Applied Soft Computing Journal 80 (2019) 525–533

Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data

Gokul S. Krishnan *, Sowmya Kamath S.

Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangaluru 575025, India

HIGHLIGHTS

- Novel GA-ELM model to predict mortality risk of ICU patients using relevant lab events.
- Genetic Algorithm wrapper for deriving relevant lab events for mortality prediction.
- Extreme Learning Machine model trained for predicting ICU patients' mortality risk.
- Patient and lab events data from MIMIC-III dataset were used for the study.
- GA-ELM model outperforms severity scoring systems and machine learning based models.

ARTICLE INFO

Article history:
Received 22 July 2018
Received in revised form 6 April 2019
Accepted 13 April 2019
Available online 29 April 2019

Keywords:
Clinical decision support systems
Healthcare analytics
Machine learning

ABSTRACT

Patient-specific mortality prediction models are an essential component of Clinical Decision Support Systems developed for caregivers in Intensive Care Units (ICUs), that enable timely decisions towards effective patient care and optimized ICU resource management. While high prediction accuracy is a fundamental requirement for any mortality prediction application, being able to so with minimal patient-specific data is a major plus point that can help in improving care delivery and cost optimization. Most existing scoring techniques and prediction models utilize a multitude of lab tests and patient events to predict mortality and also suffer from reduced performance when available patient data is less. In this paper, a Genetic Algorithm based Wrapper Feature Selection technique is proposed for determining most-optimal lab events that contribute predominantly to mortality, even for large-scale patient cohorts. Using this, an Extreme Learning Machine (ELM) based neural network is designed for predicting patient-specific ICU mortality. The proposed GA-ELM model was benchmarked against four popular traditional mortality scores and also state-of-the-art machine learning models for experimental validation. The GA-ELM model achieved promising results as it outperformed the traditional scoring systems by 11%–29% and state-of-the-art models by up to 14%, in terms of AUROC.

1. Introduction

Over the past decade, healthcare management systems have evolved to provide active decision-making capabilities through Clinical Decision Support Systems (CDSSs). Hospital systems continuously generate huge volumes of clinical data which when effectively analyzed for diagnosis support, can improve productivity and clinical care delivery. Typically, patient data is available from varied sources and is highly temporal. CDSSs help utilize this heterogeneous data for supporting intelligent applications like patient profiling and disease prediction. In critical care applications, the process of taking practical decisions on managing the care of intensive care patients can thus help augment doctors', by incorporating predictive data analysis on the large amounts of data generated while monitoring these patients.

The most important aspect of a CDSS in the ICU is, undoubtedly, its ability to accurately predicting in advance the mortality risk of a patient, so that doctors and other healthcare personnel can be prepared to intervene in time, with the resources available in ICU. Apart from measuring the severity of illness, mortality prediction can also play a crucial role in the assessment of treatment and critical care policies of a hospital. Hence, ICU mortality prediction has remained a well-researched problem over the years, a fact that is evidenced in the various severity scores developed for the purpose and also the customized, country-specific variants that are currently in use. In recent years, the application of data mining and machine learning techniques have enabled further enhancement in CDSS applications like ICU mortality prediction, length-of-stay prediction, etc. Despite this, validation studies carried out by various researchers [1,2] have shown that these scores

* Corresponding author.
E-mail addresses: gsk1692@gmail.com (Gokul S. Krishnan), sowmyakamath@nitk.edu.in (Sowmya Kamath S.).

https://doi.org/10.1016/j.asoc.2019.04.019
1568-4946/

can be further fine-tuned for better performance. Any such fine-tuning can help in reducing the time taken in collecting patient data, thus enabling earlier predictions with better accuracy than that achieved by traditional scoring systems.

Parametric scoring based prediction models typically use the perceived relevance of the clinical measurements of an ICU patient, to calculate a score in a particular range, as per a model derived by clinical experts. A popular severity scoring model called APACHE[1] [3], along with its later variants [4–6], is a physiological system that computes an ICU patient's mortality score using their clinical data. The SAPS[2] [7] scoring system and its subsequent versions [8,9] are also popular parametric scoring systems that use measured biological and clinical variables to predict the possibility of patient death in ICUs. Another scoring based model, SOFA[3] [10] tracks the extent of organ failures in ICU patients and predicts mortality risk based on only six variables (Respiration, Central Nervous System, Cardiovascular, Renal, Coagulation and Comorbidity). OASIS,[4] a recent scoring system proposed by Johnson et al. [11], uses a subset of APACHE-IV variables along with others like age, length-of-stay and elective surgery prior to ICU admission, to predict mortality of ICU patients. According to the authors, its performance is at par with that of APACHE-IV and is considered superior to APACHE-IV as it requires lesser features.

Non-parametric models incorporate soft computing techniques such as data mining, Machine Learning (ML), bio-inspired/evolutionary computing and other optimization techniques. These models have been applied across domains and have proven to be comparatively effective in solving various problems. In the field of bioinformatics, techniques like Genetic Algorithm [12,13] and memetic optimization [14] have been used to address the problem of protein multiple sequence alignment. Hybrid and modified versions of techniques like Particle Swarm Optimization, Cat Swarm Optimization [15] and Gravitational Search [16] have been applied successfully adapted for load balancing and scheduling in cloud computing environments.

In the field of Healthcare, non-parametric techniques based prediction models employ data mining and Machine Learning (ML) techniques for predicting the risk of death in ICUs. Works proposed by various authors [2,17–27] use various ML and data mining techniques such as Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees, Logistic Regression and also Deep Learning techniques like RNN for predicting mortality risk. These works have been compared with traditional severity scores, mostly SAPS, SOFA and APACHE, and were found to outperform them with respect to accuracy.

While most parametric scoring systems like APACHE-II, SAPS-II and SOFA are now considered standard for ICU mortality measurement in practice, the accuracy achieved by them is low in comparison to non-parametric methods. Moreover, the patient-specific data points considered as features by each scoring system is different and often, significantly large in number, which means that all such prescribed lab tests (lab events) have to be performed for each ICU patient before a mortality risk can be assessed. This contributes to an additional delay in making time-critical mortality decisions, while also adversely affecting cost and resource usage. It is therefore important to be able to predict mortality risk using as minimal clinical variables or lab events (features) as possible, at the earliest possible time, with high precision and accuracy. To the best of our knowledge, such an investigation focusing on the contribution of individual or group of labevent related features in mortality risk prediction has not been conducted on large-scale patient data, as a benchmark study. This paper aims to address these gaps, and our significant contributions are listed below.

1. Designing a Genetic Algorithm based Wrapper Feature Selection technique with Extreme Learning Machine as estimator (GAWFS) for effectively capturing the representative lab events for mortality prediction, and benchmarking its effectiveness against that of traditional feature selection techniques — ANOVA F-test, Mutual Information test, Recursive Feature Elimination & Sequential Feature Selection.
2. Designing an Extreme Learning Machine (ELM) based neural network architecture for building an ICU mortality prediction model trained on optimal representations of patients' clinical data for improved prediction performance.
3. Benchmarking the proposed mortality prediction model built on GAWFS and ELM against four popular traditional ICU severity scoring models — SAPS-II, SOFA, APS-III & OASIS, and state-of-the-art machine learning based models.

The rest of this paper is organized as follows: In Section 2, the proposed feature selection technique and mortality prediction model are described in detail. The experimental validation results and discussion of benchmarking experiments conducted for comparison with traditional scoring models and state-of-the-art machine learning approaches is presented in Section 3, followed by conclusions and references.

2. Materials and methods

The processes defined as part of the proposed approach for ICU mortality prediction based on a patient cohort's clinical data is depicted in Fig. 1. Each of these processes is discussed in further detail in this section.

2.1. Patient cohort selection and data preprocessing

For benchmarking the proposed methodology, we used an openly available standard dataset, MIMIC-III (v1.4) [28] for our experiments. The dataset consists of deidentified hospital data associated with multiple admissions of 46,520 distinct patients. From this data, a patient cohort was selected based on the following criteria:

1. Clinical data of only adult patients (age>15) was selected for the cohort, in accordance with previous studies. This is important as the procedures used for pediatric patients are highly specific in nature [2].
2. In cases where a patient was admitted to ICU multiple times, only the first ICU admission of each patient was considered for the study. This helps ensure the CDSS nature of a mortality prediction model which helps in predicting mortality risk with respect to earliest available data on a patient's condition.

Accordingly, a subset of 31,691 eligible patients was chosen as the patient cohort. For these patients, the results of a total of 573 lab tests performed are available in a MIMIC-III table called 'labevents'. These lab test values are extracted and modeled into a representation, where each row represents a patient, and each column represents a lab test. However, there are several missing values in some rows, as not all tests are necessarily performed on all patients. If the rows (patients) with such missing column values are directly removed from the cohort, then a large number of patients will need to be excluded from the cohort. To overcome this, we separately calculated the median values of each

[1] Acute Physiology And Chronic Health Evaluation (APACHE).
[2] Simplified Acute Physiological Score (SAPS).
[3] Sequential Organ Failure Assessment (SOFA).
[4] Oxford Acute Severity of Illness Score (OASIS).

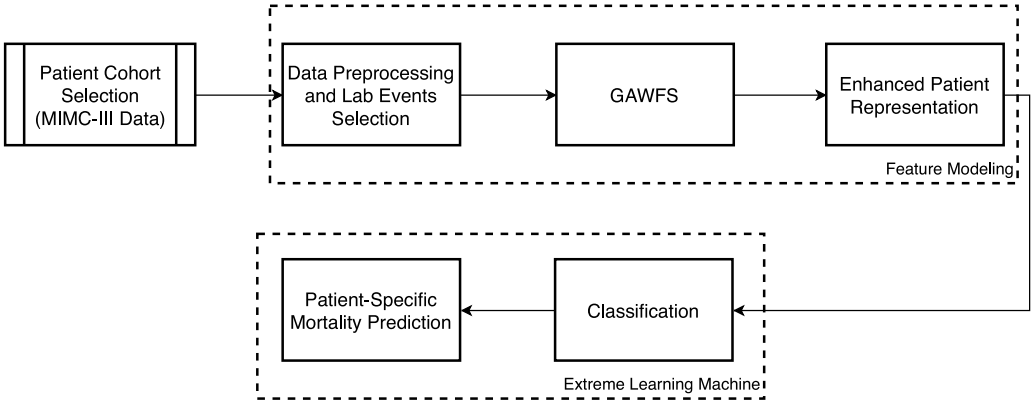

Fig. 1. Workflow of the defined methodology.

Table 1
Number of expired/alive patients in initial and selected cohorts.

Cohort	Alive	Expired	Total
MIMIC-III data	30,761	15,759	46,520
Selected Cohort	19,225	12,466	31,691

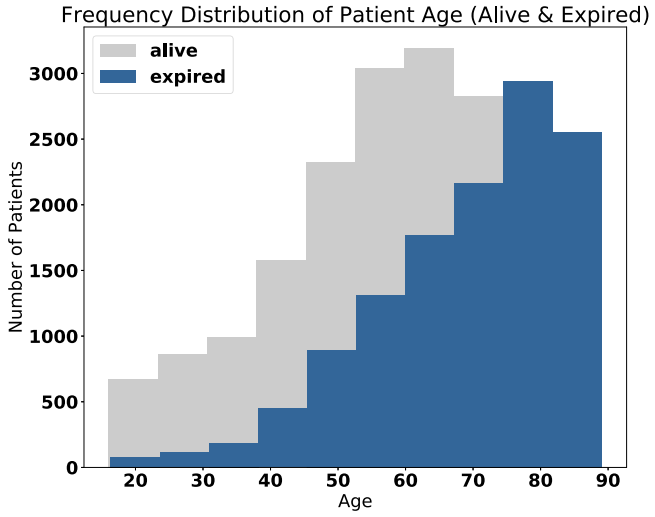

Fig. 2. Statistics relating to the age of patients in the selected cohort.

column for all alive and expired patients in the selected cohort and filled these median values in place of any missing values in that particular column. Along with these 573 features, other demographic features like *age* and *gender* were also added to the feature set. Additionally, the *ICD9* [29] disease code of a patient's first diagnosis, *length of stay* and also the *first_careunit* (the type of ICU to which the patient was first admitted to) of the patients were also considered as features. After final preprocessing tasks are applied, the patient cohort consisted of 31,691 patients (rows) and the 578 features (columns) representing them. The outcome labels are the 'expire_flag' of each patient (0 for alive and 1 for expired).

The statistics of the selected cohort is tabulated in Table 1 and Fig. 2 shows the frequency distribution of patient age (expired and alive) in the selected cohort.

2.2. Optimally modeling lab events

Laboratory tests or events are one of the several pre- and post-analytic mechanisms that help medical personnel in continuously monitoring a patient's condition. Often, medical personnel is prone to order various lab evaluation procedures for patients, some of which may be unnecessary in actual understanding of the patient's condition. Eliminating such unnecessary and wasteful lab evaluations are of significant importance, given the rapidly escalating healthcare and insurance costs as well as excessive overuse of laboratories and equipment in care delivery [30]. Moreover, the additional time spent performing such unnecessary tests required by some scoring systems can delay crucial new insights into the patient's condition, thus affecting timely intervention. For the problem of mortality prediction for ICU patients, it is critical to predict mortality risk at the earliest possible patient condition and hence, reducing the number of lab events required to predicting mortality effectively, is a matter of significant importance. In this paper, we attempt to model patient-specific lab event requirements for the two-fold objective of reducing prediction time as well as improving prediction accuracy.

To determine the optimal representation for each patient in the chosen cohort, a Genetic Algorithm based Wrapper Feature Selection (GAWFS) technique is proposed. GAWFS is used to find the most-optimal subset of feature variables of a patient (i.e., lab events) to predict mortality risk of ICU patients. This optimal feature set is used for training a learning-based risk prediction model. While feature selection techniques have been extensively used for deriving the optimal feature set, feature extraction can also be used to reduce dimensionality and increase the efficacy of the model. However, feature extraction techniques use a statistical combination of feature values to generate new features which makes it impossible to track which features (in our case lab events) were contributed in the prediction. As the purpose of our work is also to identify the most crucial lab events, we used feature selection instead of feature extraction.

2.2.1. Feature selection

Feature selection (FS) is the "process of selecting an optimal subset of relevant features for use in the construction of prediction models" [31]. Essentially, FS methods can help in reducing the dimensionality of the dataset by ignoring the unimportant or noisy features so that the prediction process can be more accurate and computationally efficient [32]. In this case, if a real-world

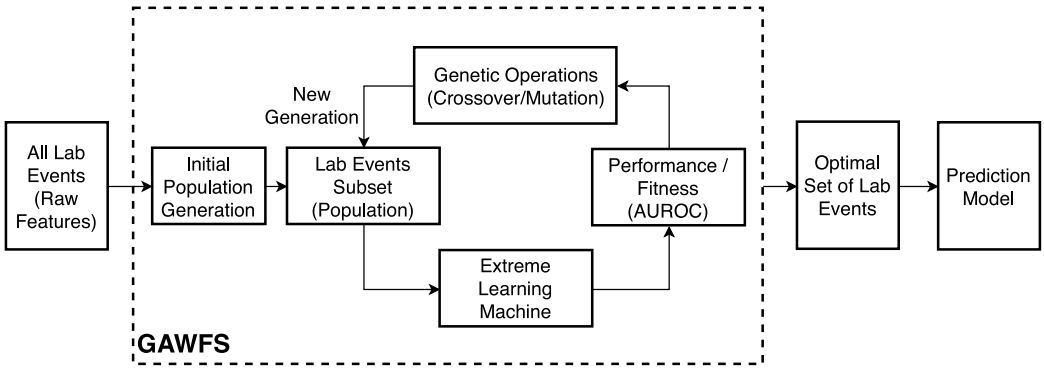

Fig. 3. Genetic Algorithm based Wrapper Feature Selection process.

Algorithm 1 Optimal Lab Events Subset Selection using GAWFS

Input: Set of all lab events & patient-specific mortality labels
Output: Optimal lab events subset of, say, n features *(Best solution)*

while iterations $\leq$ 100 **do** ▷ *No.of generations=100*
  Generate randomly a feature set (lab events) for all patients ▷ *Each feature subset represents an individual chromosome, and 100 feature subsets of patients represent the initial population*
  Select parents and perform genetic operations ▷ *Single point cross-over and mutation with probabilities of 0.5 and 0.2 respectively are used*
  Create new generation
  Calculate fitness of new generation ▷ *AUROC performance of ELM*
  if new-fitness > old-fitness **then** ▷ *New generation's fitness is better than that achieved with previous subset of features*
    Replace current generation with new generation
  else
    Retain the current generation
  end if
end while

CDSS application can make accurate predictions based on a lower number of features (e.g., lab event measurements), then it can potentially save lives, time and cost, and consequently, is more effective and valuable.

FS techniques can be mainly sub-categorized into *filter* and *wrapper* methods. Filter methods are suitable for quick feature selection based on the threshold of general characteristics of the data, such as statistical dependencies, without using any induction or classification algorithm [32,33]. Some popular examples are Analysis of Variance (ANOVA) F-test [34] and Mutual Information (MI) test. Wrapper methods generate an optimal feature subset by evaluating the quality of each feature subset, based on some classification or induction algorithm, regardless of the chosen learning method [35]. Recursive Feature Elimination (RFE) and Sequential Feature Selection (SFS) are popular examples of wrapper based methods. Although wrapper methods are computationally more expensive in comparison to filter methods, the quality of the derived feature subset is better ensured as the performance evaluation with respect to a classifier model is involved in the feature selection process.

In order to determine the reduced optimal set of features, i.e., the reduced set of lab events contributing the most towards mortality risk prediction, we propose a Genetic Algorithm based Wrapper Feature Selection (GAWFS) technique. Genetic Algorithm (GA) is an evolutionary meta-heuristic algorithm inspired by the biological process of natural selection and the theme of "survival of the fittest". GA is known to offer high-quality solutions to optimization and search problems by using the operations — Selection, Crossover and Mutation as in the process of natural selection and hence, GA is appropriate for the feature selection process for removing redundant lab events for improved mortality prediction.

The GAWFS process is depicted in Fig. 3. We made use of concepts of GA for calculation of fitness of a population (a set of individuals and chromosomes, i.e., a subset of features or lab events) and based on the fitness, a particular feature is selected if it is fit. From the full feature set consisting of 578 features, the initial population was selected as 100 random feature subsets of lab events for all patients (each feature subset is analogous to 'individuals' in the 'population'). The feature subset and the associated patient-specific mortality labels are fed into an estimator/classifier, whose fitness in terms of classification performance is then measured.

Algorithm 1 illustrates the process of deriving the optimized lab event subset using GAWFS. We use an Extreme Learning Machine (ELM) based neural network based architecture as a classifier or estimator model for the GA technique, thereby making GAWFS a wrapper based feature selection technique. ELM is a training method for a single hidden layer neural network based classifier, for which only the weights between hidden and output layer need to be learned. The ELM model is described in detail in Section 2.3. As the fitness function, the metric *Area Under the Receiver Operating Characteristic Curve (AUROC)*, as shown in Eq. (1), was used in GAWFS for calculating the fitness value associated with a particular feature subset. AUROC measures the overall quality of a classifier by varying the threshold parameter (say i), which biases the classes and returns a value between 0 and 1 (where a value of 1 indicates best classification performance possible). The number of thresholds varied is determined by the number of unique number of predicted probabilities of the ELM classifier. As AUROC measures how well a classifier has learned to classify between the majority and minority classes in the presence of class imbalance, it is apt for our problem of mortality prediction, and therefore, it was chosen as the fitness function in GAWFS and is calculated as per Eq. (1).

$$Fitness, f(x) = \sum_{1}^{N-1} (TPR_{i+1} - TPR_i)(FPR_{i+1} - FPR_i) \quad (1)$$

where FPR is the False Positive Rate, TPR is the True Positive Rate and i refers to the varying threshold parameter for which at each point FPR and TPR are determined and N is the number of thresholds which was found to be 1062 in our experiment. Eq. (1) sums all the area of all the small rectangles in the ROC curve between two FPR and TPR points for adjacent thresholds.

Single point crossover and mutation operations were performed with empirically determined probabilities of 0.5 and 0.2 respectively. During this iteration, a new generation gets generated where, least fit individuals in the population are replaced, provided their fitness is better compared to the ones in the population. In order to enable us to compare the proposed FS model with other FS techniques, the GAWFS technique was configured to select the most important 10 lab events or features.

2.3. Building the prediction model

The design of the proposed prediction model is driven by the two principal requirements of a mortality prediction CDSS. Firstly, eliminate false negative mortality predictions, i.e., a wrong low mortality risk prediction for a patient who is actually at high mortality risk should never occur, and, secondly, to ensure learnability after deployment as a real-world CDSS. To address these two major aspects, we propose an architecture built on an Extreme Learning Machine (ELM) neural network, with Rectified Linear Unit (ReLU) as the hidden layer activation function.

ELM is a learning technique for training Single hidden Layer Feedforward Neural Networks (SLFNN) [36] that are trained on finite training sets. Initially, the hidden nodes in ELM are randomly fired with random weights and learning is carried out without iterative tuning. By design, only one parameter needs to be learned in ELM, i.e., the set of weights between the hidden layer and output layer. Thus ELMs are extremely fast when compared to traditional SLFNNs and also very well-suited for further retraining for future learning [37]. Moreover, ELMs can be trained to converge to the smallest possible error with minimal magnitude of weights, due to which the generalization performance of ELMs far exceeds that of traditional feedforward neural networks (as per Bartlett's theory). Another advantage of ELM is its ability to reach solutions in a straightforward manner avoiding problems like local minima, overfitting and improper learning rate [37]. Based on these observations, we experimented with ELM as an estimator in the proposed mortality prediction model, for exploiting its advantages for the development of real-world CDSSs. After the FS process, the feature subset with the best fitness value along with the class labels is used for training the ELM model, for predicting patient-specific mortality risk.

Fig. 4 shows the ELM architecture used for training the proposed model. It is primarily a SLFNN architecture, where, the number of input nodes is governed by the number of features used for training (as described in Section 2.2.1). The hidden layer consists of 50 nodes and one node at the output layer, which predicts the mortality risk. We experimented with several variations in the number of nodes in the hidden layer, but it was observed that the performance improvement was minimal beyond 50 nodes, and moreover, this resulted in significant increase in training time. Thus, we chose the optimal number of nodes in the hidden layer as 50. We used ReLU as the hidden layer activation function in the proposed ELM architecture. ReLU, being a ramp function given by $f(x) = max(0, x)$, can trigger for any non-zero input and therefore, helps in predicting even the slightest chance of mortality, thereby eliminating any false negative predictions.

Algorithm 2 depicts the process of training the ELM network as an estimator for the proposed feature selection model, which also works as the final prediction model. The various feature sets generated by the GAWFS technique and the final optimal feature set obtained after the GAWFS process with the associated patient-specific mortality labels are used for training the ELM model. The parameters and weights are initialized randomly, and the output matrix is calculated based on the given lab events or features set and patient-specific mortality labels. During training, the weights between the hidden and output layer are iteratively optimized and finally, patient-specific mortality prediction performance is observed.

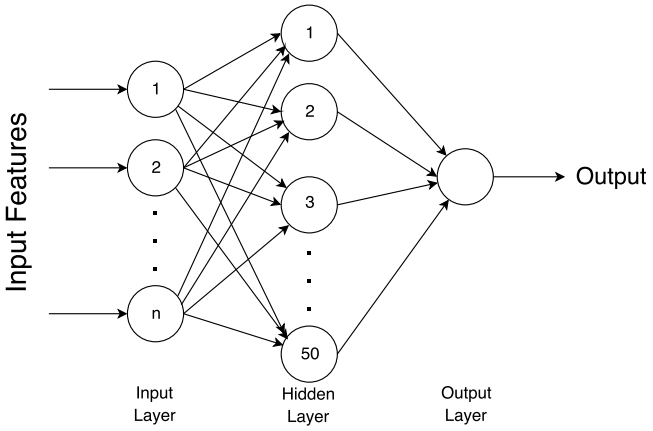

Fig. 4. Architecture of the ELM Prediction Model.

Algorithm 2 Process of training ELM as an estimator for proposed GAWFS and for the proposed prediction model

Input: A training set with N samples (from GAWFS) consisting of lab event features and mortality labels $(x_i, y_i)|x_i, y_iER, i = 1, 2, ...N$ Activation function $f(x)$ *(ReLU)* ***Output***: ICU Mortality Prediction

1: Randomly assign weights w_i and bias b_i, $i = 1, 2...N$ for N training samples
2: Compute output matrix based on the input lab event feature set, say H
3: Compute output weights based on input mortality labels and output matrix β
4: Train the ELM network using the Least Square Solution β' to the linear system $H\beta = T$
5: Analytically tune output weights
6: Perform prediction for test set and observe prediction performance

3. Experimental evaluation and results

The proposed mortality prediction approach was evaluated by a series of experiments, designed to benchmark it against both traditional and state-of-the-art learning-based approaches. The experimental evaluations were performed on an environment which consisted of a high-end server running Ubuntu Server OS with 56 cores of Intel Xeon processors, 128 GB RAM, 3 TB Hard Drive and two NVIDIA Tesla M40 GPUs. For all experiments involving training and testing, 10-fold cross-validation was performed. The proposed GAWFS technique was configured to select the 10 most important lab events that contributed towards mortality prediction from the 578 lab events (raw features). For the selected cohort of 31,691 patients, the lab events (features) that were found to be of high importance were as per proposed GAWFS technique – *Platelet Count, Red Blood Cells, Hematocrit, Sodium, Chloride, Bicarbonate, Base Excess, Urea Nitrogen, Anion Gap, Partial Thromboplastin Time (PTT)*. These features, along with the corresponding patient-specific mortality labels for the selected cohort were then used for training and validation of the designed prediction model.

3.1. Evaluation of the proposed feature selection model

This phase involved two experiments. The first experiment was designed for observing the performance of the proposed

Table 2
Comparison of ICU mortality prediction performance of the proposed GAWFS+ELM model and other traditional feature selections techniques.

Metric	*ANOVA*	*MI*	*RFE*	*SFS*	***GAWFS+ELM***
Accuracy	0.70	0.70	0.72	0.71	**0.75**
Precision	0.74	0.73	0.74	0.74	**0.81**
Recall	0.70	0.70	0.71	0.71	**0.74**
F-Score	0.71	0.71	0.72	0.72	**0.76**
AUROC	0.75	0.74	0.76	0.76	**0.80**

Table 3
Comparison of ICU mortality prediction performance comparison between ELM with no feature selection and proposed model (GAWFS+ELM).

Metric	*ELM*	***GAWFS+ELM***
Accuracy	0.70	**0.75**
Precision	0.78	**0.81**
Recall	0.71	**0.74**
F-Score	0.74	**0.76**
AUROC	0.74	**0.80**

GAWFS+ELM model, against that of conventional filter and wrapper based methods. We selected two conventional filter FS methods (ANOVA F-test and Mutual Information (MI)) and two wrapper FS techniques (Recursive Feature Elimination (RFE) and Sequential Feature Selection (SFS)) and applied them to raw features (578 in total) for deriving the respective optimal features sets. The ELM is again used as an estimator/classifier model for each FS method. Each FS technique was configured to select 10 important lab events and the respective feature sets generated by each technique were used for training a base ELM model. The performance of each of these models was compared against the proposed GAWFS+ELM model. Standard metrics like accuracy, precision, recall, F-score and AUROC were used for comparative evaluation of performance. The validated results for the ELM model trained with feature sets generated by GAWFS, ANOVA, MI, RFE and SFS feature selection techniques are shown in Table 2. Next, in the second experiment, we compared the GAWFS+ELM model to measure its performance over that of pure ELM architecture when trained with the initial feature set (578 raw features) without using any FS technique. The results of this experiment are tabulated in Table 3.

From Tables 2 and 3, it is clear that the feature selection using the proposed GAWFS technique was most effective for prediction and achieved the best performance with respect to all metrics in contrast to other FS techniques as well as over the model that used only raw features for prediction. Hence, we conclude that the reduced feature set selected by the proposed GAWFS technique is made up of the most relevant features or lab events that contribute the most significant patient-specific information, due to which a mortality prediction model trained on it can be effective in real-world scenarios too. More importantly, in the case of a real-world CDSS application, a major advantage is foreseen as only 10 features (lab tests/events) need to be measured for each patient thus eliminating wasteful or insignificant lab tests. This can result in a significant reduction in costs and unnecessary hospital resource consumption, in addition to making predictions comparatively faster, with better accuracy.

3.2. Benchmarking against traditional mortality scoring systems

Several traditional scoring methods are already in use in real-world ICUs, which are primarily parametric mortality scores. To evaluate the effectiveness of the proposed GAWFS+ELM model, we benchmarked the performance of the proposed model against that of four traditional scoring systems, SAPS-II (Simplified Acute Physiological Score), SOFA (Sequential Organ Failure Assessment), APS-III (Acute Physiological Score) and OASIS (Oxford Acute Severity of Illness Score). For each patient in the selected cohort, we implemented each traditional score using the lab event data from the MIMIC dataset [38], and the prediction results were obtained. For SAPS-II, the probability of mortality [2,8] for each patient was calculated as per Eq. (1).

$$\log(P_m/1-P_m) = -7.7631 + 0.0737 * S + 0.09971 * \log(1+S) \quad (2)$$

where P_m is the required mortality probability of a patient and S is the SAPS-II score of the patient. The threshold of classification for SAPS-II based mortality probability was taken as 0.5 as done by Patil et al. [39].

In the case of SOFA, the mortality prediction of each patient was obtained by regressing the mortality on the SOFA score using a main-term logistic regression model as per Pirracchio et al. [2]. Similarly, for APACHE-III (APS III), the mortality probability for each patient is calculated as per Eq. (2) [5], where, P_m is the required mortality probability of a patient and *APS* is the APS-III score of the patient.

$$\log(P_m/1 - P_m) = -4.4360 + 0.04726 * APS \quad (3)$$

The probability of mortality for each patient as per the OASIS scoring system is given by the in-hospital mortality score calculation [11], given by Eq. (3).

$$\log(P_m/1 - P_m) = -6.1746 + 0.1275 * OASIS \quad (4)$$

where P_m is the required mortality probability of a patient and *OASIS* is the OASIS score of the patient. The threshold of classification for APS-III and OASIS based mortality probabilities were also considered to be 0.5.

Table 4
Comparison of ICU mortality prediction performance of proposed GAWFS+ELM model with traditional severity scores — SAPS-II, SOFA, OASIS and APS-III.

Metric	***GAWFS+ELM***	*SAPS-II*	*SOFA*	*OASIS*	*APS-III*
Accuracy	**0.75**	0.65	0.63	0.62	0.62
Precision	**0.81**	0.66	0.62	0.67	0.67
Recall	**0.74**	0.65	0.63	0.62	0.61
F-Score	**0.76**	0.59	0.57	0.51	0.49
AUROC	**0.80**	0.72	0.62	0.64	0.67

The results of this experiment are summarized in Table 4. It can be observed that the proposed GAWFS+ELM model outperformed all the traditional scoring systems considered for the comparison — SAPS-II, SOFA, OASIS and APS-III, by 15%–20% in terms of accuracy, while the observed AUROC improvement was about 11%–29%. The superiority of the proposed prediction models trained on a highly relevant feature set is evident from the tabulated results in terms of all other metrics considered. A plot of ROC curves for the proposed model and also the standard scoring systems is shown in Fig. 5. It can be observed in the plot that the area under Receiver Operating Characteristic (ROC) curve for the proposed GAWFS+ELM model is significantly higher than the standard scoring models.

3.3. Comparison with state-of-the-art machine learning models

Several non-parametric approaches to ICU Mortality prediction have been proposed over the years. We conducted experiments to benchmark the performance of the proposed GAWFS+ELM model against the current state-of-the-art works in this domain, like the models proposed by Calvert et al. [23,24] and Grnarova et al. [25] in 2016, Harutyunyan et al. [26] in 2017 and Che et al. [27] (2018). These state-of-the-art ML based models were developed and benchmarked on the MIMIC-III dataset. For

Table 5
Comparison of ICU mortality prediction performance of proposed GAWFS+ELM model with state-of-the-art ML based models.

Study cohort	*Study's accuracy*	***GAWFS+ELM accuracy***	*Study's AUROC*	***GAWFS+ELM AUROC***
Calvert et al. [23]	0.80	**0.90**	0.88	**0.90**
Calvert et al. [24]	0.81	**0.92**	0.93	**0.94**
Grnarova et al. [25]	–[a]	**0.96**	0.96	**0.97**
Harutyunyan et al. [26]	–[a]	**0.92**	0.86	**0.90**
Che et al. [27]	–[a]	**0.98**	0.84	**0.96**

[a]*Note:* The authors of this study reported only AUROC performance in their paper, due to which we are unable to provide prediction accuracy comparison in Table 5.

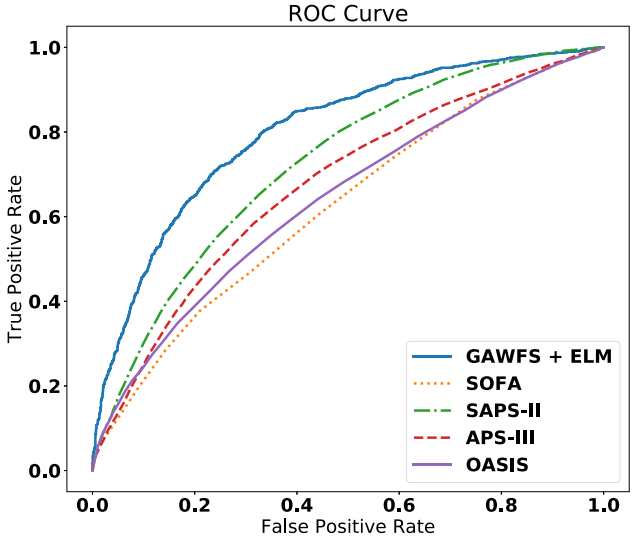

Fig. 5. Observed AUROC performance of proposed GAWFS+ELM model against various traditional severity scores.

each of these models, we re-generated the cohorts as depicted in the respective models to the highest precision possible. Cohort generation and comparison were carried out in a manner similar to that of Johnson et al. [38]. The proposed GAWFS+ELM model was then applied to the patient cohorts used by each of these works. The metrics, prediction accuracy and AUROC were considered for experimental evaluation, and the results are tabulated in Table 5.

It can be observed that the proposed GAWFS+ELM model outperformed all the state-of-the-art models in terms of both prediction accuracy and AUROC. It is to be noted that, some of the state-of-the-art models [25–27] have not reported their model's prediction accuracy values, due to which we are unable to provide these values in Table 5. Hence, we conclude that the proposed model was effective in identifying the most optimal set of lab events to be performed for each patient, to achieve cost, time and performance improvements over the existing state-of-the-art models.

3.4. Statistical significance testing of the proposed model

To further validate the proposed model's improved performance in comparison to both traditional scoring systems and the state-of-the-art machine learning models, the GAWFS+ELM model was subjected to statistical significance testing. Each model under evaluation, including the proposed as well as the state-of-the-art, was executed for a predefined number of rounds (10 rounds), and a standard-size sample of each model's results during each round, with reference to all evaluation metrics used, was collected. Interestingly, it was observed that the result samples were also normally distributed. Therefore, to check if there is a statistically significant difference between the proposed model and the models under comparison, we performed the Student's t-test [40].

The Student's t-test is a statistical hypothesis testing technique that can be used when the samples taken for a normally distributed dataset are small, and its standard deviation is not known. To start with, a null hypothesis H_0 was considered, which indicates that there is no statistically significant difference between the result samples of the proposed and existing models. The Student's t-test was performed for proposed model against each of the existing standard scoring systems, with a significance level of 5% and it was found that the p-value was lesser than 0.04 for all the metrics. Due to this, the null hypothesis H_0, was rejected for all the cases, which means that there is a statistically significant difference between the performance metrics of the proposed model and that of the existing models. It is also to be noted that the significance level is 5%, i.e., for 95% of the times, the performance of the proposed model is significantly different from that of the existing models. The results of the test of the proposed model against traditional severity scoring models are tabulated in Table 6 and that against state-of-the-art machine learning based models in Table 7.

3.5. Discussion

Based on the results of the validation experiments, several interesting observations can be made. Firstly, the proposed approach ensures that only a reduced set of features or lab events, selected by the proposed GAWFS technique, need to be measured for effectively predicting the mortality risk of a patient. For supporting this claim, we consider the patient with SUBJECT_ID: 22 in the MIMIC III dataset. The patient has spent only a single day in ICU, but the number of labevents measured amount to 80. Although this includes several labevents pertaining to the condition he or she is suffering from, the mortality risk estimation, being one of the first event performed for a patient in ICU will get delayed due to the wait time associated with other lab tests. Most existing ML based CDSS systems require a large number of features or labevents to be available to make predictions with reasonable accuracy. However, in our proposed approach, only the labevents selected by the GAWFS (10 in this case) need to be measured and input to the CDSS for mortality prediction, while still ensuring a good prediction performance (AUROC of 0.80).

Secondly, from Table 1, it is evident that the chosen patient cohort exhibits significant class imbalance. Due to the availability of lower number of samples with positive mortality labels (*expire_flag = 1*), the F-score and AUROC metrics are of crucial relevance as they actively measure the model's precision in true positive mortality prediction, i.e., patients at high mortality risk actually, predicted correctly as having high mortality risk. The high values of F-score and AUROC of the proposed model in comparison to that of traditional severity scores currently in popular use (SAPS-II, SOFA, APS-III & OASIS), means that our model can effectively capture latent relationships of features and lab events to predict mortality even in case of class imbalance exhibited

Table 6
Student's t-test results for statistical significant difference measurement between metric samples of proposed GAWFS+ELM model and traditional severity scores — SAPS-II, SOFA, OASIS and APS-III.

Metrics →		*Accuracy, Precision, Recall, F-score, AUROC*			
Method →		*SAPS-II*	*SOFA*	*OASIS*	*APS-III*
GAWFS + ELM	***P* value**	<.00001	<.00001	<.00001	<.00001
	Decision*	Reject	Reject	Reject	Reject
	Significant diff	Yes	Yes	Yes	Yes

*Significance level = 0.05.

Table 7
Student's t-test results for statistical significant difference measurement for accuracy and AUROC metrics of proposed GAWFS+ELM model and different state-of-the-art ML based models.

Metrics →		*Accuracy, AUROC*				
Method →		*Calvert et al* [23]	*Calvert et al* [24]	*Che et al* [27]	*Grnarova et al* [25]	*Harut-yunyan et al* [26]
GAWFS + ELM	***P* value**	<.02	<.04	<.04	<.01	<.01
	Decision*	Reject	Reject	Reject	Reject	Reject
	Significant diff	Yes	Yes	Yes	Yes	Yes

*Significance level = 0.05.

by the data. Our model outperformed state-of-the-art machine learning models [23–27] by a significant margin, thus underscoring its superior performance in making precise predictions for patients at higher mortality risk. Based on observed experimental results (Tables 4 and 5) and the statistical significance test results (Tables 6 and 7), it can be thus be conclusively stated that the proposed mortality prediction model can be very effective as a real-world CDSS, and can also help in effective decision towards reduced lab events. Thus, it can contribute positively to patient care and aid in making intelligent decisions in a more effective and productive way. In summary, the experimental and statistical significance test results highlight the suitability of the proposed model for use in real-world ICU mortality prediction CDSSs due to its ability to reduce lab events or features to be considered for early mortality risk prediction. Also of significant importance, is the efficacy of the ELM neural network architecture that enables higher prediction accuracy while lowering medical resource consumption footprint.

4. Conclusions

In this paper, a labevents based patient-specific ICU mortality prediction model was discussed, that is built on a Genetic Algorithm based Wrapper Feature Selection (GAWFS) and an optimized neural network architecture called Extreme Learning Machines (ELM). The GAWFS feature selection model was used to derive an optimal feature subset (i.e., a reduced set of lab events), that contribute the most towards mortality prediction of a selected patient cohort, ensuring that only these labevents be measured for prediction of mortality risk of a patient. An ELM based prediction model was built on this optimal feature set for a cohort of 31,691 patients selected from the standard MIMIC-III dataset. Performance evaluation of the proposed prediction model against models built on different feature selection techniques (ANOVA F-test, MI, RFE and SFS), selecting 10 features or labevents from a total of 578, revealed that this GAWFS+ELM model outperformed all by a margin of 4%–5% improvement in terms of prediction accuracy and 5%–8% improvement in terms of AUROC. When evaluated against four popular traditional severity scoring methods, SAPS-II, SOFA, OASIS and APS-III, the proposed GAWFS+ELM model showed a significant improvement of 15%–20% in terms of prediction accuracy and 11%–29% in terms of AUROC. Further, benchmarking against the state-of-the-art ML based mortality prediction methods applied to MIMIC-III dataset also highlighted the superior performance of the proposed GAWFS+ELM model, with an AUROC improvement of up to 14% over the state-of-the-art approach. Statistical significance testing was performed on the proposed model with Student's t-test, which confirmed the underlying statistically significant difference between the proposed and state-of-the-art models. As part of future work, we intend to benchmark the proposed models against other popular parametric scoring methods such as APACHE-IV and SAPS-3. We also intend to explore the applicability of other evolutionary algorithms and deep unsupervised feature learning methods along with other DNN based architectures and observe their performances. We also plan to validate the proposed model on real-world hospital data, so that it can be deployed and put to use in ICUs as an effective CDSS.

Funding source

This article is based on research work funded by the Govt. of India's DST-SERB Early Career Research Grant (ECR/2017/001056) to the second author. The funding agency had no role in this work; collection, analysis and interpretation of data; writing of this article and in the decision to submit this article for publication.

Conflict of interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to http://dx.doi.org/10.1016/j.asoc.2019.04.019.

References

[1] A. Awad, M. Bader-El-Den, J. McNicholas, Patient length of stay and mortality prediction: A survey, Health Serv. Manag. Res. (2017) 0951484817696212.
[2] R. Pirracchio, M.L. Petersen, M. Carone, M.R. Rigon, S. Chevret, M.J. van der Laan, Mortality prediction in intensive care units with the super ICU learner algorithm (SICULA): a population-based study, Lancet Respir. Med. 3 (1) (2015) 42–52.
[3] W.A. Knaus, J.E. Zimmerman, D.P. Wagner, E.A. Draper, D.E. Lawrence, APACHE-A physiologically based classification system, Crit. Care Med. 9 (8) (1981) 591–597.
[4] W.A. Knaus, E.A. Draper, D.P. Wagner, J.E. Zimmerman, APACHE II: a severity of disease classification system., Crit. Care Med. 13 (10) (1985) 818–829.

[5] W.A. Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, A. Damiano, et al., The APACHE III prognostic system: risk prediction of hospital mortality for critically III hospitalized adults, Chest 100 (6) (1991) 1619–1636.
[6] J.E. Zimmerman, A.A. Kramer, D.S. McNair, F.M. Malila, Acute physiology and chronic health evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients, Crit. Care Med. 34 (5) (2006) 1297–1310.
[7] J.R.G. Le, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu, P. Mercier, R. Thomas, D. Villers, A simplified acute physiology score for ICU patients, Crit. Care Med. 12 (11) (1984) 975–977.
[8] J.-R. Le Gall, S. Lemeshow, F. Saulnier, A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study, JAMA 270 (24) (1993) 2957–2963.
[9] R.P. Moreno, P.G.H. Metnitz, E. Almeida, B. Jordan, P. Bauer, R.A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.-R. Le Gall, et al., SAPS 3-From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission, Intensive Care Med. 31 (10) (2005) 1345–1355.
[10] J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C.K. Reinhart, P.M. Suter, L.G. Thijs, The SOFA score to describe organ dysfunction/failure, Intensive Care Med. 22 (7) (1996) 707–710.
[11] A.E.W. Johnson, A.A. Kramer, G.D. Clifford, A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy, Crit. Care Med. 41 (7) (2013) 1711–1718.
[12] F. Naznin, R. Sarker, D. Essam, Progressive alignment method using genetic algorithm for multiple sequence alignment, IEEE Trans. Evol. Comput. 16 (5) (2012) 615–631.
[13] M. Kaya, A. Sarhan, R. Alhajj, Multiple sequence alignment with affine gap by using multi-objective genetic algorithm, Comput. Methods Progr. Biomed. 114 (1) (2014) 38–49.
[14] Á. Rubio-Largo, M.A. Vega-Rodríguez, D.L. González-Álvarez, A hybrid multiobjective memetic metaheuristic for multiple sequence alignment, IEEE Trans. Evol. Comput. 20 (4) (2016) 499–514.
[15] R.M. Guddeti, R. Buyya, et al., A hybrid bio-inspired algorithm for scheduling and resource management in cloud environment, IEEE Trans. Serv. Comput. (2017).
[16] D. Chaudhary, B. Kumar, Cloudy GSA for load scheduling in cloud computing, Appl. Soft Comput. 71 (2018) 861–871.
[17] R. Dybowski, V. Gant, P. Weller, R. Chang, Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm, Lancet 347 (9009) (1996) 1146–1150.
[18] A. Nimgaonkar, D.R. Karnad, S. Sudarshan, L. Ohno-Machado, I. Kohane, Prediction of mortality in an Indian intensive care unit, Intensive Care Med. 30 (2) (2004) 248–253.
[19] L.S.S. Wong, J.D. Young, A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural networks, Anaesthesia 54 (11) (1999) 1048–1054.
[20] G. Clermont, D.C. Angus, S.M. DiRusso, M. Griffin, W.T. Linde-Zwirble, Predicting hospital mortality for patients in the ICU: a comparison of artificial neural networks with logistic regression models, Crit. Care Med. 29 (2) (2001) 291–296.
[21] S. Kim, W. Kim, R.W. Park, A comparison of intensive care unit mortality prediction models through the use of data mining techniques, Healthc. Inf. Res. 17 (4) (2011) 232–243.
[22] L.A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, R. Mark, A database-driven decision support system: customized mortality prediction, J. Personal. Med. 2 (4) (2012) 138–148.
[23] J. Calvert, Q. Mao, J.L. Hoffman, M. Jay, T. Desautels, H. Mohamadlou, U. Chettipally, R. Das, Using electronic health record collected clinical variables to predict medical intensive care unit mortality, Ann. Med. Surg. 11 (2016) 52–57.
[24] J. Calvert, Q. Mao, A.J. Rogers, C. Barton, M. Jay, T. Desautels, H. Mohamadlou, J. Jan, R. Das, A computational approach to mortality prediction of alcohol use disorder inpatients, Comput. Biol. Med. 75 (2016) 74–79.
[25] P. Grnarova, F. Schmidt, S.L. Hyland, C. Eickhoff, Neural document embeddings for intensive care patient mortality prediction, 2016, arXiv preprint arXiv:1612.00467.
[26] H. Harutyunyan, H. Khachatrian, D.C. Kale, G.V. Steeg, A. Galstyan, Multitask learning and benchmarking with clinical time series data, 2017, arXiv preprint arXiv:1703.07771.
[27] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (1) (2018) 6085.
[28] A.E.W. Johnson, T.J. Pollard, L. Shen, H.L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (2016) 160035.
[29] N.C. for Health Statistics (US), et al., The International Classification of Diseases: 9th Revision, Clinical Modification: ICD-9-CM, 1991.
[30] B. Chaudhry, J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S.C. Morton, P.G. Shekelle, Systematic review: impact of health information technology on quality, efficiency, and costs of medical care, Ann. Int. Med. 144 (10) (2006) 742–752.
[31] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (Mar) (2003) 1157–1182.
[32] N. Sánchez-Maroño, A. Alonso-Betanzos, M. Tombilla-Sanromán, Filter methods for feature selection–a comparative study, Intell. Data Eng. Autom. Learn. IDEAL 2007 (2007) 178–187.
[33] Z.M. Hira, D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, Adv. Bioinf. 2015 (2015).
[34] G. Forman, An extensive empirical study of feature selection metrics for text classification, J. Mach. Learn. Res. 3 (Mar) (2003) 1289–1305.
[35] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial intelligence 97 (1–2) (1997) 273–324.
[36] G. Huang, G.-B. Huang, S. Song, K. You, Trends in extreme learning machines: A review, Neural Netw. 61 (2015) 32–48.
[37] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, et al., Extreme learning machine: a new learning scheme of feedforward neural networks, in: Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, vol. 2, IEEE, 2004, pp. 985–990.
[38] A.E. Johnson, D.J. Stone, L.A. Celi, T.J. Pollard, The MIMIC code repository: enabling reproducibility in critical care research, J. Am. Med. Inf. Assoc. 25 (1) (2017) 32–39.
[39] P.A. Patel, B.J.B. Grant, Application of mortality prediction systems to individual intensive care units, Intensive Care Med. 25 (9) (1999) 977–982.
[40] B. Efron, Student's t-test under symmetry conditions, J. Amer. Statist. Assoc. 64 (328) (1969) 1278–1302.

Gokul S. Krishnan

Sowmya Kamath S.