Journal of Translational Medicine

RESEARCH

Open Access

Toward a hemorrhagic trauma severity score: fusing five physiological biomarkers

Ankita Bhat[1], Daria Podstawczyk[2], Brandon K. Walther[1,3], John R. Aggas[1], David Machado-Aranda[4,5], Kevin R. Ward[5] and Anthony Guiseppi-Elie[1,3,6,7*]

Abstract

Background: To introduce the Hemorrhage Intensive Severity and Survivability (HISS) score, based on the fusion of multi-biomarker data; glucose, lactate, pH, potassium, and oxygen tension, to serve as a patient-specific attribute in hemorrhagic trauma.

Materials and methods: One hundred instances of Sensible Fictitious Rationalized Patient (SFRP) data were synthetically generated and the HISS score assigned by five clinically active physician experts (100 [5]). The HISS score stratifies the criticality of the trauma patient as; low(0), guarded(1), elevated(2), high(3) and severe(4). Standard classifier algorithms; linear support vector machine (SVM-L), multi-class ensemble bagged decision tree (EBDT), artificial neural network with bayesian regularization (ANN:BR) and possibility rule-based using function approximation (PRBF) were evaluated for their potential to similarly classify and predict a HISS score.

Results: SVM-L, EBDT, ANN:BR and PRBF generated score predictions with testing accuracies (majority vote) corresponding to 0.91 ± 0.06, 0.93 ± 0.04, 0.92 ± 0.07, and 0.92 ± 0.03, respectively, with no statistically significant difference ($p > 0.05$). Targeted accuracies of 0.99 and 0.999 could be achieved with SFRP data size and clinical expert scores of 147[7](0.99) and 154[9](0.999), respectively.

Conclusions: The predictions of the data-driven model in conjunction with an adjunct multi-analyte biosensor intended for point-of-care continual monitoring of trauma patients, can aid in patient stratification and triage decision-making.

Keywords: Decision-making, Hemorrhage, Trauma care, DATA fusion, Risk stratification, Triage

Background

Trauma accounts for 47% of mortalities in individuals 1–46 years of age in the United States [1, 2]. Trauma-induced hemorrhage with its attendant peripheral vasoconstriction [3, 4] insulin resistance [5], hyperlactatemia, [6–8] acidosis [9], hyperkalemia [10, 11] and hypoxia can rapidly lead to death or may be followed by Multiple Organ Dysfunction Syndrome (MODS), a consequence of a "cytokine storm", which can also be fatal [9, 12]. The field triage decision scheme for the national trauma triage protocol provides guidelines to identify the status of the patient [13]. The physiological criteria includes identification of vital signs such as; systolic blood pressure (Hypotension < 90 mmHg), [14–16], abnormal respiratory rate (< 10 or > 29 breaths per minute) [13], abnormal heart rate (Tachycardia > 100 beats per minute) [17], and the Glasgow coma scale (≤ 13) [18, 19]. The Glasgow coma scale categorizes the patients according to the severity of brain injury. Simple Triage and Rapid Treatment (START) is the commonly used algorithm for mass casualty triage in the USA [20–23], which is used in conjunction with secondary triage for Secondary Assessment

*Correspondence: guiseppi@TAMU.edu
[1] Center for Bioelectronics, Biosensors and Biochips (C3B®), Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843, USA
Full list of author information is available at the end of the article

of Victim Endpoint (SAVE) when the resource supply is restricted [20]. START and SAVE employ criteria such as respiratory rate, cognitive function (ability to listen and respond to commands), and radial pulse to identify the category for triage. Another example is the Injury Severity Score (ISS) [24] based on the Abbreviated Injury Scale (AIS) system which aggregates the assessed injury to six regions of the body and establishes correlations with mortality and morbidity. American College of Surgeons Committee on Trauma (ACS COT) aims to improve care by supporting programs for injury prevention [25]. An additional data source is the MIMIC-III data set, a freely accessible critical care database attributable to Johnson et al. [26]. MIMIC-III represents global vital signs and physiological waveforms; it does not contain data for hemorrhaging patients consisting of molecular biomarkers such as glucose, lactate, potassium, pH and oxygen tension. A MODS severity score was developed by Marshall et al. [27], wherein a score (0–4) is applied following physiologic measurement of dysfunction in 6 organ systems (i) respiratory function (pO_2/FIO_2 ratio), (ii) renal function (serum creatinine), (iii) liver function (serum bilirubin), (iv) cardiovascular function (PAR), (v) Hematologic (Platelet count) and (vi) Neurologic (Glasgow Coma Score). The total number of input points were then added to achieve a score corresponding to the patient's ICU mortality %, hospital mortality %, and ICU stay.

Since the introduction of the MODS score, new rapidly deployable micro-analytical technologies have enabled measurement of key physiological indicators and the opportunity for the emergence of scores based on molecular biomarkers of physiological stress. A Hemorrhage Severity and Survivability Score (HISS) is herein introduced to allow for patient stratification. This stratification is made possible by the fusion of micro-analytical measurements of multiple physiological biomarkers [28]. HISS is a severity index intended as an adjunct to inform healthcare providers about the criticality of traumatic hemorrhage. This information would assist them in the delivery of timely and appropriate attention and care. HISS, therefore, is proposed to help in timely triage and in the stabilization of the most critically ill patients, and as a consequence, reduce patient mortality.

An adjunct device in the form of an indwelling biosensor system, the Physiologic Status Monitoring (PSM) Biochip, has been proposed and is under active development to help healthcare providers of trauma care in mass military and civilian triage situations [29, 30]. A dual-responsive biosensor for glucose and lactate has been proposed, designed, fabricated and successfully tested in rodent and piglet animal models of hemorrhaging trauma [31]. The PSM Biochip is a bio-SONDE; an indwelling device which measures, monitors and wirelessly transmits physicochemical information from within a victim of hemorrhaging trauma [29]. The bio-SONDE capable of acquiring the relevant physiological data pertinent to hemorrhagic shock states is the potential source of the data for subsequent fusion. When implanted intramuscularly, the PSM- Biochip enables the continuous, real-time monitoring of the patient's physiological status via the key biomarkers; glucose, lactate, pH, potassium and oxygen tension. Such a system has the potential to go beyond single immediate datum (stat) capability to reveal evolving and predicted temporal trend status. This bio-SONDE is combined with a wireless processing hardware and a software algorithm to enable data fusion from the five identified biomarker analytes. This system would potentially guide evidence-based decision-making [32] derived from the real-time pathophysiological profile of the patient.

The present work evaluates multiple data fusion algorithms and seeks to identify the minimum patient and expert data sets needed to arrive at accurate predictions. The goal is to arrive at reliable and confident patient stratification decisions using the HISS Score. The present focus is on molecular biomarkers of pathophysiology as supplements to the traditional gross indicators for the development of biomarker monitoring systems at this scale for disease states such as traumatic shock [33]. Molecular indicators such as changes in oxygen tension may be earlier indicators of physiological stress than global vital signs. There are clear biochemical interactions among the identified variables. For example, glucose and lactate are related via the Cory Cycle. Lactate and acidosis (pH) are directly related. Indeed, there are statistical interactions among these variables. For example, some variables are known to swing quite widely during hemorrhage, such as glucose and lactate. Other variables are early onset indicators while others are late onset indicators, such as acidosis and potassium.

The main contributions of this study are (i) the establishment of a novel HISS score, (ii) the generation and use of Sensible Fictitious Rationalized Patient (SFRP) data, (iii) the prediction of the size of the patient and expert data set needed to achieve 99.0 and 99.9% accuracy in the predictions of HISS scores, and (iv) recognition of the inter-expert variability and strong intra-expert consistency in expert scoring data. Here, multiple fictitious patient physiological status data are produced and multiple individual experts score the data. A key consideration is thus the real-time fusion of disparate pathophysiological data to yield an actionable HISS score. Such data integration has medical and biomedical engineering applications such as in rapid, wearable health monitoring and internet of things (IoT) monitoring [34–35]. Data fusion can also be applied to implantable devices

to generate data telemetry systems [36] with patient profiles [37]. Decision trees [38], support vector machine, neural networks [39], uncertainty index [40] and hybrid intelligent systems consisting of fuzzy logic and genetic algorithms [41] have been employed as classification approaches for data fusion in medicine. Decision tree classifiers were used to build a classification model in the form of a tree from the patient biomarker data [42]. The classifier provides a score for the data by testing each attribute and sorting and classifying particular instances in the data [43]. Ensemble bagged decision trees helped to reduce variance by the 'bagging' effect [44]. The support vector machine classifier [45] makes use of an optimal hyperplane and calculates the margin or the distance of the points from the hyperplane [46]. The points closest to the hyperplane are called the support vectors [47]. Support vector machines are often used because they are robust [48] and fast [45]. Neural networks mimic the structure of biological neurons, have input, output and/or hidden layers, and propagate to adjust the weights between the elements of the networks [49]. They are often used because of their value in tuning of data [50]. Genetic algorithms are employed to find an optimal solution for systems based on natural evolution [51] and have been used in time-series based neural networks [52] and in steady-state gene regulatory networks [53]. Similar approaches for the application of artificial intelligence in medicine and for developing a score for patients in the ICU [54] include the DeepSOFA [54], an automated alert functions for the patient status [55]. Decision support-systems employing an artificial intelligence clinician for sepsis in the ICU have also been generated [56].

Expert scoring of pathophysiological data may be incomplete or uncertain due to fuzziness or imprecision, and in some cases may be erroneous. Possibility theory [57] is a framework that is particularly applicable to expert knowledge. Unlike probability theory, possibility theory uses a pair of dual set-functions, namely possibility and necessity measures which make it capable of representing partial ignorance [58]. The possibility rule-based classification using function approximation (PRBF) algorithm has been shown to successfully handle the uncertainty in class labels of data and make an efficient use of the available data provided in the incomplete expert evaluation, a condition which is generally neglected in traditional supervised learning techniques [59]. Possibility labels may be directly extracted from an expert [60] by (1) the expert weighting the possibility of data belonging to each of the given c classes by a number between 0 and 1, or (2) to use possibility histograms from an empirical distribution of multiple expert opinions.

In the absence of actual viable penta-analyte patient data, synthetic data must be developed. Thus, a secondary objective of this work was to develop a synthetic data generation algorithm that produces Sensible Fictitious Rationalized Patient (SFRP) data. The SFRP algorithm creates a hidden seed layer and then generates biomarker values with filters to add noise/fuzziness and introduce variance to the five physiological biomarkers of interest. Practitioner input was sought in refining the filters and noise/fuzziness for each biomarker. The five biomarker values for each SFRP maps to a single output, the HISS score. The SFRP data were then shared with practicing physician experts who provided their individually rationalized HISS scores. Thus, the physicians' scores serve as the ground truth but carry the inherent uncertainty born from disagreement among experts. Multiple SFRP data sets scored by a single expert, allowed an assessment of intra-professional variance. Correspondingly, multiple physicians providing ground truths of a single SFRP data set allowed accommodation of inter-professional variations. Multiple physician experts, given the results of a single set of measurements of physiological biomarkers, evaluate the status of patients in the form of a HISS score. In the decision-making processes, which incorporates bioanalytical diagnostic data and expertly sourced scores, uncertainty is inevitable. That is, given a reported set of measurements of the five biomarkers for a patient, different physicians may provide different evaluations, i.e. different scores, to represent the status of the patient. In such cases, it is possible to represent the uncertain scores in the form of a range of values. The generated data were used to make predictions for the status of the hemorrhaging patients by training a decision tree classifier and rule-based evolutionary classifier [58] to handle uncertainty in scores. The results of training models are presented in terms of their prediction accuracies. Furthermore, this allowed forecasting of the size of the patient data set and the number of clinician experts required to achieve stratification accuracies of 99% and 99.9%.

Materials and methods

On-line data engines were searched for the availability of anonymized actual patient biomarker data for the hemorrhaging trauma patient (glucose, lactate, pH, potassium and oxygen tension). Public databases are available with specific datasets such as vital signs but they did not contain sufficient biomolecular data elements for the traumatic hemorrhage. Owing to necessary HIPAA-based security policies at hospitals, actual data for hemorrhaging patients could not be directly accessed. Access to diagnostic data sets under appropriate approvals is being pursued. Accordingly, the classification methods were

each employed on the synthetically derived Sensible Fictitious Rationalized Patient (SFRP) data.

Patient data generation-Sensible Fictitious Rationalized Patient data and evaluation by practitioners

In lieu of actual patient data, synthetic (SFRP) data sets were generated via a scripted algorithm in Python 3.7.0. The flowchart for the SFRP data set generator is based on the pathophysiological data in Table 1. The general algorithm begins with a seeded hidden layer of HISS scores that ranged from low(0) to severe(4). The initial seeding distribution for trauma scores was evenly distributed among the five levels. Each of the five biomarker values associated with each level was subsequently filled by randomly selecting a value from within a pathophysiological range that can be attributed to that trauma level (based on normal physiologic values and specific trauma and hemorrhage perturbations). The noise was introduced by controlling the relative level of deviation from initially seeded values into other trauma regimes—i.e. letting initially chosen values drift into other regimes not originally occupied by the primary, hidden trauma seed score. Glucose noise was based on potentially convoluting scenarios (adrenergic response) or by a simple, tunable probability of taking on a value, not within the seeded range. Lactate was similarly assigned. Potassium noise was added via post mathematical calculation. Acidosis (pH) noise was introduced by allowing for physiologically normal values to be taken at any hidden seed (with the rationale being that pH is a late and severe biomarker). Oxygen tension (pO_2) noise was introduced via a convoluting scenario (respiratory compensation based on pH—determined randomly) and simple, random noise. The algorithm, in the most direct sense, allowed for initial seed values to bleed over into other regimes and create data that was confounded. For proof of concept, random number generators of no bias were used—although extension into Gaussian and other distributions may be readily implemented.

The algorithm may be run for any number of synthetic patients to generate SFRP data sets for each. For assessment, the initial hidden layer seeding was not output and was not revealed to any evaluator of the datasets. Potassium was determined via the empirical relationship from Burnell et al. [61], which details that a 0.1 unit drop in pH raises the $[K^+]$ by 0.6 mM. This is implemented in pseudocode in the following way for each output in (1):

$$\left[K^+\right]_i = random\left(\left[K^+\right]_{normal}\right) + (7.35 - pH[i]) * 6 \quad (1)$$

where $\left[K^+\right]_i$ denotes the potassium of the ith patient, and $pH[i]$ that patient's pH level generated earlier. The random function action yields a normal potassium concentration within physiologic ranges, and is then altered if the pH of the patient is abnormal via the relation described. This concept is illustrated for the entries shown in italic in Table 1.

In this way a complete set of Sensible Fictitious Rationalized Patient data for n + 25 = 100 fictitious avatars (not patients) were created and ported into an excel spreadsheet for expert scoring and fusion considerations. Empirical relationships among the biomarker variables are possible and are being explored to enhance the robustness of the SFRP data sets. Table 2 shows a possible outcome for generating the training and testing data sets using the SFRP data generator. Accordingly, 100 unique Sensible Fictitious Rationalized Patient (SFRP) data sets were scored by five clinical experts. Each of the five experts assigned a HISS score, valued 0–4, to each pentaanalyte data set while providing a rationale for their selection of the assigned score for a particular patient (0 = LOW, 1 = GUARDED, 2 = ELEVATED, 3 = HIGH, 4 = SEVERE). This resulted in a multi-class/expert framework [62] for the model-based predictions.

Table 1 Bounded pathophysiological ranges of key biomarkers of physiological stress in the hemorrhaging trauma patient

Pathophysiological range			
Analyte	**Low**	**Normal**	**High**
Glucose	*Hypoglycemia* *< 3.88 mM* *< 70 mg/dL*	Euglycemia 3.88–5.50 mM 70-99 mg/dL	Hyperglycemia >5.50–10.00 mM 99–180 mg/dL
Lactate	Hypolactatemia <0.50 mM	Eulactatemia 0.50–2.00 mM	*Hyperlactatemia* *> 2.00–4.00 mM*
Potassium	Hypokalemia (< 3.50 mM)	Eukalemia 3.50-5.50 mM	*Hyperkalemia* *(> 5.50 mM)*
pH	*Acidosis* *(< 7.35)*	7.35-7.45	Alkalosis (> 7.45)
pO_2	*Hypoxia* *< 5.18 mM* *< 100 mmHg*	5.18-6.22 mM 100-120 mmHg	Hyperoxia (> 6.22 mM) >120 mmHg

Italicized entries relate to an example implementation of the SFRP data generator, explained further in the text

Classification algorithms

Data from different sources can be fused via estimation, association and decision fusion [63]. Multi-class [64] linear support vector machines (SVM-L), ensemble bagged decision tree (EBDT), artificial neural network with Bayesian regularization algorithm (ANN:BR) and possibility rule-based using function approximation (PRBF) classifiers were used to classify the SFRP data sets. Figure 1 shows the concept for a fused score from the data of the five biomarkers. Five unique data sets, each of size 100, corresponding to the pathophysiological profile of 100 fictitious patients

Table 2 Partial data set for "fictitious patients", including training data set (1 to n) and testing data set (n + 1 to n + 25) generated using the Sensible Fictitious Rationalized Patient (SFRP) data generator and corresponding expert assigned Hemorrhage Intensive Severity and Survivability (HISS) score

Fictitious Patient	Sensible Fictitious Rationalized Patient (SFRP) Data					HISS				
	Glucose (mg/dl)	Lactate (mmol/l)	pH	Potassium (mmol/l)	pO_2 (mmHg)	D1	D2	D3	D4	D5
1	70	2.7	7.42	5.10	78	1	1	1	1	0
2	160	6.0	7.11	6.14	44	4	2	3	3	3
n	41	9.7	7.26	4.84	97	3	3	4	3	3
..	..	..	..	..	..	..	..	..	..	..
n + 1	123	3.3	7.41	5.00	86	UD	UD	UD	UD	UD
n + 2	49	8.7	7.13	5.92	53	UD	UD	UD	UD	UD
..	..	..	..	..	..	..	..	..	..	..
n + 25	220	8.6	7.23	4.52	92	UD	UD	UD	UD	UD

UD = undeclared i.e. assigned by the experts but predicted by the algorithms

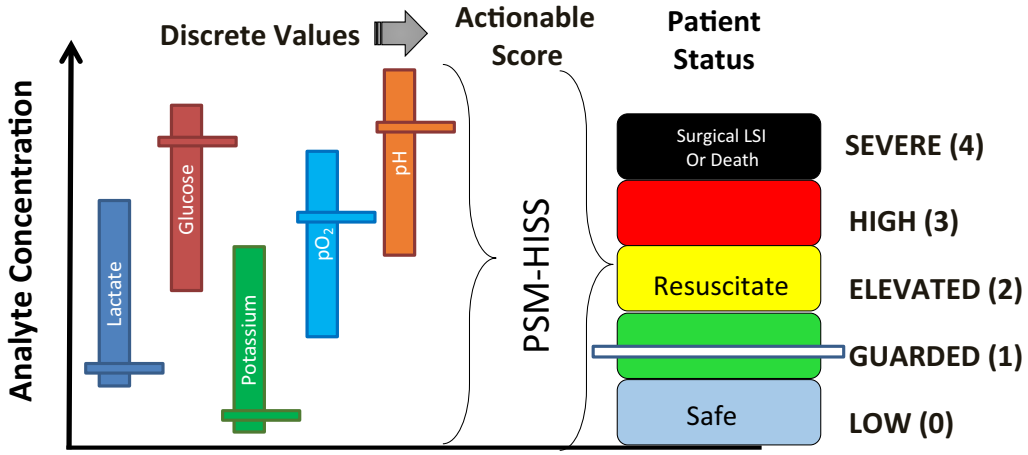

Fig. 1 Immediate and continual measurement of key biomarkers may serve as a "gauge" for identifying shock states. Discrete values from the five indwelling biosensors are fused into a single actionable Hemorrhage Intensive Severity and Survivability (HISS) score [1]. This score is further stratified into five, color-coded levels, 'SEVERE (4)' being the most critical and 'LOW (0)' corresponding to expectant

and along with the HISS scores of five healthcare provider experts: [100][D1], [100][D2], [100][D3] [100][D4] and [100][D5] were thus created from the available 100 penta-biomarker, patient data sets.

The multi-class ensemble bagged decision tree and linear support vector machine classifiers were used for predictions over the entire data sets (D1–D5). Neural networks [65] were used to determine the adequate number of training size for accurate predictions over the five data sets. Possibility rule-based classifiers were used to capture the uncertainty in the responses of the experts over the five data sets.

Multi-class linear support vector machine and ensemble bagged decision tree classifiers

For both support vector machine and decision tree classification models, a set of hyper-parameters was tuned and the model with the highest test accuracy was chosen to be reported. The cross-entropy was employed as the selection criterion at each node. Moreover, for the bagged decision tree algorithm the number of estimators was selected from [6, 20] with step 2. The ensemble technique (bagging) was applied in order to reduce an error of DT, as the combination of several weak predictors into

one high-quality ensemble model improves predictive performance [66]. For the SVM model, different kernel functions (linear, polynomial, radial base function, and sigmoid) were tested. In the case of the polynomial kernel, the degree of the polynomial was selected from [2, 6]. To train each model, fivefold cross-validation was employed and the average test accuracies along with the standard deviation of the accuracies was reported. Classifiers were trained using Python Scikit-learn [67] library.

The computations were performed using MATLAB R2019b Classification Learner App run on a PC.

Artificial Neural network with Bayesian regularization algorithm for sorted and unsorted data

ANN:BR has the advantage of tuning the incoming patient data sets. The term epoch in ANN is defined as the measure of the number of times all of the training vectors were used to update the weights [68]. The softmax activation function was used to introduce non-linearity into the model. The inputs were turned into a linear model (ωx + b), where ωx is the matrix multiplication of weights (ω) and inputs (x) and b is the bias. The scores obtained from this step were fed into the softmax function (2) which converts them into probabilities.

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} for j = 1, 2 \ldots, k \tag{2}$$

Softmax function maps the set of outputs onto inputs. In this case, there are five outputs which when passed through the softmax function get distributed according to probability (0,1). This was useful for finding the most probable occurrence or classification for a particular output.

A Bayesian regularized neural network (ANN:BR) capable of classifying patients using an assigned HISS score was developed in MATLAB 2018a Neural Network Pattern Recognition App run on a PC [69]. The neural network was trained to a max epoch size of 100 using Bayesian regularization algorithm [70] (training stops according to adaptive weight minimization) using Mean-Square Error as the performance metric. Responses from the five experts were used to create a single label by calculating the mathematical mode as the best metric of central tendency. The mode was chosen over the mean due to possible skew in HISS scores. The NN was trained using (i) sorted and (ii) unsorted data. Sorted data served to ensure that HISS scores were normally distributed among the training and test data. Sorting established groups of 5 different patient data using the 80:20 rule (e.g. 80% of HISS score "1" was used in the training set, while 20% of HISS score "1" was used in the test set). Unsorted data employed no such grouping and hence carried the risk that the test data could be unbalanced in its representation of certain HISS scores. Neural network performance was measured by using a constant test set size of 25 with 4 or fivefold cross-validation, where the training set size varied from 15 to 75% of the total data set size. In yet a totally different and additional approach, the mean test and mean training accuracy were determined by varying the size of the training set between 30 and 80 instances with steps of 5. To test the trained models, a fixed set of instances of size 20 was used. The experiment for each training set size was repeated twenty times to reduce the effect of variance on the results and the mean accuracy was reported. This served as a self-consistent approach across all classifier algorithms.

Possibility rule-based classifier

In the possibility rule-based classifier system using function approximation (PRBF) [71], possibility theory is used to handle uncertainty in expert knowledge. The degree of belonging of an instance to the kth class may be characterized by $u^k \in [0, 1]$. Different theoretical frameworks have been proposed to solve problems that suffer from uncertainty [72] including probability theory, set theoretic functions, and possibility theory. Under the possibility theory [71, 73] framework, u^k is the level of possibility that the given data point belongs to the class of score k and the following representation holds for the set of possibilistic classes assigned to the ith instance:

$$\boldsymbol{u}_i = \left(u_i^1, u_i^2, \ldots, u_i^c\right) \forall u_i^k \in [0, 1] \tag{3}$$

In (3), c is the number of scores defined for the problem, i.e. in this case five scores, being LOW (0), GUARDED (1), ELEVATED (2), HIGH (3), SEVERE (4). Unlike the probabilistic labels, the values of the vector $\boldsymbol{u}_i$ do not have to sum up to unity. Instead, each parameter takes a value ranging from 0 to 1. The classification scheme proposed by Nazmi and Homaifar [71], namely, possibility rule-based classifier using function approximation (PRBF), employs this definition of a class assignment and trains a rule-based evolutionary model that given a data point, predicts the degree of possibility to which the SFRP data set belongs to each of the possible classes.

The possibility rule-based classifier was implemented using Python 3.7.5 run on a PC. A fivefold cross-validation was used with population size = 4000, stretch = 25, learning rate = 0.1, and training iterations = 100,000. Having the assigned scores from five expert physicians for the generated SFRP data sets, it is probable that any two physicians might disagree on the score of any one patient's values or the same physician assigns different scores to patient's values that are nearly similar. This problem may be addressed with the use of possibility

theory [71], capturing the inherent intra-expert and inter-expert variation in the responses of physicians. More specifically, scores provided by the physicians for each set of measurements, the SFRP data set, were converted into possibility values that were values between 0 and 1. For a given measurement vector $\mathbf{x}$ and a hypothetical class, ω^k, the possibility distribution, $\pi_{\mathbf{x}}$, defined for $\mathbf{x}$ represents the knowledge contribution of an information source about the actual state of $\mathbf{x}$. In other words, $\pi_{\mathbf{x}}(\omega) = 0$ means that state ω is rejected as impossible, and $\pi_{\mathbf{x}}(\omega) = 1$ means that state ω is totally possible (plausible). In a machine learning framework, this concept is employed to solve classification problems by taking $\pi_{\mathbf{x}}$ to represent the degree of belonging of SFRP data to classes which are provided by the expert(s) [60].

PRBF has two main mechanisms to generate a problem solution; a rule-based evolutionary algorithm to approximate possibility labels, and an information fusion method to make plausible inferences for unseen data. When trained on a dataset with possibility labels, PRBF iteratively evolves a population of overlapping rules which are piece-wise linear approximations of the target possibility distributions. Moreover, the data fusion technique employed in PRBF combines the data provided by multiple sources, i.e., rules of the model, and calculates the most plausible values for the class membership of the unseen data set. Consequently, for an unseen patient data set, the model generates a possibility distribution (π). This distribution may then either be interpreted by an expert for decision-making purposes or processed to extract a crisp class by taking the one with the highest possibility. To demonstrate the benefit of employing a model that is robust in the presence of HISS score uncertainty, the same training data that were generated using SFRP data generator were used to train the PRBF algorithm and the trained model was evaluated against the 100 instances used in the previous sections for the model evaluation. The disagreement among the physicians' evaluations, was captured by repeating the process for all of the 100 SFRP samples by calculating a set of possibility labels as well as a class label based on the majority vote.

Performance metric, cross-validation, adequacy of patient data size and predicted patient data size with the number of experts

In general, the performance of a multi-class classification can be measured using accuracy, precision and F-score [74]. A confusion matrix plot can be used to evaluate the quality of the classifier [75]. The matrix contains values corresponding to true labels and predicted labels. The values in the major diagonal of the confusion matrix can determine how well the classifier has performed. In this work accuracy was used as a performance metric to report the prediction performances, which can be obtained from the major diagonal elements of a confusion matrix as follows in (4),

$$\begin{aligned} Accuracy &= \frac{\#\ Correct\ predictions}{\#\ predictions} \\ &= \frac{\Sigma\ of\ elements\ in\ the\ major\ diagonal}{\#\ of\ elements} \end{aligned} \tag{4}$$

Cross-validation [76, 77] helps with using all the available data for model training and hence in making more robust predictions. To do so, the data were randomly split into equal sets for training of multiple models. Here a fivefold cross-validation [78] was used. The adequacy for the patient data size was tested with the minimal point for stabilizing validation accuracy. The adequacy for the number of experts and the prediction for the patient data size for a test accuracy of 0.99 and 0.999 with the predicted number of experts necessary to achieve that accuracy was arrived at using the regression model fit and application of predictive modeling in JMP Pro software version 14.0 run on a PC.

Comparison of classification algorithms

The classification algorithms employed in the previous sections were compared for their respective accuracies. In this case, DT, SVM and PRBF classifiers were trained according to the approach presented in "Multi-class linear support vector machine and ensemble bagged decision tree classifiers" and "Possibility rule-based classifier" sections, respectively. While for ANN:BR, the number of nodes was selected from [5, 63] with step 5. Different activation functions were tested and a 'tanh' function was selected. The solver that was used to train the models was the 'adam' solver and the model was trained for 100,000 iterations. Finally, for the PRBF model, the maximum number of rules was selected from {500, 1000, 3000, 4000, 5000, 6000}. The maximum condition stretch was selected from [20, 27, 32, 34, 38] which modifies the proportional size of the rule condition and effects the accuracy of the rules. The learning rate was set to 0.1, and the number of training iterations was 30,000.

To train the PRBF algorithm, the uncertain labels (u) were used and the other classification algorithms were trained on D1–D5 and using the majority vote of the labels obtained from the five physician experts. For the decision tree classifiers, support vector machine and the neural network, their Python implementation that was available in Scikit-learn [67] library was used. For the PRBF algorithm, its implementation in Java was used. All experiments were carried out on a 2.70 GHz Windows 10 machine with a 16.0 GB RAM. One-way Analysis

of Variance (ANOVA) was used to determine the significance levels for the performance of these algorithms using JMP Pro software version 14.0 run on a PC.

Results

Classification via linear support vector machine and ensemble bagged decision tree

Two well-established classifier algorithms, namely linear support vector machine (SVM-L) and ensemble bagged decision tree (EBDT), were used in the classification of SFRP data. Figure 2a, b provide the accuracy versus

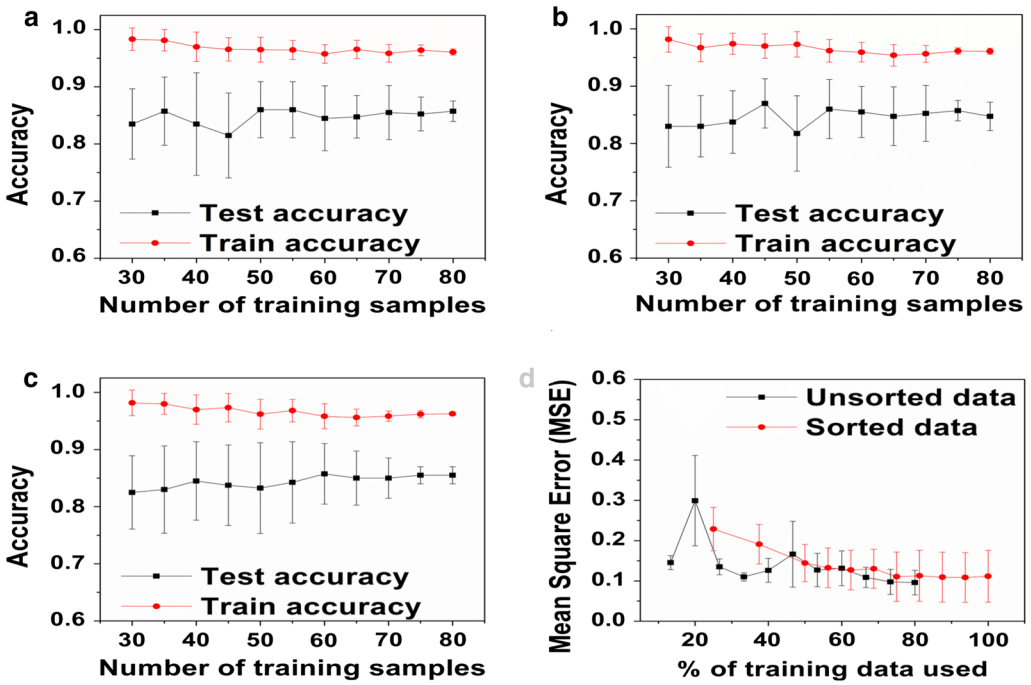

Fig. 2 Evaluation of the mean test and mean train accuracy versus the number of training samples for the training size varied from 30-80 in steps of 5. Accuracy vs. training dataset size for **a** SVM-L, **b** EBDT, **c** ANN:BR, and **d** Evaluation of the influence of the size of the training set, expressed as a % of available data, on the performance of the ANN:BR as expressed in the Mean-Square-Error for maximum epochs of 100. For unsorted data, trained ANNs were tested with a sliding window of 25 validation data set. For sorted data, trained ANNs were tested with a sliding window of 20 validation data set

Table 3 Application of two different algorithms (linear support vector machine and ensemble bagged decision tree) to the five(5) unique SFRP data sets; [100][D1], [100][D2], [100][D3], [100][D4] and [100][D5]

Class	Frequency (%)				
	D1	D2	D3	D4	D5
0	56	43	37	43	53
1	14	20	27	18	17
2	5	18	7	15	13
3	19	17	11	24	17
4	6	2	18	0	0
SVM-L accuracy (%)	78.3 ± 0.5	92.7 ± 0.5	78.3 ± 2.4	88.3 ± 0.5	86.7 ± 0.9
EBDT accuracy (%)	83.3 ± 1.2	96.3 ± 0.9	72.3 ± 0.9	90.0 ± 0.0	87.7 ± 1.2
Class with the highest confusion (TPR—sensitivity for EBDT)	4 (17%)	4 (0%)	2 (14%)	2 (60%)	2 (77%)

The table presents the fraction of the total observations for each class for each dataset and corresponding cross-validated accuracies for both classifiers. The confusion (true positive rates (TPR)) was correlated to the percentage of observations of the class

the number of training samples for SVM-L and EBDT. Table 3 presents the findings of the SVM-L and EBDT classifiers in terms of their validation accuracy.

Analysis of each experts' model individually, revealed that EBDT generally performs better than SVM-L (Table 3). Although the differences between both predictors for each dataset were slight (in the range of 2–6%), when it comes to patient stratification decisions, small improvements may be consequential to the therapeutic intervention for a patient. The highest cross-validated accuracy was achieved for the expert D2 dataset and the EBDT classifier ($96.3 \pm 0.9\%$). However, a confusion matrix revealed that D2 failed completely to predict Class 4 (Severe), as 100% of labels were misclassified. Among all experts, the highest confusion (TPR) occurred for Class 2 (Elevated), which was the most frequently misclassified as either Class 1 (Guarded) or 3 (High) and for Class 4 (Severe) misclassified as Class 3 (High). There was no single instance where all the five experts concurred on the score of 2. This is due to the fact that 2 is a score in the mid-range of 0–4 and hence higher variability for this score was introduced compared to the extremities [79]. As shown in Table 3, a high level of misclassification may result from an imbalanced number of instances in each class. For example, for expert D1, only 6 instances out of 100 were labeled with Class 4 (Severe), which leads to only 17% TPR. For D3, only two data rows were labeled as Class 4 (Severe), which caused complete misclassification of this score (0% TPR). While for D4 and D5, despite the high performance, none of the input instances were scored as Class 4 (Severe), leading to a model which will fail to make predictions of this class for the new data. Bagging classifiers may reduce the misclassification rate and improve overall accuracy of algorithms. Thus, the EBDT classifier, while being more time-consuming, performed with high accuracy compared to the SVM-L. However, the support vector machine classifiers had a higher accuracy for the data set D3 whereas the decision tree algorithm was less effective in capturing the localized accuracy of D3. From the literature, accuracies of 83–88% for SVM [80], and accuracies of 70–83% have been reported for decision trees in medical applications [81].

Classification via artificial neural network classifier

With a constant size of 25 validation data sets, it was observed that the error increased with increase of the SFRP training data sets. From the literature, it is known that the error should have stabilized or be shown a decrease to some extent with increasing training data sets [82]. This was attributed to the difference in the opinions of the experts. Consequently, a sliding window of validation data sets was used. Using the approach of Mode to allow the doctors to vote together along with a sliding window of validation data sets showed a decrease in error with an increase in the training data sets, in agreement with the literature as shown in Fig. 2d [83]. Here it was observed that sorting improved output quality with a smooth trend towards equilibrium or limiting error. However, the unsorted data appeared chaotic with stochastic noise.

Unsorted data at a very low training set size (15–20%) showed a very high standard deviation of MSE due to a lack of heterogeneity within the training set. The probability of less frequently available HISS scores, such as 4, being withheld from the training set was very high. For example, at a training set size of 20%, the probability of a HISS Score of 4 showing up in the training set was 0.008. Unsorted data tended to the same MSE as sorted data (~ 0.12), but was variable in its descent due to probabilities of scores not being included in the training set because if low frequency (e.g. HISS Score 4). As shown in Fig. 2d, improvement in the test accuracy of the ANN:BR was insignificant for the number of SFRP training samples larger than 75. Based on Fig. 2c, d, 75 SFRP data sets were established to be an adequate data size to build a model for prediction. From the literature, prediction accuracies of 44% and training accuracies of 50% and above have been reported for neural networks in medical applications [84, 85]. When the accuracies were on a lower side, the neural network approaches were often combined with hybrid fuzzy systems [86].

Performance of PRBF

One simple way to resolve conflict among class evaluations from multiple experts is to take the class label that was most frequently identified. An alternative approach, which makes better use of the rich data provided by the experts, is to calculate a set of possibility labels using (3), which is expected to reflect the disagreement among experts better than solely taking the majority vote. Table 4A depicts the fivefold training and validation accuracies from experts D1–D4, for the PRBF algorithm. For the sample presented in Table 4B, the majority vote opts for class zero to represent the patient's status, as shown in Table 4. The possibility labels calculated using the (3) for the same patient data are provided in column 'Uncertain labels' in Table 4B. The uncertain labels assume the association of the patient data to class zero and one. The degree of possibility that the sample belongs to each class is different however and is equal to 1 and 0.5 for class zero and one, respectively. This graded association reflects the disagreement among the experts in deciding the true status of the patient.

A confusion matrix plot was used to represent the performance of the possibility rule-based classifier using

Table 4 Results from PRBF algorithm from experts D1-D4. A) Cross-validation model training results for PRBF algorithm for Population size = 4000, stretch = 25, learning rate = 0.1, and training iterations = 100,000, B) True labels and predicted uncertain labels for the tested SFRP sample of fictitious patient number 72

A

	Training accuracy	Test accuracy
Fold-1	0.95	0.90
Fold-2	0.96	0.90
Fold-3	0.98	0.95
Fold-4	0.95	0.95
Fold-5	0.94	0.90
Mean accuracy	0.96	0.92
Standard deviation	±0.01	±0.03

B

Fictitious patient	Majority vote	Uncertain label (u)	PRBF prediction (π)
72	0	[1,0.5,0,0,0]	[0.979,0.321,0,0,0]

function approximation (PRBF) [75]. Folds indicate the division of the data set to confirm that each of the folds has been used as a set. Uncertainty or error information has been utilized to support medical diagnostics where a prediction accuracy of 87% has been reported for a training accuracy of 90% [87]. Common approaches like possibility rule-based classification for handling error include fuzzy probabilities [88], and hybrid fuzzy-NN systems [89].

The PRBF model was able to predict HISS scores with 92% accuracy for a testing and training accuracy of 96% when using D1–D4. These results confirm that the idea of integrating evaluations from multiple experts and modeling them with a proper uncertainty handling tool, which is possibility theory in this work is beneficial for decision-making. Note that by increasing the number of training samples of the SFRP data sets, the model will be better trained and able to produce more accurate predictions.

Comparison of the test accuracies of classification algorithms

The performance of the four classification algorithms, linear support vector machine (SVM-L), ensemble bagged decision tree (EBDT), artificial neural network with Bayesian regularization algorithm (ANN:BR) and possibility rule-based using function approximation (PRBF) were compared for their ability to accurately classify the SFRP data sets. Figure 3a lists the test accuracies and Fig. 3b shows the misclassification rates for the classification algorithms and the uncertainty labels of PRBF algorithm for different experts and the majority vote. The highest accuracy is highlighted in bold for each

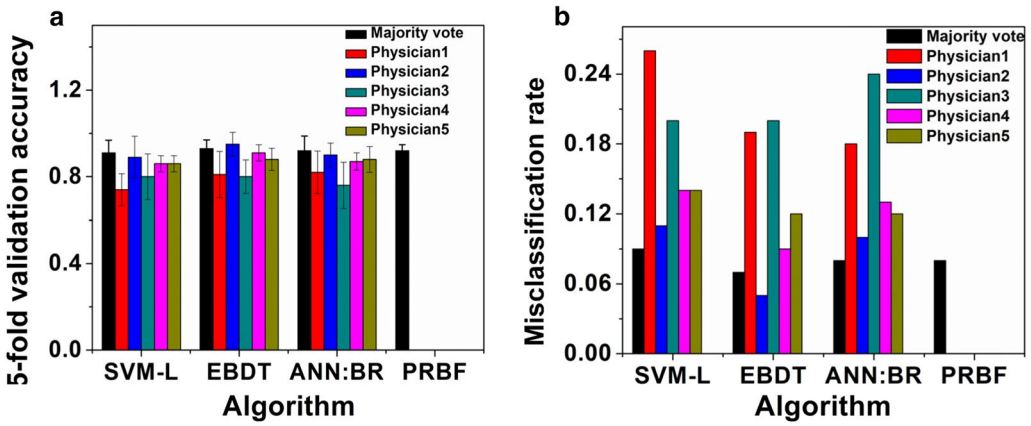

Fig. 3 Comparison of **a** Test accuracies, and **b** Misclassification rate, of SVM-L, EBDT, and ANN:BR for experts D1–D5 as well as the majority vote, along with the uncertainty labels of PRBF algorithms for experts D1–D4

algorithm. SVM-L, EBDT, ANN:BR and PRBF generated score predictions with testing accuracies (majority vote) corresponding to 0.91 ± 0.06, 0.93 ± 0.04, 0.92 ± 0.07, and 0.92 ± 0.03, respectively, with no statistically significant difference ($p > 0.05$) in their means for $\pm 95\%$ confidence interval (C.I).

Predictions for the adequacy of the patient data size and number of experts for improved accuracy

It is reasonable to ask, given the scoring accuracies obtained for the 100 patients and 5 physician experts 100 [5], what data set size and how many experts will be required to improve scoring accuracies? Targeted accuracies of 99% and 99.9% could be achieved with SFRP data size and clinical expert scores of 147 [7] (99%) and 154 [9] (99.9%), respectively. The model fit for 99% was for a $R^2 = 0.96$ with the Total Sum of Squares (SS_{total}) as 0.04 with a statistical significance of $p \leq 0.05$ for a $\pm 95\%$ confidence interval (C.I). The model fit for 99.9% was for a $R^2 = 0.89$ with the Total Sum of Squares (SS_{total}) as 0.11 with a statistical significance of $p \leq 0.05$ for a $\pm 95\%$ confidence interval (C.I).

Discussion

Evaluation of individual classifiers

The collection, labeling and archiving of medical data is usually time-consuming, expensive and fraught with security concerns, appropriately so, [77]. Therefore, it is a challenge to build predictive models based on limited available training data. Moreover, the labels are often provided by multiple experts, who may have different opinions about the same patient's health status. Such disagreement may result from differences in experts' knowledge and clinical experience. As a worst-case scenario, differences in opinions may lead to patient misclassification, which may have serious consequences to their health [90]. The goal of the present study was to produce and use 100 instances of Sensible Fictitious Rationalized Patient data in the development of predictive models for patient stratification, to use expert opinion to achieve the same stratification in order to ground truth the predictive models and to engage cognizance of intra-expert consistency and inter-expert variability. The study revealed that the more imbalanced the input data, the higher the misclassification penalty. The similarity in misclassification (high level of misclassification of Scores 2 and 4) for each dataset and for both EBDT and SVM-L classifiers, may be the result of insufficient information provided to perform reliable labeling. It is, therefore, extremely important to compare various classifiers in terms of not only their accuracy but also their level of misclassification as performed in this study.

Qualitative evaluation of experts' HISS scoring

As a pilot study, the opinions of five experts, D1– D5, were obtained. Expert 1 based his bias weighing decisions on the abnormal levels of biomarkers, being driven by the extremes. For example, when the lactate levels were high, with potassium elevated but compensated, but with a normal pH, this produced a HISS score of 1. It is observed that a score of 2 was assigned when the lactate level does not correlate with other values (normal pH, Eukalemia, Euoxia). High lactate, very low glucose, low pH and normal oxygen produced a HISS score of 3. All values very deranged with pH almost out of physiologic non-recoverable range; hypoxia below 60, elevated lactate, potassium elevated suggesting cell injury, resulted in a HISS score of 4. While providing the scores, expert 2 was able to pick the ones that were similar. Hence, his scores were consistent across all different profiles. The scoring pattern of expert 3 was not localised. Expert 4 localised his scores from 0 to 3. While this paper is not concerned with expert performance, and the data set was far too small to allow the analysis of experts, the very low intra-expert variability (8.0%) and larger inter-expert variability (20.6%) is worthy of mention.

Comparison of classifiers in terms of cross-validation accuracy

By the majority vote, SVM-L, EBDT, ANN:BR, and PRBF had cross-validated accuracies of 0.91 ± 0.06, 0.93 ± 0.04, 0.92 ± 0.07, and 0.92 ± 0.03 respectively. The results for SVM-L, EBDT, and ANN:BR were statistically significant. The misclassification is more prominent among the middle classes of 2 and 3. For example from Fig. 4, misclassification rates were 71% for the class of 2 for SVM-L This is because the experts converge upon the extreme values but may have an overlap in the middle classes. There is a need to better delineate the intermediate shock states and so successfully intervene with appropriate resuscitation measures. The accuracy of these intermediate states can be improved by expanding upon the number of expert participants. We hope to improve accuracy of these intermediate states by expanding upon the number of expert participants. Linear support vector machine (SVM-L) and ensemble bagged decision tree (EBDT) classifiers provided for classification in a simple hierarchy of a tree structure and SVM-L provided robust classification. An unquestionable advantage of the presented decision tree classifiers is that they are simple and rapid prediction tools which establishes the trauma severity score with a high accuracy. The results showed that the decision tree classifiers constitute a reasonable basis for the further extensive studies on more specific and complex prediction approaches which may overcome the

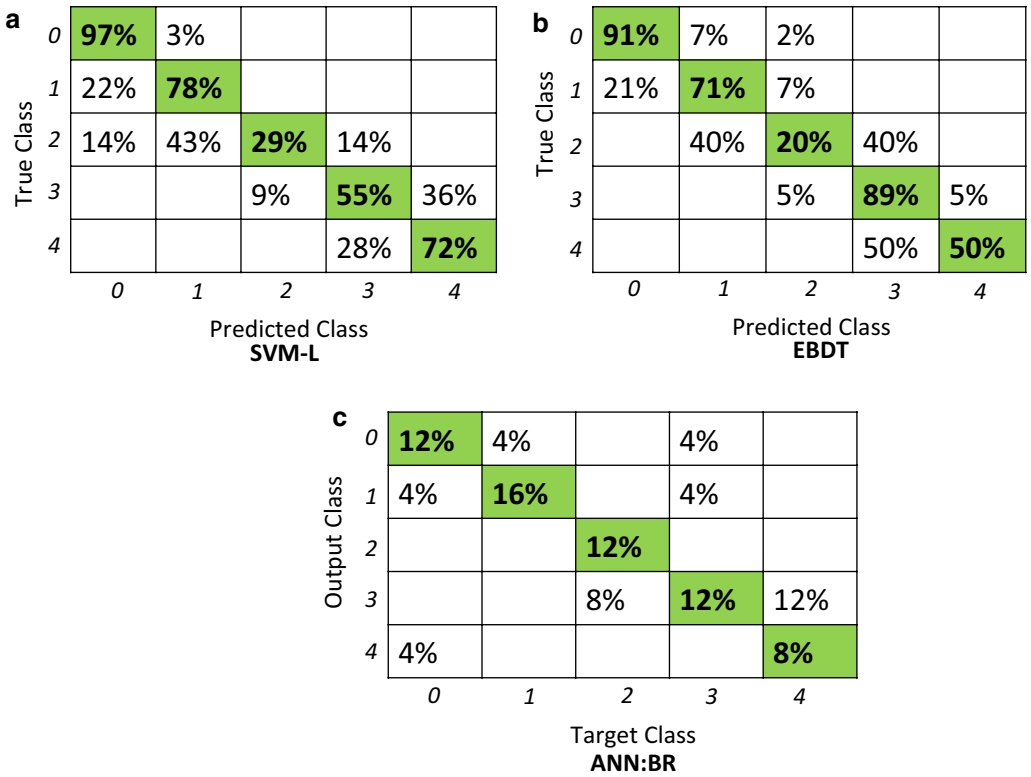

Fig. 4 Representative confusion matrices for SVM-L, EBDT, ANN:BR

limitations of the current methods such as a lack of external validation of the model, experts' opinion, or variation.

Performance of PRBF relative to other classifiers

The PRBF classifier added a layer to the intra- and inter-expert variabilities addressed by the other classifiers by tapping into the votes (either majority or individual) of the experts for a particular patient data set and reporting the number of times a physician's label agrees with the consensus. It was interesting to note that expert 4 had the highest concurrence from his scoring pattern localized for 0–3. This coincides with real-life scenarios when expert physicians try to categorize the patients from 0 to 3 and try to save them. Comparatively, the score of 4 corresponding to severe was rare. From the fivefold cross-validation results, the improvement in the test accuracy is insignificant for the number of training samples larger than 70. The increase in the training samples from 30 to 70% improved the accuracy from 71 to 78.5%.

The PRBF model seeks to incorporate the inherent disagreement among the physician experts into the model training procedure. According to Fig. 3a, integrating evaluations from multiple physicians through the possibility theory resulted in a better performance than SVM-L, EBDT, ANN:BR, and PBRF trained using the majority vote. This implies that employing different tools of modeling the uncertainty, allows for capturing different forms of uncertainty and potentially leads to better prediction accuracy. Moreover, training a model using PRBF allows for an additional level of interpretation of the model prediction during the decision-making process. To illustrate this point, consider the example of Fictitious Patient 72 presented in Table 4B. When the trained PRBF model was elicited for predicting a label for this sample, it was able to correctly predict association to both classes with different degrees of belonging, as shown in Fig. 3a. For each test sample, the PRBF model provides a degree of possibility to belong to each class. The possibility values can be used to gain more insight into the prediction process of the model and provides the decision-maker with more information about the potentially over-lapping classes. Figure 3b shows the misclassification rates for than SVM-L, EBDT, ANN:BR, and PBRF. PRBF has the least misclassification rates. As per the majority vote, SVM-L seems to have high misclassification followed by ANN:BR and then the EBDT. Representative confusion matrices have been shown in Fig. 4.

Prediction for an adequate patient data size and predicted patient data size with the number of experts

An adequate testing patient data size of 75 was found beyond which the Mean Square Error and the validation accuracy were both stabilized for ANN:BR. This therefore establishes the minimum patient data set needed to conduct predictive patient classification. The present patient data size of 100 and five scoring experts produced accuracies of 0.93. The patient data size needed to obtain an improved accuracy of 0.99 was predicted to be 147 with the predicted number of 7 experts. Similarly, for an accuracy of 0.999, the predicted size of the number of patient data was 154 with 9 scoring experts. From the model, R^2 was 0.96, with the Total Sum or Squares (SS_{total}) as 0.04 with $p \leq 0.05$ for a $\pm 95\%$ confidence interval (C.I). Increasing the number of scoring experts from 5 to 7 can yield an accuracy of 99% but necessitates an increase in patient data set size from 100 to 147 ($R^2 = 0.96$). Likewise, increasing the number of scoring experts from 5 to 9 can yield an accuracy of 99.9% but necessitates an increase in patient data set size from 100 to 154 ($R^2 = 0.89$). There is less certainty in the prediction in going from 99 to 99.9% because of the limitations of the resent data set.

Limitations of the current approach and improvements to the existing model-based on a substantial number of experts

The HISS score is not intended to replace existing approaches but rather augment present decision making. This is a preliminary evaluation of the multiple approaches for the fusion of discrete patient sensor data into an actionable HISS score. The interactions among variables for metabolic biomarkers across dataset features and the effects would be captured using machine learning algorithms. Hence, the model-based predictions along with the evaluations of the experts' opinions form a baseline and serve as a precursor to a larger study for which the following improvement strategies can be implemented:

Number of experts

The current study uses five experts, D1-D5, with 100 SFRP data sets. The robustness of the probability theory and capacity to ascertain and account for physician variance was tested by means of uncertainty in the experts' opinions. From the results, it is observed that the self-consistency in the scoring of 4 experts can overcome the scoring inconsistency of 1 expert. Hence, a ratio of 4:1 is suggested for the number of experts. This aids in substantiating the robustness of the machine learning approaches to ascribe an accurate and actionable HISS score despite the presence of inter-physician variance.

Overlap of intermediate shock states

Due to the preliminary small number of experts, there is overlap of the intermediate states. Increasing the number of experts can help to delineate these states and improve accuracy. Future work would entail (1) improving the predictive accuracy of HISS by growing the expert data set and (2) engaging in a small scale preliminary clinical trial to assess feasibility. There is tremendous opportunity to build a longitudinal assessment data set, particularly as it relates to data on extreme resuscitations and long-term patient outcomes.

The confidence level of expert scores

It is believed that experts assign the patients to a particular class with a certain confidence level, in this case, 100%. However, they can be requested to reveal their confidence level in scoring each patient. Alternatively, the statistical confidence can be extracted by capturing the variability in the responses of the physicians using approaches like ANOVA. This could be implemented for a substantial number of experts (e.g. 100).

The relative weights of each patient attribute

In arriving at the class assignment, the expert physician reviews the five relevant physiological attributes. In its implementation, the classifier algorithms accept a single score with the assumption that each attribute is equally weighted in that decision-making assignment by the expert. In reality, experts inherently weigh each attribute and the weight is often influenced by that value and the values of other attributes. Based on their experience, the expert physician may treat certain biomarker attributes as being more or less important/influential than others when assigning the patient to the selected class. This can be extended for a substantial number of experts (e.g. 100), where a methodology can be developed to extract the relative weighting of each attribute. This relative weighting is thus a global factor assigned to the attribute. From the multiple expert physician responses obtained, a statistical assessment of the significance of each attribute can be determined. Techniques such as "leave one out analysis" and ANOVA will allow the extraction of the relative sensitivity of each attribute to the class assignment.

Temporal variation in HISS scores

In the present implementation, SFRP data were presented to each classifier algorithm as STAT data. However, patients are known to display temporal changes or trends in these biomarker values during hemorrhage progression such as during evacuation from theatre to the Green Zone. There is increasing attention being given to the diagnostic relevance of trend data in patient stratification. Temporal variations in the data to reveal

physiological trends can be explored using algorithms like recurrent neural network; however, this is outside the scope of the present study.

Clinical significance and integration of HISS

HISS is proposed for use as an adjunct device in conjuction with the state-of-the-art emergency protocols. It would help trauma surgeons and emergency physicians in mass triage and point of care stabilization of traumatic hemorrhage patients. Presently used prediction models like the TRISS or RISC-2 have been useful, being aggregates of gross physiological trends. HISS would serve as a supplement to current clinical algorithms as these can be improved upon especially by developing approaches that are capable of measuring and indicating real-time high resolution time series data that can be used for real-time decision making. Furthermore, this study may expand upon biological interactions and the relative contributions of each clinical/physiological parameter such as to inform decision making. This could be applied as a precision health approach to the diagnosis and treatment of victims of traumatic shock.

Conclusions

In this study, the Sensible Fictitious Rationalized Patient (SFRP) synthetic data generator was introduced for hemorrhaging trauma patients wherein five biomarkers; glucose, lactate, pH, potassium, and oxygen tension, served as the basis for an actionable HISS score rendered by five experts. This score is intended to serve as an adjunct and be complementary to current measures. The focus is on metabolic biomarkers over the traditional gross physiological data. Normalization of these values may greatly assist in preventing the under and over-resuscitation of victims. Several classification algorithms; linear support vector machine (SVM-L), ensemble bagged decision tree (EBDT), artificial neural network with bayesian Regularization algorithm (ANN:BR) and possibility rule-based using function approximation (PRBF) were evaluated for their ability to accurately classify the 100 entries of the SFRP data set. These data-driven predictions are presented as an adjunct to help the decision-making of physicians regarding the status of the hemorrhaging patient during triage and uses a severity scale of (0 = LOW, 1 = GUARDED, 2 = ELEVATED, 3 = HIGH, 4 = SEVERE). A training data set size of 75 has been identified as adequate to achieve the best performance by minimizing the Mean Square Error. This approach has the advantage of high validation accuracies from the ensemble bagged decision trees and linear support vector machines ($93 \pm 0.04\%$ and $91 \pm 0.06\%$) with the tunability of neural networks ($92 \pm 0.07\%$), and the ability to capture the uncertainty in the responses of experts with the help of a possibility theory-based approach ($92 \pm 0.03\%$). The predictions generated using the classification methods would assist in an adjunct device in the form of a biosensor system for point-of-care monitoring of the trauma patient, especially in mass casualty situations.

Improvement strategies are discussed with an increase in the number of experts to 100 scoring the SFRP data sets. This paper has a clinical utility in terms of classification by grouping data, prediction for incoming data and regression by means of prediction of continuous data. The predicted patient data size to obtain a test accuracy of 0.99 has been identified to be 147 with a predicted number of 7 experts. Refined prediction model disclosed a predicted patient data size of 154 with a predicted number of 9 experts for a test accuracy of 0.999. Similarly, the adequacy of the patient data size has been identified to be 75 and of the number of experts has been noted as 5 to allow training and validation. Intermediate states reveal more overlap when compared to the extreme states of LOW and SEVERE. HISS may be clinically relevant as it relates to the translation of physiologic states to severity and outcomes. Future work will entail improving the predictive accuracy of HISS and delineating the intermediate shock states by growing our expert data set and engaging in a small scale preliminary clinical trial to assess feasibility.

Abbreviations

HISS: Hemorrhage Intensive Severity and Survivability score; SFRP: Sensible Fictitious Rationalized Patient data; SVM-L: Linear support vector machine; EBDT: Ensemble Bagged Decision tree; ANN:BR: Artificial Neural Network with Bayesian Regularization; PRBF: Possibility rule-based using function approximation; MODS: Multiple Organ Dysfunction Syndrome; START: Simple Triage and Rapid Treatment; SAVE: Secondary Assessment of Victim Endpoint; ICU: Intensive care unit; ISS: Injury severity score; AIS: Abbreviated Injury Scale; PSM: Physiologic Status Monitoring; DeepSOFA: Deep Sequential Organ Failure Assessment; HIPAA: Health Insurance Portability and Accountability Act; ANOVA: One-way analysis of variance; TPR: True positive rates; MSE: Mean square error; NN: Neural network; C.I: Confidence interval; SS_{total}: Sum of squares.

Acknowledgements

Thanks to Urel Perfect Djiogan for discussions on neural networks. Thanks to Shabnam Nazmi and Dr. Abdollah Homaifar of North Carolina A&T State University for the useful discussions on the PRBF approach.

Authors' contributions

Idea, resources, and supervision: AGE, and KRW; Investigation: AB, DAP, BKW, JRA, DMA, KRW; Draft preparation, review, and editing: AB, JRA, DAP and AGE. All authors read and approved the final manuscript.

Funding

This work was supported by a Texas A and M Engineering Experiment Station (TEES) (Grant Number TEES-246413) Research Professorship to Anthony Guiseppi-Elie at Texas A and M University.

Availability of data and materials

All data used in the manuscript will be provided upon request to the corresponding author.

Ethics approval and consent to participate
Not applicable.

Consent for publication
All the authors agree for the publication.

Competing interests
Prof. Guiseppi-Elie is founder, president and scientific director of ABTECH Scientific, Inc., manufacturer of microfabricated electrodes and devices used in the measurement of physiological data.

Author details
[1] Center for Bioelectronics, Biosensors and Biochips (C3B®), Department of Biomedical Engineering, Texas A&M University, College Station, TX 77843, USA. [2] Department of Process Engineering and Technology of Polymer and Carbon Materials, Wroclaw University of Science and Technology, Norwida 4/6, 50-373 Wroclaw, Poland. [3] Department of Cardiovascular Sciences, Houston Methodist Institute for Academic Medicine and Houston Methodist Research Institute, 6670 Bertner Ave, Houston, TX 77030, USA. [4] Departments of Emergency Medicine and Biomedical Engineering, Michigan Center for Integrative Research in Critical Care, University of Michigan, Ann Arbor, MI 48109, USA. [5] Department of Surgery, Division of Acute Care Surgery, University of Michigan, Ann Arbor, MI 48109, USA. [6] Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA. [7] ABTECH Scientific, Inc, Biotechnology Research Park, 800 East Leigh Street, Richmond, VA 23219, USA.

Received: 31 May 2020 Accepted: 4 September 2020
Published online: 14 September 2020

References

1. Kotanen CN, Guiseppi-Elie A. Monitoring systems and quantitative measurement of biomolecules for the management of trauma. Biomed Microdevice. 2013;15(3):561–77.
2. Williams M, Lockey A, Culshaw M. Improved trauma management with advanced trauma life support (ATLS) training. Emerg Med J. 1997;14(2):81–3.
3. Geeraedts LMG, Kaasjager HAH, van Vugt AB, Frölke JPM. Exsanguination in trauma: a review of diagnostics and treatment options. Injury. 2009;40(1):11–20.
4. Tsuei BJ, Kearney PA. Hypothermia in the trauma patient. Injury. 2004;35(1):7–15.
5. Wilcox G. Insulin and insulin resistance. Clin Biochem Rev. 2005;26(2):19–39.
6. Chima RS, Hake PW, Piraino G, Mangeshkar P, Denenberg A, Zingarelli B. Ciglitazone ameliorates lung inflammation by modulating the inhibitor kappaB protein kinase/nuclear factor-kappaB pathway after hemorrhagic shock. Crit Care Med. 2008;36(10):2849–57.
7. Weil MH, Tang W. Forty-five-year evolution of stat blood and plasma lactate measurement to guide critical care. Clin Chem. 2009;55(11):2053–4.
8. Paladino L, Sinert R, Wallace D, Anderson T, Yadav K, Zehtabchi S. The utility of base deficit and arterial lactate in differentiating major from minor injury in trauma patients with normal vital signs. Resuscitation. 2008;77(3):363–8.
9. Luchette FA, Robinson BR, Friend LA, McCarter F, Frame SB, James JH. Adrenergic antagonists reduce lactic acidosis in response to hemorrhagic shock. J Trauma Acute Care Surg. 1999;46(5):873–80.
10. Rocha Filho JA, Nani RS, D'Albuquerque LAC, Malbouisson LMS, Carmona MJC, Rocha-e-Silva M, et al. Potassium in hemorrhagic shock: a potential marker of tissue hypoxia. J Trauma Acute Care Surg. 2010;68(6):1335–41.
11. Aboudara MC, Hurst FP, Abbott KC, Perkins RM. Hyperkalemia after packed red blood cell transfusion in trauma patients. J Trauma Acute Care Surg. 2008;64(2):S86–91.
12. Keel M, Trentz O. Pathophysiology of polytrauma. Injury $V. 2005;36(6):691–709.
13. Sasser SM, Hunt RC, Faul M, Sugerman D, Pearson WS, Dulski T, et al. Guidelines for field triage of injured patients: recommendations of the National Expert Panel on Field Triage, 2011. Morbidity Mortality Weekly Rep Recommendations Rep. 2012;61(1):1–20.
14. Cocchi MN, Kimlin E, Walsh M, Donnino MW. Identification and resuscitation of the trauma patient in shock. Emerg Med Clin North Am 2007;25(3):623–42, vii.
15. Vandromme MJ, Griffin RL, Weinberg JA, Rue LW, 3rd, Kerby JD. Lactate is a better predictor than systolic blood pressure for determining blood requirement and mortality: could prehospital measures improve trauma triage? J Am Coll Surg 2010;210(5):861–7, 7–9.
16. Lin G, Becker A, Lynn M. Do pre-hospital trauma alert criteria predict the severity of injury and a need for an emergent surgical intervention? Injury. 2012;43(9):1381–5.
17. Brasel KJ, Guse C, Gentilello LM, Nirula R. Heart rate: is it truly a vital sign? J Trauma Acute Care Surg. 2007;62(4):812–7.
18. Teasdale G, Jennett B. Assessment of coma and impaired consciousness: a practical scale. Lancet. 1974;304(7872):81–4.
19. Teasdale G, Jennett B. Assessment and prognosis of coma after head injury. Acta Neurochir. 1976;34(1–4):45–55.
20. Benson M, Koenig KL, Schultz CH. Disaster triage: START, then SAVE–a new method of dynamic triage for victims of a catastrophic earthquake. Prehosp Disaster Med. 1996;11(2):117–24.
21. Garner A, Lee A, Harrison K, Schultz CH. Comparative analysis of multiple-casualty incident triage algorithms. Ann Emerg Med. 2001;38(5):541–8.
22. Jenkins JL, McCarthy ML, Sauer LM, Green GB, Stuart S, Thomas TL, et al. Mass-casualty triage: time for an evidence-based approach. Prehosp Disaster Med. 2008;23(1):3–8.
23. Sacco WJ, Navin DM, Fiedler KE, Waddell RK 2nd, Long WB, Buckman RF Jr. Precise formulation and evidence-based application of resource-constrained triage. Acad Emerg Med. 2005;12(8):759–70.
24. Baker SP, O'Neill B, Haddon W Jr, Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. J Trauma. 1974;14(3):187–96.
25. Latifi R, Ziemba M, Leppäniemi A, et al. Trauma system evaluation in developing countries: applicability of American College of Surgeons/Committee on Trauma (ACS/COT) basic criteria. World J Surg. 2014;38:1898–904. https://doi.org/10.1007/s00268-014-2538-7.
26. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3:160035. https://doi.org/10.1038/sdata.2016.35.
27. Marshall JC, Cook DJ, Christou NV, Bernard GR, Sprung CL, Sibbald WJ. Multiple organ dysfunction score: a reliable descriptor of a complex clinical outcome. Crit Care Med. 1995;23(10):1638–52.
28. King RC, Villeneuve E, White RJ, Sherratt RS, Holderbaum W, Harwin WS. Application of data fusion techniques and technologies for wearable health monitoring. Med Eng Phys. 2017;42:1–12.
29. Guiseppi-Elie A. An implantable biochip to influence patient outcomes following trauma-induced hemorrhage. Anal Bioanal Chem. 2011;399(1):403–19.
30. Guiseppi-Elie A. Implantable biochip for managing trauma–induced hemorrhage, Patent No. US2012 0088997 A1, Apr. 12, 2012.
31. Lymberis A. Advanced wearable sensors and systems enabling personal applications. In: Lay-Ekuakille A, Mukhopadhyay SC, editors. Wearable and autonomous biomedical devices and systems for smart environment: issues and characterization. Berlin: Springer; 2010. p. 237–57.
32. Bal M, Amasyali MF, Sever H, Kose G, Demirhan A. Performance evaluation of the machine learning algorithms used in inference mechanism of a medical decision support system. Sci World J. 2014;2014:15.
33. Schiller Alicia M, Howard Jeffrey T, Lye Kristen R, Magby Christian G, Convertino Victor A. Comparisons of traditional metabolic markers and compensatory reserve as early predictors of tolerance to central hypovolemia in humans. Shock. 2018;50(1):71–7. https://doi.org/10.1097/SHK.0000000000001034.
34. Paradiso R, Loriga G, Taccini N. A wearable health care system based on knitted integrated sensors. IEEE Trans Inf Technol Biomed. 2005;9(3):337–44.
35. Majumder S, Mondal T, Deen MJ. Wearable sensors for remote health monitoring. Sensors. 2017;17(1):130.
36. Salinas J, Nguyen R, Darrah MI, Kramer GA, Serio-Melvin ML, Mann EA, et al. Advanced monitoring and decision support for battlefield critical care environment. US Army Med Dept J; 2011.

37. Gerst KS, Somberg BL, Jain BK, Canady LD. System and method for providing automatic setup of a remote patient care environment, Patent No. US 9773060 B2, Sep. 26; 2017.
38. Chern C-C, Chen Y-J, Hsiao B. Decision tree-based classifier in providing telehealth service. BMC Med Inform Decis Mak. 2019;19(1):104.
39. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
40. Bond RR, Novotny T, Andrsova I, Koc L, Sisakova M, Finlay D, et al. Automation bias in medicine: the influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. J Electrocardiol. 2018;51(6):S6–11.
41. Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. Ann R Coll Surg Engl. 2004;86(5):334–8.
42. Walker PB, Mehalick ML, Glueck AC, Tschiffely AE, Cunningham CA, Norris JN, et al. A decision tree framework for understanding blast-induced mild Traumatic Brain Injury in a military medical database. J Def Model Simul. 2017;14(4):389–98.
43. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21(3):660–74.
44. Zhu D. A hybrid approach for efficient ensembles. Decis Support Syst. 2010;48(3):480–7.
45. Krooshof PW, Üstün BL, Postma GJ, Buydens LM. Visualization and recovery of the (bio) chemical interesting variables in data analysis with support vector machine classification. Anal Chem. 2010;82(16):7000–7.
46. Moguerza JM, Muñoz A. Support vector machines with applications. Stat Sci. 2006;21(3):322–36.
47. Steinwart I, Christmann A. Support vector machines. Berlin: Springer Science & Business Media; 2008. p. 287.
48. Trafalis TB, Gilbert RC. Robust support vector machines for classification and computational issues. Opt Methods Softw. 2007;22(1):187–98.
49. Amato F, López A, Peña-Méndez EM, Vaňhara P, Hampl A, Havel J. Artificial neural networks in medical diagnosis. J Appl Biomed. 2013;11(2):47–58.
50. Masumoto H, Tabuchi H, Nakakura S, Ohsugi H, Enno H, Ishitobi N, et al. Accuracy of a deep convolutional neural network in detection of retinitis pigmentosa on ultrawide-field images. PeerJ. 2019;7:e6900.
51. Maulik U, Bandyopadhyay S. Genetic algorithm-based clustering technique. Pattern Recogn. 2000;33(9):1455–65.
52. Kordmahalleh MM, Sefidmazgi MG, Homaifar A, KC DB, Guiseppi-Elie A. Time-series forecasting with evolvable partially connected artificial neural network. In: Proceedings of the Companion Publication of the 2014 Annual Conference on Genetic and Evolutionary Computation. Vancouver, BC, Canada: ACM; 2014. p. 79–80.
53. Larvie JE, Sefidmazgi MG, Homaifar A, Harrison SH, Karimoddini A, Guiseppi-Elie A. Stable gene regulatory network modeling from steady-state data. Bioengineering. 2016;3(2):12.
54. Davoudi A, Malhotra KR, Shickel B, Siegel S, Williams S, Ruppert M, et al. Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning. Sci Rep. 2019;9(1):8020.
55. Armen SB, Freer CV, Showalter JW, Crook T, Whitener CJ, West C, et al. Improving outcomes in patients with sepsis. Am J Med Qual. 2016;31(1):56–63.
56. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716–20.
57. Zadeh LA. Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst. 1978;1(1):3–28.
58. Dubois D, Prade H. Possibility theory and its applications: where do we stand? Springer handbook of computational intelligence. Berlin: Springer; 2015. p. 31–60.
59. Nazmi S, Ramyar S, Homaifar A. Determination of the driver at-fault using possibility theory-based classification. In: 2019 Transportation Research Board Annual Meeting (TRB). TRB; 2019.
60. Denœux T, Zouhal LM. Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets Syst. 2001;122(3):409–24.
61. Burnell JM, Scribner BH, Uyeno BT, Villamil MF. The effect in humans of extracellular pH change on the relationship between serum potassium concentration and intracellular potassium. J Clin Invest. 1956;35(9):935–9.
62. Valizadegan H, Nguyen Q, Hauskrecht M. Learning classification models from multiple experts. J Biomed Inform. 2013;46(6):1125–35.
63. Castanedo F. A review of data fusion techniques. Sci World J. 2013;2013:19.
64. Newby D, Freitas AA, Ghafourian T. Comparing multilabel classification methods for provisional biopharmaceutics class prediction. Mol Pharm. 2015;12(1):87–102.
65. McLauchlan L, Mehrübeoğlu M. Neural network-based watermark embedding and identification. SPIE; 2008.
66. Pomares A, Martínez JL, Mandow A, Martínez MA, Morán M, Morales J. Ground extraction from 3D lidar point clouds with the classification learner App. In: 2018 26th Mediterranean Conference on Control and Automation (MED). 2018. p. 1–9.
67. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
68. Sexton RS, Gupta JND. Comparative evaluation of genetic algorithm and backpropagation for training neural networks. Inf Sci. 2000;129(1):45–59.
69. Rangelova V, Tsankova D, Dimcheva N. Soft computing techniques in modelling the influence of ph and temperature on dopamine biosensor. Intelligent and biosensors. Rijeka: InTech; 2010. p. 101.
70. Burden F, Winkler D. Bayesian regularization of neural networks. In: Livingstone DJ, editor. Artificial neural networks: methods and applications. Totowa: Humana Press; 2009. p. 23–42.
71. Nazmi S, Homaifar A. Possibility rule-based classification using function approximation. SMC; 2018.
72. Iaccarino G, Petrone G, Witteveen J, Quagliarella D, Nicola CD, Axerio-Cilies J. Wind turbine optimization under uncertainty with high performance computing. In: 29th AIAA applied aerodynamics conference; 2011.
73. Petrone G, Axerio-Cilies J, Quagliarella D, Iaccarino G. A probabilistic non-dominated sorting GA for optimization under uncertainty. Eng Comput. 2013;30:1054–85.
74. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.
75. Muh HC, Tong JC, Tammi MT. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. PLoS ONE. 2009;4(6):e5861.
76. Nasrabadi NM. Pattern recognition and machine learning. J Electron Imaging. 2007;16(4):049901.
77. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Stat Surveys. 2010;4:40–79.
78. Tibshirani R, Walther G. Cluster validation by prediction strength. J Comput Graph Stat. 2005;14(3):511–28.
79. Thornton AR, Raffin MJM. Speech-discrimination scores modeled as a binomial variable. J Speech Hear Res. 1978;21(3):507–18.
80. Rehme AK, Volz LJ, Feis D-L, Bomilcar-Focke I, Liebig T, Eickhoff SB, et al. Identifying neuroimaging markers of motor disability in acute stroke by machine learning techniques. Cereb Cortex. 2014;25(9):3046–56.
81. Bashir S, Qamar U, Khan FH. IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. J Biomed Inform. 2016;59:185–200.
82. Forcellese A, Gabrielli F, Ruffini R. Effect of the training set size on spring-back control by neural network in an air bending process. J Mater Process Technol. 1998;80–81:493–500.
83. Twomey JM, Smith AE. Bias and variance of validation methods for function approximation neural networks under conditions of sparse data. IEEE Trans Syst Man Cybern Part C Appl Rev. 1998;28(3):417–30.
84. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PLoS ONE. 2014;9(2):e88225.
85. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: a scoping review. PLoS ONE. 2019;14(2):e0212356.
86. Al-Absi HRH, Abdullah A, Hassan MI, Bashir Shaban K. Hybrid intelligent system for disease diagnosis based on artificial neural networks, fuzzy logic, and genetic algorithms. In: Abd Manaf A, Zeki A, Zamani M, Chuprat S, El-Qawasmeh E, editors. Informatics engineering and information science. Berlin: Springer; 2011. p. 128–39.
87. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. Sci Rep. 2017;7(1):17816.
88. Moshtagh-Khorasani M, Akbarzadeh-T M-R, Jahangiri N, Khoobdel M. An intelligent system based on fuzzy probabilities for medical diagnosis- a study in aphasia diagnosis. J Res Med Sci. 2009;14(2):89–103.

89. Kumar RN, Kumar DMA. Enhanced Fuzzy K-NN approach for handling missing values in medical data mining. Indian J Sci Technol. 2016;9:1–7.
90. Chan V, Pole JD, Mann RE, Colantonio A. A population based perspective on children and youth with brain tumours. BMC Cancer. 2015;15:1007.

Publisher's Note