal CT scans
36 excluded
- 15 CPA at arrival
- 2 child under 10 years old
- 2 pregnant
- 2 lack of pupil findings
- 15 lack of FDP measures
232 randomly divided
80%
20%
185 training sample
47 testing sample
Hyperparameter optimization with 5-fold cross-validation
Final model selection based on statistical measures (sensitivity, specificity, prediction accuracy, and AUC) in cross-validation
Final model validation with Bootstrap method (n=47:1000 replicates)

Figure 2

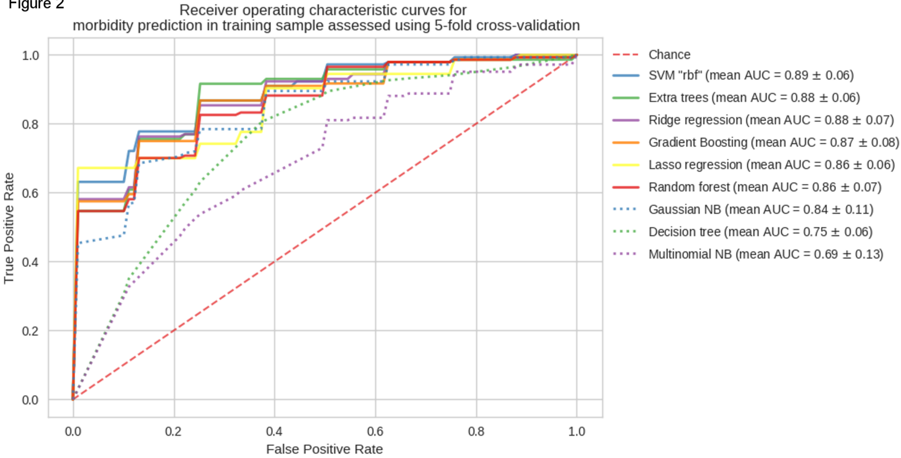

Receiver operating characteristic curves for
morbidity prediction in training sample assessed using 5-fold cross-validation
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Chance
SVM "rbf" (mean AUC = 0.89 ± 0.06)
Extra trees (mean AUC = 0.88 ± 0.06)
Ridge regression (mean AUC = 0.88 ± 0.07)
Gradient Boosting (mean AUC = 0.87 ± 0.08)
Lasso regression (mean AUC = 0.86 ± 0.06)
Random forest (mean AUC = 0.86 ± 0.07)
Gaussian NB (mean AUC = 0.84 ± 0.11)
Decision tree (mean AUC = 0.75 ± 0.06)
Multinomial NB (mean AUC = 0.69 ± 0.13)

Figure 3

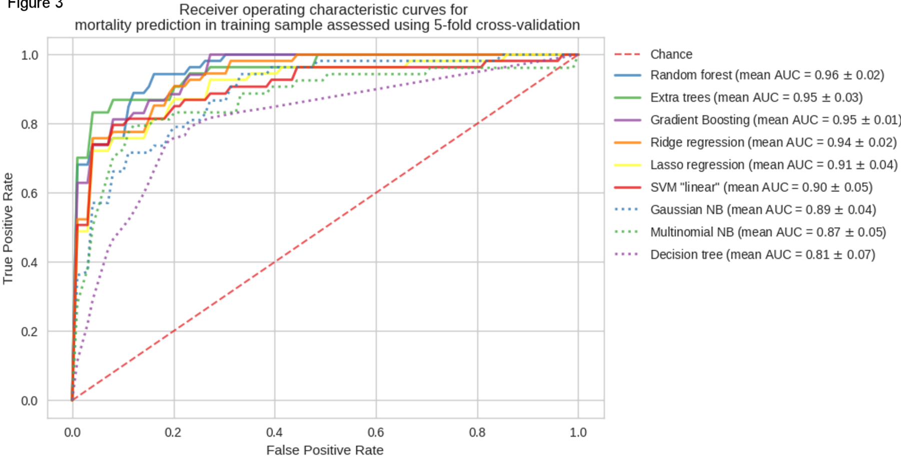
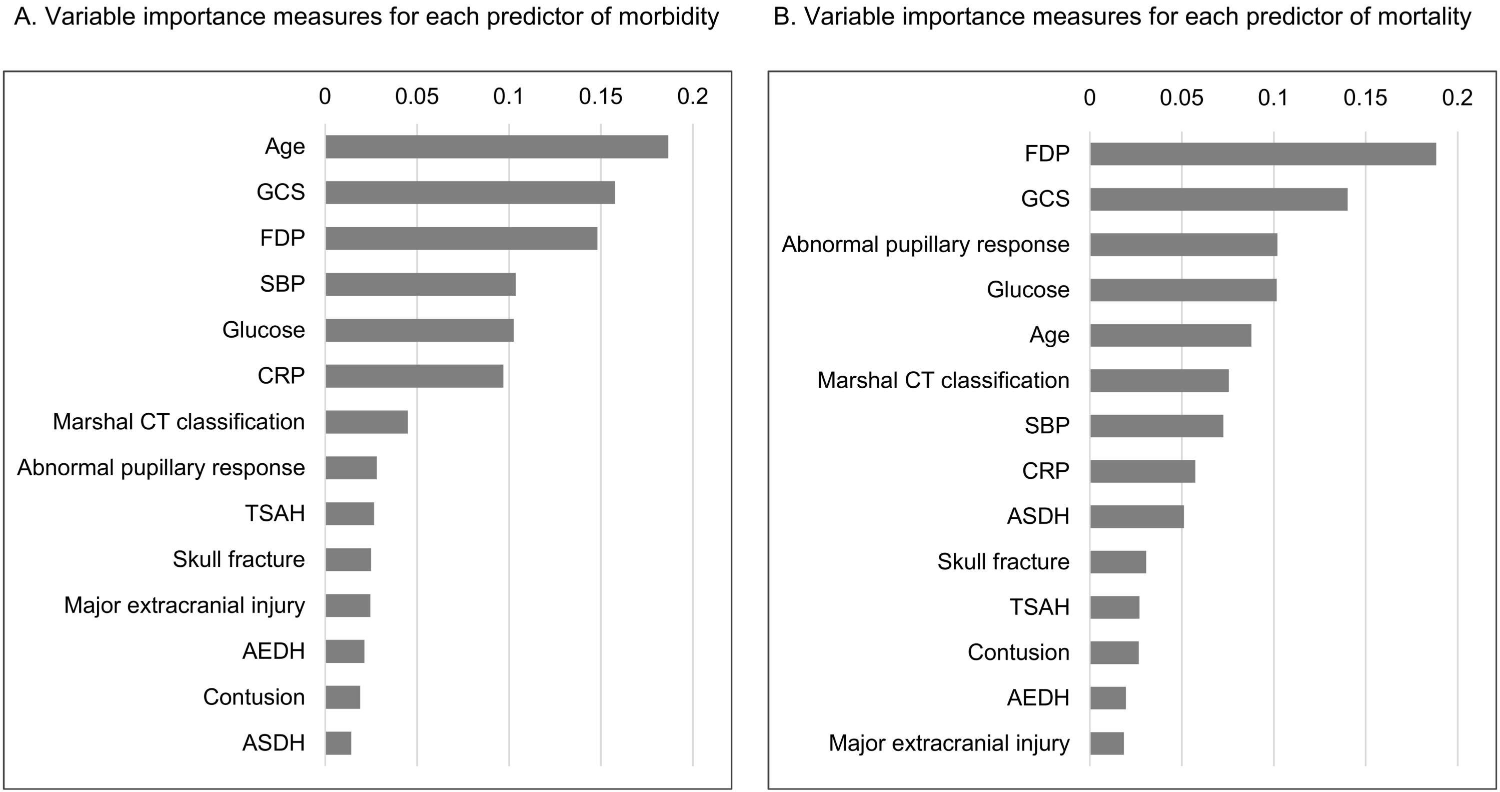

Receiver operating characteristic curves for
mortality prediction in training sample assessed using 5-fold cross-validation
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Chance
Random forest (mean AUC = 0.96 ± 0.02)
Extra trees (mean AUC = 0.95 ± 0.03)
Gradient Boosting (mean AUC = 0.95 ± 0.01)
Ridge regression (mean AUC = 0.94 ± 0.02)
Lasso regression (mean AUC = 0.91 ± 0.04)
SVM "linear" (mean AUC = 0.90 ± 0.05)
Gaussian NB (mean AUC = 0.89 ± 0.04)
Multinomial NB (mean AUC = 0.87 ± 0.05)
Decision tree (mean AUC = 0.81 ± 0.07)

Figure 4



A. Variable importance measures for each predictor of morbidity

B. Variable importance measures for each predictor of mortality

Appendix Table 1. Marshal CT classification score

Diffuse Injury 1	No visible intracranial pathology seen on CT scan
Diffuse Injury 2	Cisterns are present with midline shift 0-5 mm and/or lesion densities present; no high- or mixed-density lesion>25ml; may include bone fragments and foreign bodies
Diffuse Injury 3 (swelling)	Cisterns compressed or absent with midline shift of 0-5mm; no high- or mixed-density lesion>25ml.
Diffuse Injury 4 (shift)	Midline shift>5mm; no high- or mixed-density lesion>25ml
Evacuated mass lesion 5	Any surgically evacuated lesion
Non evacuated mass lesion 6	High- or mixed-density lesion>25ml; not surgically evacuated

Appendix Table 2. Tuned hyperparameters for ML models for in-hospital morbidity prediction

ML algorithm	**Hyperparameter**
Gaussian NB	no hyperparameters to tune

Gradient boosting	n_estimators = 250
	loss = "deviance"
	max_features = 2
	max_depth = 1
	learning_rate = 0.1

Random forest	n_estimators = 500
	max_features = ”log2”
	max_depth = 10
	criterion = "entropy"

SVM	kernel = "rbf"
	C = 10
	gamma = 0.01

ML: machine learning; NB: naïve Bayes; SVM: support vector machine.

Appendix Table 3. Tuned hyperparameters for ML models for in-hospital mortality prediction

ML algorithm	**Hyperparameter**

Random forest	n_estimators = 1500
	criterion = "entropy"
	max_features = 1
	max_depth = 5

Ridge regression	C = 0.01
	fit_intercept = True

SVM	kernel = “linear”
	C = 10

ML: machine learning; SVM: support vector machine.