IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 11, 2020, accepted December 21, 2020, date of publication December 24, 2020,
date of current version January 7, 2021.

*Digital Object Identifier 10.1109/ACCESS.2020.3047186*

# A Novel and Reliable Framework of Patient Deterioration Prediction in Intensive Care Unit Based on Long Short-Term Memory-Recurrent Neural Network

**TARIQ I. ALSHWAHEEN[1], YUAN WEN HAU[1], NIZAR ASS'AD[2], AND MAHMOUD M. ABUALSAMEN[3]**

[1]UTM-IJN Cardiovascular Engineering Center, Faculty of Engineering, School of Biomedical Engineering and Health Sciences, Universiti Teknologi Malaysia, Johor 81310, Malaysia
[2]School of Nursing and Midwifery, The University of Newcastle, Newcastle, NSW 2300, Australia
[3]Department of Family and Community Medicine, Faculty of Medicine, The University of Jordan, Amman 19328, Jordan

Corresponding author: Tariq I. Alshwaheen (iatariq@live.utm.my)

This work was supported in part by the Ministry of Higher Education Trans-Disciplinary Research Grant Scheme (UTM vote no. R. J130000.7845.4L842) under Grant TRGS/1/2015/UTM/02/3/3, and in part by the UTM International Doctoral Fellowship.

**ABSTRACT** The clinical investigation explored that early recognition and intervention are crucial for preventing clinical deterioration in patients in Intensive Care units (ICUs). Deterioration of patients is predictable and can be preventable if early risk factors are recognized and developed in the clinical setting. Timely detection of deterioration in ICU patients may also lead to better health management. In this paper, a new model was proposed based on Long Short-Term Memory-Recurrent Neural Network (LSTM-RNN) to predict deterioration of ICU patients. An optimisation model based on a modified genetic algorithm (GA) has also been proposed in this study to optimize the observation window, prediction window, and the number of neurons in hidden layers to increase accuracy, AUROC, and minimize test loss. The experimental results demonstrate that the prediction model proposed in this study acquired a significantly better classification performance compared with many other studies that used deep learning models in their works. Our proposed model was evaluated for two tasks: mortality and sudden transfer of patients to ICU. Our results show that the proposed model could predict deterioration before one hour of onset and outperforms other models. In this study, the proposed predictive model is implemented using the state-of-the-art graphical processing unit (GPU) virtual machine provided by Google Colaboratory. Moreover, the study uses a novel time-series approach, which is minute-by-minute. This novel approach enables the proposed model to obtain highly accurate results (i.e., an AUROC of 0.933 and an accuracy of 0.921). This study utilizes the individual and combined effectiveness of different types of variables (i.e., vital signs, laboratory measurements, GCS, and demographic data). In this study, data was extracted from MIMIC-III database. The ad-hoc frameworks proposed by previous studies can be improved by the novel and reliable prediction framework proposed in this research, which will result in predictions of more accurate performance. The proposed predictive model could reduce the required observation window (i.e., a reduction of 83%) for the prediction task while improving the performance. In fact, the proposed significant small size of observation window could obtain higher results which outperformed all previous works that utilize different sizes of observation window (i.e., 48 hours and 24 hours). Moreover, this research demonstrates the ability of the proposed predictive model to achieve accurate results (>80%) on 'raw' data in an experimental work. This shows that the rule-based pre-processing of clinical features is unnecessary for deep learning predictive models.

**INDEX TERMS** Genetic algorithm, long short-term memory, patient deterioration, prediction framework, and recurrent neural network.

The associate editor coordinating the review of this manuscript and approving it for publication was Ran Cheng.

## I. INTRODUCTION

Identifying patients of ICU who have a high deterioration risk is vital so that treatment decisions, quality assurance, and resources usage management can be guided to reduce mortality rate. Patients who are admitted to ICUs and survive hospitalization have a high mortality rate in the six months after discharge [1]. A lot of these post-discharge deaths are within patients transferred to other acute-care hospitals [2] or long-term acute care facilities [3]. Unidentified deteriorations could delay the ICU transfer of patients, which would necessitate resuscitation in as much as 67% of cases or eventually result in deaths [4]. A report by the American Health Association (AHA) in 2015 showed that about 209,000 in-hospital cardiac arrests occur annually in the United States of America (USA) [5]. There are approximately 2,300 annual cases of cardiac arrests in Swedish hospitals as reported by the Swedish Resuscitation Council, which oversees 95% of Swedish hospitals [6]. It was also found by the 2010 USA government investigation that 44% of adverse events could have been clearly or likely prevented [7]. Some researchers in New Zealand [8], UK [9], and Canada [10] used deterioration as defined by the result of health care management instead of the underlying disease process in the assessment of more than 25,000 patient records, from which 8% - 17% of admissions were related to unfavourable events, preventable deteriorations made thought to be around 37% - 51%, and 7% - 19% ended in disability or death.

To this end, several works have put forward different definitions of deterioration that are dependent on the various causes and the involved critical procedure. For instance, some studies [11]–[14] defined deterioration as the patient being transferred to an ICU or experiencing a cardiac arrest, while there are other researchers related the term to patients who are admitted, transferred to another specialized hospital for emergency surgical treatment, or died after revisiting the emergency department (ED) [15]. Quinten *et al.* [16] demonstrated that deterioration is primarily connected with organ dysfunctions like liver failure, kidney injury, respiratory failure, ICU admission, or death at a hospital. Further, deterioration have also been defined by several studies [4], [17] to be a patient's sudden transfer from the general ward to an ICU with positive pressure ventilation, vasopressors, fluid resuscitation, or any immediate procedure that may be conducted between 2 hours pre or 12 hours post transfer. Henriksen *et al.* [18] has also defined deterioration as a patient deviating from the specified normal range in the 2–24 hours interval after hospital admission. Nevertheless, in the present, the physiological importance of deterioration is appreciated and the exact definition of it is still vague among the scientific community [19].

Deterioration of patients in ICUs can be avoided by utilising technologies that detect deterioration in a timely manner, by logging several data types in health informatics systems, and processing the data by utilising software analysis models with accurate performance [20]–[23]. There are many excellent data-driven learning models could be implemented in clinical decision support system by the implementation of electronic health records (EHRs), Markov models [24] and dynamic Bayesian networks [25] to study disease development through modelling the temporal characteristics of EHRs. Moreover, preventing the occurrence of patients' deterioration in an adequate time window turns into a need in medicinal services communities and biomedical research fields. It is also imperative that hospital care quality is enhanced significantly so that unwanted results are reduced. The notable hypothesis is recent technology can be used so that models that were developed using dynamic variables (e.g., vital signs and/or lab tests) and static variables (e.g., age, gender, and admission type) are utilized to build and strengthen an automated classification algorithm that can predict deterioration accurately.

In this study, patient deterioration is defined as the patients either suddenly being transferred to ICUs from general wards (i.e., urgent admission type), or ICU patients suddenly dying [13], [14], [26], [27]. Studies [28], [29] showed a sudden ICU transfer is related with worse outcomes and increased mortality. The complex patterns in patients' longitudinal data affect the clinical interventions and ICU deaths [30]. As such, this study intends to forecast these events more reliably prior to their occurrence so suitable pre-emptive action can be taken by the hospital staff.

The main findings and contributions are as follows 1) The proposed model will assist in building a prediction model based on ''Big Data'' which has enhanced prediction accuracy. 2) This study contributes by revealing previously unknown relationships between many variables (predictors) which could result in useful diagnostic or prognostic insights. 3) This research uses definitions of deterioration, where its endpoint measure will be either mortality or sudden transfer to ICUs, which is used by researchers to obtain a better classification of patients. 4) The proposed predictive model is implemented using the state-of-the-art GPU virtual machine. This work proposes an advanced hardware that overcome challenges in gain, estimation time and testing processing time via using a virtual GPU. Moreover, the study uses a novel time-series approach, which is minute-by-minute. This novel approach enables the proposed model to obtain highly accurate results. The novel deep learning predictive model's ability to identify patterns in multivariate time-series of different clinical measurements is empirically evaluated by this research. 5) This research proposes an LSTM-RNN deep learning model that does not require feature engineering, it also proposes an optimisation model based on GA to enhance the performance metrics. 6) The ad-hoc frameworks proposed by previous studies can be improved by the novel and reliable prediction framework proposed in this research, which will result in predictions of higher accuracy. The proposed predictive model could reduce the required observation window for the prediction task while improving the performance. The rest of the paper is organized as follows.

In Section II, related works that interested in predictive models of deterioration of patients are introduced. The top-level proposed prediction framework for patient's deterioration in ICU is presented in Section III. The design idea and steps of establishing the sequential model of LSTM-RNN are introduced in Section IV. In Section V, a modified multi objective genetic algorithm is used to optimize the hyperparameters of the proposed LSTM-RNN predictive model. The results of the proposed models with LSTM-RNN and GA and comparison with other works are presented in Section VI. A conclusion is given in Section VII.

## II. RELATED WORKS

A revolutionary development in information technologies such as cloud computing [31], web hosting [32], social networks [33], and bioinformatics [34] has caused a rapid expansion in health data and many research fields. As data normally originates from multiple sources, there is a substantial opportunity to be heterogeneous. In fact, clinical elements based on patient's vital signs can be standardized across different institutions. However, the other types of clinical elements such as acuity and nursing assessments differed across the institutions because of different EHR systems and/or different customizations made by each institution.

A recent research indicated that 90% of global data today has been updated in the preceding 2 years [35]. The modern application field is facing a crucial challenge of extracting useful information from data to carry out beneficial actions. Deep learning's capacity for extracting high-level, complex abstractions, and data representations from massive data repositories, for both unsupervised data and sufficient volume of supervised data, makes it a valuable tool in Big Data analytics [36], [37]. More specifically, ''Big Data'' problems like prediction of patients' deterioration in ICU and/or general wards as well as fast information retrieval can be better tackled with the aid of deep learning [38], [39].

Deep learning has rapidly become so popular due to the reason that this approach promises a better performance (i.e. accuracy and/or AUROC) in solving several problems [40]. When as initial input data is entered by humans, they are required to work so much harder to ensure that these initial input data have good responds towards the targeted problem by a process so-called feature engineering. Problem solving has been made much easier by deep learning as the most critical step in a machine-learning workflow (i.e. feature engineering) has been completely automated [41]. Using deep learning, all features can be learnt by models in one pass instead of being reengineered through a repetitive and iterative cycle. Conventional machine learning workflow has been highly simplified by this ability, which frequently results in sophisticated multistage pipelines being replaced by a single and end-to-end deep learning model. Thus, a deep learning model learns all layers of representations jointly and simultaneously instead of in succession [42].

There has been increased interest recently in time-series data availability task [43]. This interest causes hundreds of time-series classification models to be implemented. The definition of a time-series classification problem is a classification problem that uses data that is registered by considering any notion of ordering [44]. Common deep learning models are convolutional neural networks (CNNs) [45] and recurrent neural networks (RNNs) [46]. The following sub sections describe the predictive models of patient's deterioration based on several well-known deep learning models such as CNN and LSTM. It also provides an overview of optimisation methods applied in past research that used genetic GA based on deterioration prediction

### A. PREDICTIVE MODELS BASED ON CONVOLUTIONAL NEURAL NETWORKS

Visual data processing and other healthcare problems are frequently use convolutional neural network (CNN) as a deep learning model [47]. A CNN is conceptually similar to a multilayer perceptron in concept (MLP) [48]. Each neuron in the MLP has a function for activation that maps the weighted inputs to the output [49]. When more than one hidden layer is added to the network, an MLP becomes a deep MLP. In a similar manner, a CNN is considered as an MLP that has a special structure. The CNN's special structure allows it to be both translation and rotation invariant due to the model's architecture [50]. The model's architecture has three basic layers i.e., a convolutional layer, a pooling layer, and a fully connected layer with a rectified linear activation function. Therefore, a CNN comprises one or more convolutional layers (often with a subsampling step), followed by one or more fully connected layers [51].

Rafiq *et al.* [52] developed a deep learning-based methods to identify the factors contributing to hospital readmissions of patients within 30 days, by multiple chronic concurrent (MCC) conditions and using sequential EHRs gathered from 610 patients undergoing treatment at Danderyd Hospital in Stockholm, Sweden. The study illustrated that physicians often document their communication about the patient's condition along with treatment plans and outlines as an unstructured, free-flowing text in clinical EHR notes, which makes later assessment of these EHR data tedious and time-consuming. In their study, Word2Vec approach [53] was used to convert the non-sequential records in EHR data into a vector form, and then a CNN was used to reorder and make the EHR sequential. The EHR data that has been sequenced was then applied in a recurrent neural network deep learning architecture to predict the hospital readmissions. The main disadvantage of this proposed work is that the prediction accuracy derived from texts using CNN still requires improvement.

Wickramasinghe [54] suggested that CNN and a logistic regression-based deep learning method called Deepr (abbreviation of Deep record) be combined in a hybrid method to predict unplanned readmissions after hospitalized patients were discharged. Their method was based on the concepts

used in natural language processing which involve the conversion of electronic medical records (EMRs) into a ''sentence'' of multiple phrases (with each phrase representing a visit to the hospital) separated by unique ''words'' that represent the time gap between phrases. Converting patient's medical information into a sentence makes it possible to analyze their information accurately and efficiently, as shown by the study utilising a validation dataset containing 300,000 patient records divided into three subsets based on the unplanned readmission of patients within different periods. Brand *et al.* [55] proved that one-dimensional CNNs has the potential to predict mortalities using vital signs data of variable length. This proposed CNN model appraises patient risk hourly with minimal equipment use as it needs only low-frequency vital signs. It provides patient risk scores that are periodically automatically updated. This study provided a guideline for using a proper database for implementing predictive models, which is the MIMIC-III database. The study also illustrated that implementing accurate predictive models requires more data types in addition to vital signs, which was the only type used in this model.

Chen *et al.* [56] put forward a new CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm that uses structured and unstructured data from a hospital. This study gives a guideline to obtain benefits from advances in computing hardware, particularly graphics processing units (GPUs) that can enable larger, deeper networks to be trained, as well as obtaining more accurate and less training time results. Data was divided into a training set and a test set using a ratio of 6:1. There were 606 patients in the training data set and 100 patients in the test data set. This study's primary weaknesses are the absence of a validation data set and the test data set is in low number. As a result, the prediction method of [56] cannot generate a solid conclusion and outperformed the methods used in [52], [54], [55], even though the method can achieve higher prediction accuracy.

Yi?it and I?ik [57] proposed a deep learning model based on non-invasive neuroimaging biomarkers to diagnose Alzheimer's disease (AD) and dementia. The predictive model used structural magnetic resonance (MR) brain images as the input. Two different pre-processed data sets of brain images were used to train and test the CNN models. The models achieved approximately 80% accuracy values in diagnosing both Alzheimer's disease and mild cognitive impairment. It can be concluded from this work that CNN is a valuable approach in image processing methods and can achieve acceptable prediction accuracy when it is used to predict the diagnosis of certain diseases. The model, which is trained with the back-propagated algorithm, could make predictions on the pixels of the image without feature extraction.

To conclude, the concept of weight sharing (or weight replication) in CNN reduces the number of trainable network parameters, so the network complexity is reduced and generalization is enhanced as compared to ANN [58]. CNNs are easy to train with backpropagation as compared to other ANNs due to sparse connectivity in each convolution layer [59]. CNNs are widely used in the area of deep learning because of the availability of application-oriented large databases and efficient parallel computing in GPUs [60]. CNNs deliver better performance in image resolution as compared to traditional sparse representation because it possesses higher representation capability [61]. On the contrary, CNN optimizes its weight of the convolution masks via gradient-based training scheme, which fundamentally considers self-similarity in the entire set of patches available in a relatively large number of training sets [62]. It also it suffers from a problem of gradient explosion and fails to converge quickly [63].

### B. PREDICTIVE MODELS BASED ON RECURRENT NEURAL NETWORKS

RNN presents an elegant infrastructure to process ever-evolving streams of clinical data due to loops that permit them to persist information from the past (time) [64]. RNN repeatedly going through the sequence elements for processing and sustaining a condition which comprises information relative to what it has seen so far (sequence) [65]. In general, RNNs consist of four main architectures, which are simple RNN [66], Gated Recurrent Unit (GRU) [67], LSTM [68], and Bidirectional Recurrent Neural Network (BRNN) [69]. A simple RNN is the first implemented architecture of RNNs with simple multiplication of inputs and previous outputs without any control gates. To solve the vanishing gradient problem of a simple RNN, GRU utilizes, so-called, update gate and reset gate. Fundamentally, these are two vectors that determine what information should be passed to the output. The special thing about them is that they can be trained to keep information from long ago, without washing it through time or remove information which is irrelevant to the prediction. In this architecture, mathematical operations are done on the same inputs. A GRU is unidirectional and less complex than both LSTM and BRNN architectures. It is also considered as a simplified form of LSTM architecture [70]. This study implemented a predictive model using this architecture (i.e., GRU) and benchmarked the results with the proposed predictive model based on LSTM-RNN deep learning model.

LSTM is a special type of RNNs with memory cells. In the basic architecture of LSTM, the network is given two additional gates, i.e., the forget gate and the output gate together with the existing update gate. This architecture requires more mathematical operations but offers the most controllability and better flexibility in controlling the outputs to provide better performance results. LSTM should, in theory, remember longer sequences than GRU and outperform GRU in tasks requiring modelling of long-distance relations [71]. In BRNN, the output of each layer learns from the previous layers and the next layers. In addition, every hidden layer is comprised of two opposite layers, i.e., a forward layer and a backward layer. The gating mechanism in this type of neural networks makes it as a perfect choice for long-term dependencies [72], [73]. The BRNN can be trained

without the limitation of using input information just up to a pre-set future frame. This is accomplished by training it simultaneously in positive and negative time direction. It is mostly useful for sequence embedding and the estimation of observations given bidirectional context. Thus, gradients will have a very long dependency chain. Also, it is very costly to train due to long gradient chains.

Hochreiter and Schmidhuber [74] were the first to introduce LSTMs after many researchers had studied RNNs for sequence learning. Rumelhart *et al.* [75] also conducted a previous work that was fairly significant and they proposed backpropagation through time. Elman [76] is notable for training RNNs to conduct supervised machine learning tasks with sequential inputs and outputs. The new LSTM memory cell design has maintained its closeness to the original, where forget gates have been added and normally utilized along with peep-hole connections [77], [78]. In LSTM, memory is built up by feeding the previous hidden state as an additional input into the subsequent step. This makes modelling dynamic information in time-series variables particularly suited to this model since there is a strong statistical dependency between medical events over long-time intervals. The identification of early signs of physiological deterioration can be achieved using this dependency. LSTM also permits gradients to be efficiently propagated in the training phase, alleviating the problem of vanishing gradients common in recurrent neural networks [79].

Many extensions to develop LSTM have been proposed. Che *et al.* [80], for example, concentrated on handling missing values and time irregularities. DeepCare [81] modelled the effects of time irregularities via forget gate activation. The work also showed that the interactions between disease progression and interventions were confusing. Zhang *et al.* [82] proposed an LSTM-based framework that has two levels of imperfect but informative labels to jointly learn septic shock's distinct patterns. A framework on a variant of LSTM models was proposed by the work to demonstrate septic shock's temporal progression during a long visit. The experiment results demonstrated that the proposed framework's dominance and LSTM significance in comparison with various baselines. The proposed LSTM was also shown to be robust and this was validated by the test data with three different ground-truth labels.

Lin *et al.* [83] proposed a model that combined static and dynamic features for early diagnoses and prediction of sepsis shock using convolutional-LSTM where it used in learning optimal features directly from the data itself with no human guidance. This allows latent data relationships to be automatically discovered instead of being unknown. The proposed framework was compared against other classic machine learning models that are commonly used with a prediction window size of 2 to 4.5 hours and 85.17% AUC, and with a prediction window size of 5 to 24 hours with 72.65 % AUC. It can be noted from this study that large prediction window sizes can negatively influence the performance of a predictive model. The study provides a guideline to utilize different data types and combine static and dynamic features for implementing predictive models.

Harutyunyan *et al.* [84] supported the adoption of a public benchmark suite that includes four different clinical prediction tasks inspired by the opportunities for ''big clinical data'' as discussed in Bates *et al.* [85], which are in-hospital mortality [86], physiologic decompensation [87], length of stay (LOS) [88], and phenotype classification [89] taken from the public MIMIC-III database [90]. Their research was an attempt to supply public benchmarks that could decrease the entry barrier and allow novice researchers to start with no need to acquire data access or recruit expert collaborators. This study also determined a rule to implement a predictive model for different clinical prediction tasks.

Lipton *et al.* [89] supported LSTMs proficiency in distinguishing patterns in clinical measurements with multivariate time-series. The work specifically considered multi-label classification of diagnoses. A model was trained to classify 128 diagnoses with 13 commonly but irregularly sampled clinical measurements. The work initially measured a simple LSTM network's efficiency in clinical data modelling. It then built a training strategy that was straightforward and effective where the targets were duplicated at every sequence step. The model that was proposed had a superior performance to various strong baselines, including a multilayer perceptron that was trained on hand-engineered features.

A warded patient's unplanned readmission indicates patient risk exposure and unnecessarily and avoidably wastes medical resources. Lin *et al.* [91] suggested a solution to the problem by proposing an LSTM model on comprehensive, longitudinal clinical data from MIMIC-III database to predict patient ICU readmission within 30 days of their discharge. The research combined various feature types such as chart events, and demographics. The study showed that the proposed LSTM-based solution can better predict ICU patients' high volatility and unstable status that are crucial factors in ICU readmission.

Most of the dynamic prediction models proposed by previous work do not handle multi-period data with different intervals, and patient hospital records are large scale data that have not been used efficiently to improve the prediction performance. Therefore, Junwei *et al.* [92] put forth a new traditional LSTM-based model for predicting cardiovascular disease as a type of deterioration. The irregular time interval is smoothed in this work to achieve the time parameter vector, and it adopts the LSTM's forgetting gate input for solving the prediction obstacle resulting from the irregular time interval. A guideline is also presented by this study to use the weighted summation of the prediction loss of all time slices and the prediction loss of the last time slice as a loss function of the entire model to update its parameters. In addition, the importance of test loss as a performance metric is required in modelling of the predictive model.

Long-term dependencies in LSTM are done via the additional use of an input, an output, and a forget gate, allowing every neuron to 1) select which measurements update its

current state, 2) choose which values to output to its future state, and 3) determine which previous values to forget, respectively. This is done at each input time and can be regarded as the memory cell of the neural network [93]. Plate *et al.* [94] put forward a model for predicting clinical deterioration at the intermediate care unit (IMCU) via multiple repeated measurements being incorporated using either a joint modelling approach or an LSTM-RNN. The study illustrated that the LSTM model requires the measurements to be taken at regular time intervals, which asked for inter- and extrapolation of some of the data points in this study dataset. This study provides a guideline that the primary outcome for clinical deterioration is the sudden transfer of patients to ICU or death. However, the clinical application of the proposed model is limited as the overall performance is unsatisfying.

Doctor AI [95] utilizes discretized medical codes (e.g., diagnosis, medication, procedure) from longitudinal patient visits via a purely supervised setting. The proposed model was developed and applied to longitudinal time stamped EHR data from 260K patients using GRU deep leap learning model. Encounter records were input to GRU-RNN to predict (all) the diagnosis and medication categories for a subsequent visit. The proposed model was tested on a large real-world EHR datasets and achieved 79.58% recall@30. Thus, in medical practice, incorrect predictions can sometimes be more important than correct predictions as they can degrade patient health and it would be more useful to learn to perform better than average. To avoid overfitting, the study used dropout between the GRU layer and the prediction layer (i.e., code prediction and time duration prediction). Hence, this gives a guideline to utilize such technique (i.e., dropout) to prone overfitting.

In time series prediction, it has been notified that missing values and their missing patterns are repeatedly correlated with the target labels, a.k.a., informative missingness. Che *et al.* [80] proposed a novel deep learning models, namely GRU-D, as one of the early attempts that is based on GRU. It adopts two representations of missing patterns which are masking and time interval that efficiently integrates them into a deep model architecture so that it not only captures the long-term temporal dependencies in time series, but also uses the missing patterns to acquire better prediction results. The study concludes that if the missingness is not informative at all, or the inherent correlation between the missing patterns and the prediction tasks are not clear, then the proposed model possibly will gain limited improvements or even fail. The proposed models are only evaluated in retrospective observational study settings, which is due to the inherent limitation of the publicly available datasets utilized in the study. Thus, it is important to consider the generalization task of an adopted predictive model.

Peiffer *et al.* [96] proposed a predictive model based on a Bidirectional Long Short-Term Memory (BiLSTM) network and bidirectional recurrent neural network (BiRNN) and utilizes different monitored parameter sequences from over 2000 ICU admissions to predict the presence of sepsis. Temporal models have a slight advantage in predicting the outcome close to the time the blood sample test was taken but are noticeably better than other models in predicting this test many hours upfront. In the advanced BiLSTM network, hidden states from all time steps have direct influence on the output node. While in the simpler BiLSTM network, the hidden state values first must ripple through the whole LSTM chain.

Pan *et al.* [97] performed different experiment using various deep learning models for comparison to predict AD. The study proposed a simple RNN model based on time series that utilizes the common AD diagnostic attribute values and combines the data of three-time nodes. It also proposed an LSTM model that uses the same input data and function. Moreover, the study proposed a GRU model simplified by LSTM that has the same input data and the same model function. Further, the study utilizes a bidirectional LSTM plus Attention mechanism as a model. In addition, the patient's basic information, genetic data and three time points of neuropsychology scale were used as input to predict the development trend of the patient's condition, which is normal (NL), mild cognitive impairment (MCI) or Alzheimer's disease (AD).

It can be observed that RNNs of different varieties have become popular in modelling clinical time series, as they are able to learn complex nonlinear functions of their input without the need for extensive domain knowledge or feature engineering. This better allows for learning expressive representations and discovering unforeseen structure than methods that rely on hand-crafted features. Among the previous studies that are based on prediction of deterioration of patients, it can be observed that LSTM cells are with more complex form of module that can be used to alleviate issues of vanishing and exploding gradients when trained via gradient-based methods. LSTM-RNNs are standard baselines as they tend to work well in practice and often give competitive performance compared with more sophisticated architectures. Unlike naive RNNs, RNNs built using LSTM cells can capture long range dependencies and nonlinear dynamics. However, LSTM networks have internal contextual state cells that act as long-term or short-term memory cells. The output of the LSTM network is modulated by the state of these cells. This is a particularly important property for the prediction of the neural network to depend on the historical context of inputs, rather than only on the very last input. The related works provide a guideline to the proposed elements of the baseline deterioration prediction model in this study. These elements involve the choice of data pre-processing, appropriate database, and methodology to implement the proposed models.

### C. OPTIMSATION METHODS APPLIED IN LITERATURE THAT USED IN GENETIC ALGORITHM

It is important to provide an overview of optimisation methods applied in past research that used genetic algorithm (GA) based on deterioration prediction. Over the past decade, GA has been used in various fields at varying success rates (Ding and Fu 2016). Jiang *et al.* (Jiang, Peng *et al.*

2017) proposed the Probability Distribution Patterns Analysis (PDPA) method to derive significant information from the continuous blood pressure time series. The research then used a machine learning model by merging GA and SVM to identify the representative features for successful classification. The acquired accuracy for classifying and predicting hypotension was 80.8%, 78.2% for sensitivity, and 81.5% for specificity when applied to the validation cohort. Moreover, Choudhury *et al.* (Choudhury and Greene 2018) proposed a study that considered patient readmission risk as the objective for optimisation and utilised a valuable risk prediction approach to tackle unplanned readmissions. Additionally, GA and Greedy Ensemble algorithm were used to optimize the developed model constraints.

Unexpected disease deterioration causes a lot of sudden health issues to patients. Lai *et al.* .(Lai, Tan *et al.* 2020) suggested a model of clinical support to predict readmissions for patients admitted with all-cause conditions. Specifically, the research designed an online web service system based on integrated GA and SVM (IGS) to assist physicians in detecting patients that have potentially a higher risk of all-cause readmissions after discharge. When necessary, the patients are given appropriate interventions to reduce their morbidities and mortalities, and to lower their healthcare costs. GA was used to select significant variables and adjust SVM parameters (C and $\gamma$). Pre-defined GA parameters were used in the study, for example the initial chromosome population number that was set to 10.

Extracting clinical phrases from nurses' notes has been recently used as an advanced method to find patient deterioration risk factors. Korach *et al.* .(Korach, Yang *et al.* 2020) put forward the use of multiple natural language processing approaches which contain language modelling, word embeddings, and two phrase mining methods (TextRank and NC-Value) to identify the quality phrases that are clinically significant from unannotated nursing notes. The study combined the two mining methods in GA to complement each other. The study suffered from low performance with average precision of 0.890 to 0.764, and it was also a single-objective optimisation approach.

Incapacitated patients are assigned temporary emergency medical service (EMS) centres based on their geographical locations and final mortality risk value. Gao *et al.* .(Gao, Zhou *et al.* 2017) utilised GA with modified fuzzy C-means clustering algorithm to find the temporary emergency medical service centres and their allocations. The study aimed to minimise the total mortality risk value. However, the parameters and settings of the predictive model were not optimised. The mortality of cardiovascular patients in an ICU has been predicted using different models with some extent. However, several models require many patient registrations, which in the majority of cases is not possible to record all data.

Moridani *et al.* .(Moridani, Setarehdan *et al.* 2018) employed a roulette wheel to benefit from the selection operator. The prediction of coronary artery disease also used secondary input and output membership functions (optimised with GA). Sensitivity and specificity were shown to be the best findings with a prediction window ranging between 0.5 hour and 1 hour before the patients with coronary artery disease died with area feature. Table 2.5 provides recent previous studies (Gao, Zhou *et al.* 2017, Jiang, Peng *et al.* 2017, Choudhury and Greene 2018, Moridani, Setarehdan *et al.* 2018, Korach, Yang *et al.* 2020, Lai, Tan *et al.* 2020) that have adopted GA to optimise classification performance of their deterioration proposed diagnosis models. Their application, fitness function, strengths, and weaknesses are summarised.

## III. TOP-LEVEL PROPOSED PREDICTION FRAMEWORK FOR PATIENT'S DETERIORATION IN ICU

Big health data has presented more opportunities to health data analysis and health service development through innovative approach. This study proposes a novel and reliable prediction framework to enhance ICU patient deterioration prediction based on the paradigms of big data analysis and modern GPUs to increase the efficiency of data processing [98]. The framework consists of five layers, namely dataset layer, exploration layer, prediction layer, optimisation layer and performance evaluation layer. In this study, the patient deterioration is defined as the patients either suddenly being transferred to ICUs from general wards (i.e., urgent admission type), or ICU patients suddenly dying [13], [14], [26], [27]. The main novelty of this proposed prediction framework is the novel and reliable prediction framework offers the transformative characteristic that allows a model to learn all layers of representation jointly and simultaneously, instead of in succession. This proposed novel framework is based on the combination of LSTM deep learning algorithm as prediction model, together with the optimisation approach based on multi-objective genetic algorithm to supply the LSTM predictive model with more comprehensive and reliable time-series data to achieve better prediction performance for higher accuracy and reliability.

The novel and reliable prediction framework and all the associated layers are applied in a manner which fully utilizes the advantages of deep learning as shown in Figure 1 below.

The preliminary knowledge and basic concept that related with each layer will be presented in detail in the following subsections.

### A. DATASET LAYER

The adoption of EHR system is increasing in recent years to encourage paperless system in the medical field [99]. In the USA, the number of non-federal severe medical centres that utilize EHR system increased more than eight times to reach 75.4% during the interval between 2008 and 2014 [100]. With EHRs providing ideal support, it enables real-time trend analysis and score visualisation with an easy clinical deterioration display and adjustable thresholds per patient. In a constant monitoring environment particularly, this could probably lead to earlier interventions and better patient outcomes [101]. The structure of the dataset layer on the proposed prediction framework is based on the utilization of publicly available

open-source research dataset, so called MIMIC-III database [102]. The main function of the dataset layer is to provide all the necessary input data to the prediction layer for the patient deterioration prediction in the proposed prediction framework. There are four main categories of input parameter, namely static parameter which involve three demographic attributes, dynamic parameter which involves seven vital signs, dynamic parameter consists of eight laboratory test and dynamic parameter of the level of consciousness. As a result, all four categories contribute to total of 19 attributes as summarized in Table 1.

The latest version of MIMIC-III (i.e., version 1.4) database contains rich de-identified data which include 46,476 patients with 61,532 ICU stays. There are 53,423 distinct hospital admission records for adult patients and 8,100 neonates. The detailed data is associated with patients admitted to several care units between June 2001 and October 2012. The median age of adult patients is 65.8 years old (Q1–Q3: 52.8–77.8), with female patients make up 44.1% of all patients. Each hospital admission has a mean of 4,579 charted observations and about 380 laboratory measurements [100]. The database contains about 30 parameters recorded every minute with an amplitude resolution of up to 16 bits.

Supplementary details are acquired from different sources to establish the complete database, such as demographic data like gender, date of birth, ethnicity, and in-hospital mortality. Furthermore, the database comprises laboratory measurements, such as chemistry tests and microbiology results. Other than billing-related information, other info such as International Classification of Diseases, 9$^{th}$ Edition (ICD-9) codes, Diagnosis Related Group (DRG) codes, and Current Procedural Terminology (CPT) codes are also available. Moreover, Echo reports, ECG reports, and radiology reports are available for both inpatient and outpatient stays. More specifically, in MIMIC-III database CATEGORY and DESCRIPTION define the type of note recorded. For example, a CATEGORY of 'Discharge summary' implies that the note is a discharge summary, and the DESCRIPTION of 'Report' implies a full report while a DESCRIPTION of 'Addendum' indicates an addendum (additional text to be added to the previous report). However, before the data was integrated into the database, it was first de-identified in conformity with all of the 18 rules defined in the MIMIC-III database has 26 different tables. Tables are merged with each other using different identifiers (IDs), such as SUBJECT_ID, hospital admission identifier (HADM_ID), and ICUSTAY_ID. Correspondingly, each of these tables contains detailed information about specific patient's data. These tables are categorised into four main categories. The first category determines, and tracks patients' stays. The second category provides ICU data and involves 6 tables. The third category provides hospital data and contains 10 tables. The fourth category implies dictionaries for cross referencing codes versus their competent definitions [90], [100]. All the categorized and associated identifiers are summarized in Table 2.

As this study utilizes different types of data (i.e., dynamic, and static data) to achieve higher predictive performance based on the MIMIC-III database. This framework utilizes many tables, such as CHARTEVENTS and LABEVENTS for dynamic data, and ADMISSIONS and PATIENTS for static data, respectively. All these data are then fetched to the proposed prediction model and optimisation model for ICU patient deterioration prediction. In this dataset layer, some significant tools and libraries are used, such as PostgreSQL that utilizes and expands the Structured Query Language (SQL) combined with many features that safely store and scale the most complicated data workloads.

### B. PREDICTION LAYER

This layer is the most crucial layer in the proposed patient deterioration prediction framework. This layer consists of different steps which are windowing, data pre-processing, feature selection, data merging, and prediction model based on LSTM-RNN as shown in Figure 2. In fact, windowing is the extraction of successive data with size N from a time-series data [103]. This step involves observation window (OW) (i.e., duration of observation before prediction window from which data are used [104], [105]) and prediction window (PW) (i.e., the amount of time before clinical diagnoses [104]). Moreover, data pre-processing is an essential step in this work to group and arrange a biological dataset into a proper manner before further data processing [106]. Besides, feature selection enables the predictive model to process and apply various transform that converts datasets into a much more usable and desirable form. This study modified the RNN architecture to contain both the dynamic as well as the static data by data merging to combine various types of different data for performance enhancement of the proposed prediction model based on LSTM-RNN to predict deterioration of patients in ICU due to its capacity of capturing long-term sequential patterns, by employing LSTM in a deep learning architecture.

In fact, the LSTM architecture has three gates (input $i$, forget $f$, and output $o$) and a cell memory vector. The input gate determines how the state of the memory cell is varied by the incoming vectors $x_t$. The output gate permits the memory cell to affect the output. Lastly, the forget gate lets the cell remember or forget its previous state [107]. Moreover, the memory unit has two valves, which are the forget valve and the new memory valve [108].

A single layer neural network controls the forget valve. Besides, the inputs of the neural network, a bias vector $b_o$, is also added to the forget valve. The activation function is the sigmoid function $\sigma$ and the forget valve acts as the output vector. The forget valve will be provided to the old memory $C_{t-1}$ by an element-wise multiplication [108]. A second neural network represents the new memory valve that forms the second valve. This valve takes the same inputs as the forget valve and its activation function is the hyperbolic tangent *tanh* [109]. In fact, this valve controls how much new memory affects the old memory. The output will be multiplied with the

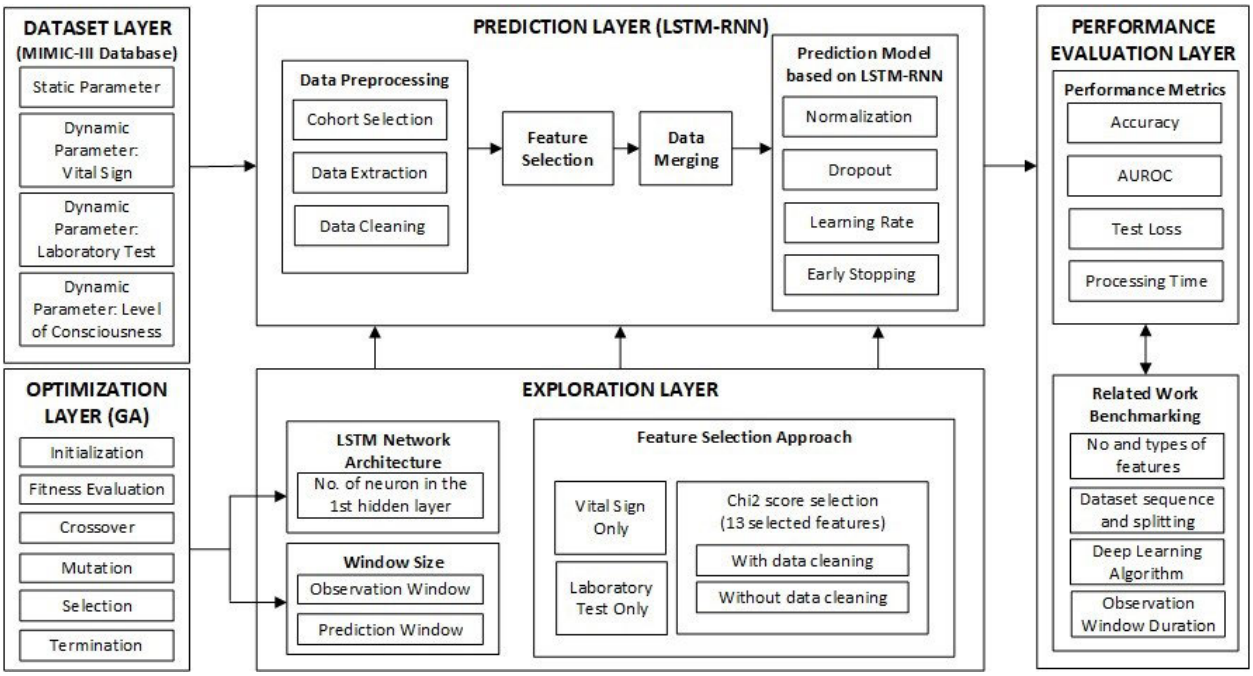


**FIGURE 1.** Top-level prediction framework for patient deterioration.

**TABLE 1.** Input data summary of dataset layer

| Input Data | Attributes |
|---|---|
| Static parameter: demographic data | First admin age, admission type, gender |
| Dynamic parameter: Vital Sign | Systolic blood pressure (SysBP), diastolic blood pressure (DiasBP), mean blood pressure (MeanBP), body temperature (T), respiratory rate (RR), Blood oxygen saturation (SpO2), heart rate (HR) |
| Dynamic parameter: Laboratory test | Glucose, Creatinine, Potassium, Sodium, Blood Urea Nitrogen (BUN), International Normalized Ratio (INR), Partial Thromboplastin Time (PTT), Prothrombin Time (PT). |
| Dynamic parameter: Level of consciousness | Glasgow Coma Scale score |

new memory valve and added to the old memory to produce the new memory [108]. In the end, the output for the LSTM network will be generated. The new memory is where the inputs and a bias vector control the output valve. This valve is responsible for the memory that will be generated to the next LSTM unit [108].

In this model, several techniques are also applied, namely normalization, dropout, learning rate and early stopping. In fact, dropout is a technique where randomly selected neurons are ignored during training. They are ''dropped-out'' randomly. Moreover, normalization is an adopted technique to change the value of the numeric variable in the dataset to a typical scale, without misshaping contrasts in the range of value, this has the effect of settling the learning process and significantly reducing the number of training epochs needed to train a deep neural network. On the other hand, learning rate is defined as the amount that the weights are updated during training. Particularly, it is a configurable hyperparameter utilized in the training of neural networks with a small positive value, often in the range between 0.0 and 1.0. However, early stopping is a form of regularization utilized to avoid overfitting when training a learner with an iterative method, it implies a guidance as to how many iterations can be run before the learner begins to over-fit.

Moreover, time series data in real applications are often collected over a long span of time such as electronic health records. However, this is the main layer, and its aim is predicting ICU patient deterioration by employing LSTM in a deep-learning architecture. The LSTM predictive model has individual layers that are trained to produce a higher-level representation of the patterns observed, based on the data it receives as the input from the layer below. This main layer can accumulate time-series data to make accurate predictions of future unseen data.

This subsection briefly introduces the structure of LSTM which is a variant of RNN. Rather than conducting classification at each time step separately, LSTM introduces an LSTM cell to model the transitions between several time steps. Every LSTM cell involves a cell state $C_t$ which serves as memory and controls the information flow, added,

**TABLE 2.** Description of tables in MIMIC-III database

| No | Table Name | Category | Description |
|---|---|---|---|
| 1 | ADMISSIONS | Patient tracking | Defines the role of caregivers |
| 2 | PATIENTS | Patient tracking | Defines each SUBJECT_ID in the database |
| 3 | ICUSTAYS | Patient tracking | Defines each ICUSTAY_ID in the database |
| 4 | CALLOUT | Patient tracking | Provides information about ICU discharge planning |
| 5 | TRANSFERS | Patient tracking | Physical locations for patients throughout their hospital stay |
| 6 | CHARTEVENTS | ICU data | Contains all charted data for all patients |
| 7 | INPUTEVENTS_CV | ICU data | Input data for patients |
| 8 | INPUTEVENTS_MV | ICU data | Input data for patients |
| 9 | DATETIMEEVENTS | ICU data | Contains all date formatted data |
| 10 | OUTPUTEVENTS | ICU data | Output data for patients |
| 11 | PROCEDUREEVENTS_MV | ICU data | Contains procedures for patients |
| 12 | CAREGIVERS | Hospital data | Defines the role of caregivers |
| 13 | CPTEVENTS | Hospital data | Involves current procedural terminology (CPT) codes, that facilitate billing for procedures performed on patients. |
| 14 | DIAGNOSES_ICD | Hospital data | Involves ICD diagnoses for patients, most notably ICD-9 diagnoses |
| 15 | DRG_CODES | Hospital data | Contains diagnosis related groups (DRG) codes for patients |
| 16 | LABEVENTS | Hospital data | Contains all laboratory measurements for a given patient |
| 17 | MICROBIOLOGYEVENTS | Hospital data | Contains microbiology information |
| 18 | NOTEEVENTS | Hospital data | Contains all notes for patients |
| 19 | PRESCRIPTIONS | Hospital data | Contains medication related order entries, i.e., prescriptions |
| 20 | PROCEDURES_ICD | Hospital data | Contains ICD procedures for patients, most notably ICD-9 procedures. |
| 21 | SERVICES | Hospital data | Lists services that a patient was admitted/transferred under |
| 22 | D_CPT | Dimension table | High-level definitions for current procedural terminology (CPT) codes |
| 23 | D_ICD_PROCEDURES | Dimension table | Definition table for ICD procedures |
| 24 | D_ITEMS | Dimension table | Definition table for all items in the ICU databases |
| 25 | D_ICD_DIAGNOSES | Dimension table | Definition table for ICD diagnoses |
| 26 | D_LABITEMS | Dimension table | Definition table for all laboratory measurements |

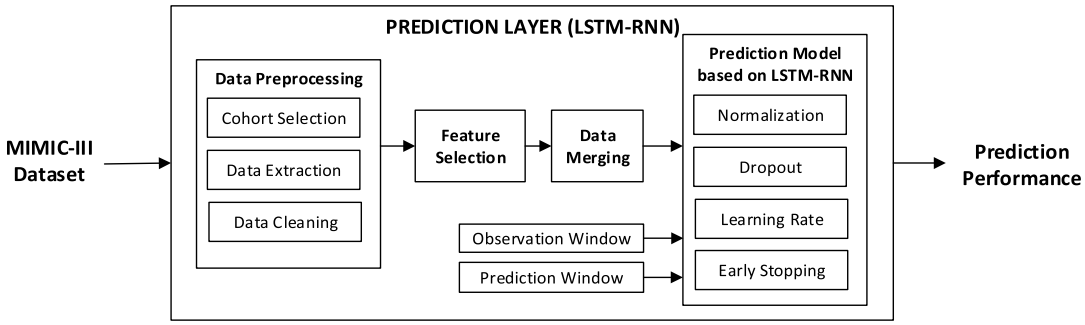


**FIGURE 2.** Prediction layer.

removed, or unchanged. Every LSTM cell also outputs a hidden representation $h_t$ which is a high-level representation of information at current time step, and can be utilized for classification at time $t$.

The cell state $C_t$ is generated by combining memory from previous block $C_{t-1}$ and output of previous block $h_{t-1}$. Specifically, the LSTM structure first decide what information should be added to the current cell state by generating a new candidate cell state. The LSTM structure also generates an input gate to filter the new added information. The gating variables in LSTM cell such as input gate are computed by a sigmoid function with combination of $x_t$ and $h_{t-1}$, while the candidate cell state is generated by a *tanh* function as follows:

$$a_t = tanh(W_a{}^* x_t + U_a * h_{t-1}) \quad (1)$$

$$i_t = \sigma(W_i{}^* x_t + U_i * out_{t-1}) \quad (2)$$

where $Ua \epsilon R^{H\times H}$, $Wa \epsilon R^{H\times D}$ and $Wa$, $Ua$, $Wi$, and $Ui$ denotes two sets of weight parameters for generating the input gate (i.e., a way to ''learn'' new information) and candidate gate (i.e., a way to ''ignore'' new information) respectively. Hereinafter, the previous formulas omit the bias terms as they can be absorbed into weight matrices. After that, the LSTM structure creates a forget gate using a sigmoid function to remove information from the past:

$$f_t = \sigma(Wf^* x_t + U_f{}^* h_{t-1}) \quad (3)$$

where $W_f$, $U_f$ denotes the weight parameters utilized to generate the forget gate layer.

In the way of forgetting old information from the old state $C_{t-1}$ and filtering new information from the candidate cell state at time $t$, the new cell state is obtained as follows

$$C_t = f_t{}^* state_{t-1} + a_t{}^* i_t \quad (4)$$

At the end, the hidden representation is generated by filtering the obtained new cell state by an output gate layer $O_t$

$$O_t = \sigma(W_o{}^* x_t + U_o{}^* h_{t-1}) \quad (5)$$

$$h_t = tanh(C_t)^* O_t \quad (6)$$

where $Wo$, $Uo$ are weight parameters that are utilized to generate the output gate layer. The output gate determines the information to output from $C_t$ to $h_t$. Thus, the LSTM architecture has three gates (input $i$, forget $f$, and output $o$) and a cell memory vector. The input gate determines how the state of the memory cell is varied by the incoming vectors $x_t$. The output gate permits the memory cell to affect the output. Lastly, the forget gate lets the cell remember or forget its previous state [107].

### C. OPTIMISATION LAYER BASED ON GENETIC ALGORITHM

Deep learning algorithms utilizes novel methods for classification that tries to improve the training speed, while still striving for a comparable or better classification performance. In this study, a modified Genetic Algorithm (GA) is applied along with the proposed deep learning predictive model based on LSTM-RNN to identify the potential informative hyperparameters for predicting patient deterioration.

GA is a search heuristic algorithm that tries to mimic biological process of natural selection. It is used widely to generate useful solutions for optimisation and search problems. GA can be broken down into four sub parts which are individual, chromosome, population, and fitness. In fact, individual chromosome can be defined as a single entity that closely resembles a possible solution amongst a pool of solutions. However, every individual has a chromosome that is generally bit encoding of the representative solution in a binary format. Moreover, a population is defined as a group of individuals (i.e., solutions) amongst which reproduction takes place. Finally, fitness which is a characteristic representative of how fit or good the individual is.

The basic process for a genetic algorithm is consists of several phases, which are initialization, fitness evaluation, selection, mutation, and termination. Initialization creates an initial population which is usually randomly generated and can be of any desired size, from only a few individuals to thousands. Fitness evaluation is another step where each member of the population is estimated and a 'fitness' for that individual is calculated. The fitness value is calculated by how well it fits with the desired requirements. Selection is another important step in the process of GA since there is a need to constantly improving the populations overall fitness. Selection helps to do this by discarding the bad designs and only keeping the best individuals in the pop-

ulation. The fourth step in the process of GA is crossover where new individuals are created by combining aspects of selected individuals. By combining certain traits from two or more individuals, the aim of crossover is to create an even 'fitter' offspring which will inherit the best traits from each of its parents. Mutation induces randomness into the chromosomes of the solution. It aims to bringing back some characteristic genes which were lost during the process of crossover.

The same processes are iteratively repeated to generate the next generation of population, go through the fitness evaluation, selection, crossover, and mutation until the termination criteria is reached. GA is considered as one of the extremely innovative approaches to perform optimisation. GA can be utilized better than conventional approaches. It is also able to handle datasets that have lots of features. GA also does not require specific knowledge of the problem being studied. Also, GA could be simply parallelized in computer clusters. The aforementioned steps describe the primary steps required to design a chromosome. Based on the literature review, the appropriate parameters and settings are selected. Then, they are encoded into the binary string with a certain bit length and put into the chromosome. For the OW, 10 bits are chosen. 5 bits are selected as the number of bits to perform the chromosome (i.e., in the first hidden layer). 10 bits are chosen for the PW. These bits are put into the whole chromosome. Here, in the step, a modification was performed to obtain the targeted hyperparameters namely length of observation window, number of hidden neurons in the first hidden layer, and the size of the prediction window.

There are three basic termination criteria of a GA algorithm, (i) there is no further improvements, (ii) the given maximum number of generations is reached, and (iii) the objective function reaches minimum value. In this work, the basic prediction framework is padded with this optimisation layer to give more composite representations to the predictive model for ICU patient deterioration. The aim of the proposed model is to optimize certain framework parameters, such as the size of the observation window and the prediction window, as well as the number of neurons in the first hidden layer, which will be further discussed in exploration layer in next sections.

### D. EXPLORATION LAYER

To get the best performance of the proposed prediction framework, many parameters are explored to identify the best configuration at the prediction layer from data pre-processing until final prediction model based on LSTM-RNN as shown in Figure 3. These parameters include the window size of the observation window and the prediction window, the number of neurons in the first hidden layer of the prediction model as well as feature selection approach. As feature selection is one of the most crucial process in prediction layer to train, validate and test the prediction model, this work has applied three different approaches of feature selection for performance comparison. There are features based on vital sign only, features based on laboratory only, and finally the feature selection based on ''selecting a percentile'' using Chi2 score. The performance evaluation proves that the Chi2 score feature selection approach is the best approach to strike the balance between prediction accuracy and the computation performance.

It is also important to note that in the exploration layer, the exploration of window size and LSTM network architecture is an automated process based on GA algorithm, whereas the feature selection approach is executed manually by comparing the framework performance.

### E. FRAMEWORK EVALUATION LAYER

As the name implied, the framework evaluation layer as shown in Figure 4 is meant to measure the performance of the proposed patient deterioration prediction framework in terms of prediction accuracy, AUROC, test loss and processing time. To evaluate the performance, the MIMIC-III dataset at the dataset layer is divided into three parts, i.e., training dataset, validation dataset, and testing dataset, with the ratio of 70%, 15% and 15%, respectively.

Training dataset is a set of data used to fit the model parameters, such as weights of connections between neurons in the proposed prediction model. Firstly, the weights of the network are allocated random values, consequently the network simply implements a series of random transformations. Certainly, its output is far from what it should preferably be, and the loss score is therefore very high. However, with every single example of the network processes, the weights are adjusted in the correct direction to minimize the loss score. This is the training loop, which, repeated an adequate number of times (typically tens of iterations over thousands of examples), produces weight values that reduce the loss function. A network with a minimal loss is one for which the outputs are as close as they can be to the targets: a trained network. On the other hand, the validation dataset evaluates the loss at the end of each epoch. It is used to give an estimation of model skill while tuning model's hyperparameters [110]. The dataset utilized to assess the final model performance is called the ''test set''.

Conventionally, most of the studies in literature split the dataset into just two sets which are training and testing datasets. However, this will give a biased framework performance as using information from the test set during the model training in any way is considered as a ''peeking'' behavior. As a result, many researchers strongly suggested to totally separate and lock the test set from model training until the model tuning is completed [111].

In the framework evaluation layer, the performance of the proposed prediction framework will also benchmark with many previous related works in terms of the number and types of features obtained from different database targeted for different prediction task, the dataset sequence and the splitting for performance evaluation, different prediction model based on various deep learning algorithm, as well as the observation window size.

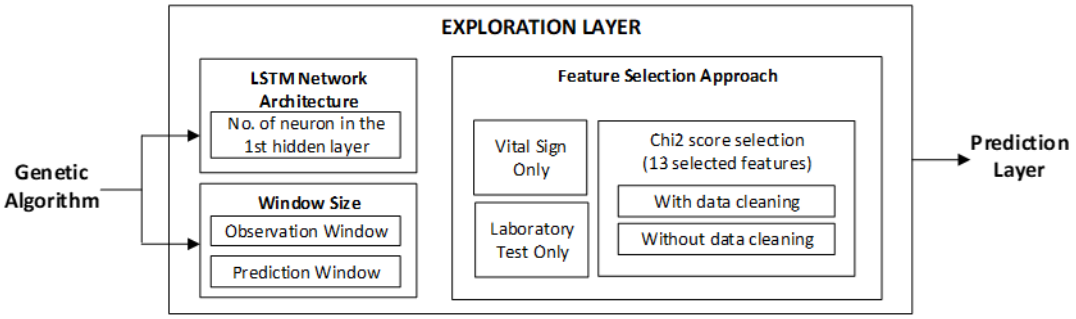


**FIGURE 3.** Exploration layer.

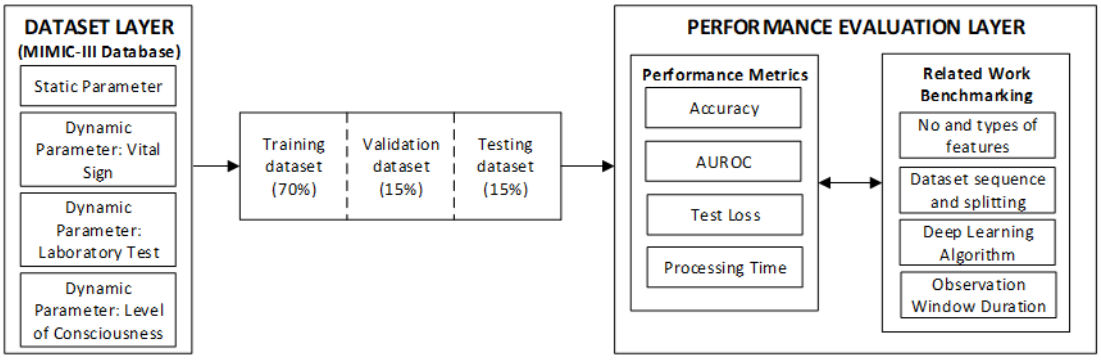


**FIGURE 4.** Framework evaluation layer.

## IV. MODELLING AND ALGORITHMIC DEVELOPMENT IN THE PROPOSED PREDICTIVE MODEL

In achieving the aim of implementing a novel and reliable prediction framework that enables different techniques for data preparation, training, and prediction to be deployed to carry out experiment, a deep learning approach of LSTM-RNN to predict patient deterioration in ICU was chosen in this study as the suitable method for a baseline analysis. The hypothesis that changing window sizes will result in a balance between the accuracy and time prediction of different variables for deterioration prediction is adopted. In addition, an experimental set using different types of data was carried out to demonstrate the influence exerted by data types to obtain better results. A generic experimental procedure that followed the conventional design cycle of pattern recognition was developed. This includes vital components of the raw variables pre-processing tasks; dividing the data into training, validation, and testing data sets; performing experiments at different sizes of observation window and prediction window; and optimisation and calculation results. Under this framework, the impact of different types of data time-series and different sizes of windows on the prediction performance will be analysed. The implementation of the baseline prediction framework to perform the experimental works is shown in Figure 5.

The long size of observation window (i.e., 24 hours and/or 48 hours) requires more data storage and more computations to obtain acceptable results. Therefore, this work aims to decrease the size of observation window from 24 hours to only 4 hours (i.e., an 83% reduction) while still acquiring an acceptable accuracy rate in its prediction. It also aims to predict deterioration before a time that enables the medical team to save lives, increase morbidity, and decrease mortality. In other words, predict deterioration before 1 hour, 2 hours, 3 hours, 4 hours, 5 hours, and 6 hours. These goals show the importance of this study to predict deterioration before an acceptable prediction window and using small observation window. This study acquires key results that illustrate the strength points performed in this study.

Based on the works of Purushotham *et al.* [112] and Johnson *et al.* [113], adult patients are considered to comprise any patient whose age was above 15 years at the ICU admission time. This research targeted this group. Many patients in MIMIC-III database have more than one ICU admission. Hence, in this study, the data associated with the first admission is only utilized and this is done to avoid potential details leakage in the upcoming analysis, as well as to maintain comparable experimental settings [112], [113]. For every admission, all records of vital signs and laboratory values comprised an interval of 5 hours starting from the admission of patients to the ICU for both control and case groups (i.e., 5 hours prior to sudden death or sudden transfer for the group that suffered from clinical deterioration), such as [11]. However, patients who stayed under 5 hours in the

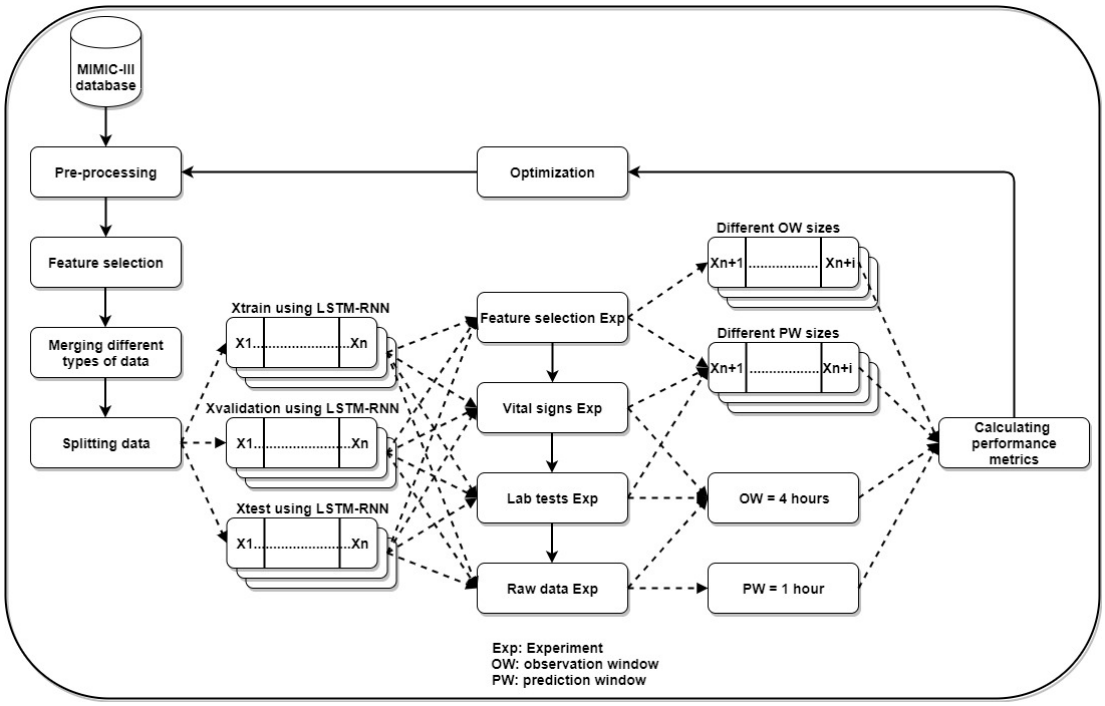


**FIGURE 5.** Implementation of the baseline prediction framework to perform the experimental works.

ICU were placed under the exclusion criteria. Thus, admissions that met these criteria were eligible for inclusion. The study employs minute-by-minute time-series, which involves 8,031 kilobytes (KB).

This study defines a dataset involving 399 patients like [114]. The dataset contains 2,274,300 samples organised as 119,700 rows and 19 columns in addition to the time-series column and the label column. The dataset is used to define training, validation, and testing data sets. Building solid models requires selecting the significant features that affect the performance of the model while implementing it [115]. Feature selection is performed to obviate overfitting and develop model performance. Besides, it is done to enhance the speed of models in terms of time consumption. Feature selection can also imply deeper insight into the embedded techniques that provided the data. To select correlated features, this research applied a methodology, namely Selecting a Percentile of Features using Chi2 methodology [116]. This methodology selects features as stated by a percentile of the highest scores. The selected features involve dynamic (i.e., sequential) features and static (non-sequential) features. The sequential features involve vital signs (i.e., HR, SysBP, MeanBP, RR, $SPO_2$), GCS, and laboratory test results (i.e., Glucose, PTT, PT, and BUN). The selected static features were age and admission type. In the proposed LSTM-RNN predictive model, each of the dynamic features was represented by the sequences of minute-by-minute values corresponding to 5 hours in the first calendar day after ICU admission. If there was no value reported during a particular minute, a missing value was set.

In clinical datasets, static information like age, gender, blood type, and admission type combined with dynamic information (i.e., sequences of data) are recorded for hospitalized patients, either in general wards or in ICUs. The study uses the data collected from MIMIC-III database that contains ICU patients' complete information. It is notable that the medical data sets have become longer (i.e., more samples are collected through time) and wider (i.e., store more variables). Therefore, it is necessary to merge various data types prior to their being imported to any predictive model that analyses the complex relationships between many time-evolving variables. In this study, the static information is concatenated with the dynamic information in a merge step using a category of joins called ''many-to-one'' joins. This is one of the strength points involved in this study. In fact, some joins are those in which one of the two key columns has duplicate entries. For the many-to-one case, the DataFrame that results will preserve these duplicate entries as appropriate. The output then undergoes concatenation with LSTM's hidden states at each time step. These LSTM hidden layers output will be fed into a dense hidden layer before the output layer.

## V. PROPOSED OPTIMISATION MODEL BASED ON MULTI OBJECTIVE GENETIC ALGORITHM

Many widespread applications of learning models from customer targeting to medical diagnosis, evolve due to sophisticated relationships among settings and parameters. In addition, optimisation is a method to discover input that are most significant for a predictive model. This particular

method can be used to identify and remove parameters and settings that are unwanted, insignificant, and superfluous, and which have no effect or minimise the predictive model's accuracy. Also, the approach determines the most appropriate values that provide higher performance metrics. The process of selecting appropriate sizes of windows and number of neurons in hidden layers needs much computational effort and, if the numbers of neurons and sizes of observation window and prediction window are significant, the process becomes impractical.

Consequently, the requirement of rational approaches that select appropriate values of important settings in practice is urgent. One of the most innovative models for optimisation is genetic algorithm. GA is a method that stochastically optimises functions and was created based on natural genetics and biological evolution mechanics [117]. These models could be utilized to optimise the performance of a predictive model via assigning the most significant values of different settings. GA is an empirical and population-based search approach that is presented by Holland [118]. It is introduced based on the Darwinian natural evolution in biological systems. In GA, a list of potential solutions for an optimisation problem is characterised via what is called a population. The population comprises a finite number of individuals (or called chromosomes). Every individual comprises a list of genes that signifies a point in the search space that sequentially signifies a potential solution.

The flow chart of the GA-based optimisation model as recommended by Huang and Wang [119] and this study has adopted it with some modifications. The model proposed by Huang and Wang is utilized as a reference for describing the optimisation process. The dataset is first split into the training set and the testing set. The GA solution is decoded for every chromosome in the population so that the optimal values for OW, PW, and number of neurons in a hidden layer can be achieved. The algorithm returns the fitness score of the predefined value of 100 if OW, PW, and number of neurons are zero. The selected settings are utilized collectively with the training dataset. Then, the algorithm splits the data based on new optimised settings.

Time-series data are used in this work, so an LSTM model has been designed to train the training data and predict the testing data. An LSTM model involves an input layer with several neurons equal to the new optimised window size. It also contains a hidden layer with several neurons equal to the number of optimised number of neurons. The LSTM model additionally has one output dense layer that provides the predicted values of OW, PW, and number of units in the first hidden layer. Root Mean Square Error (RMSE) score is calculated as a fitness score of GA. DEAP Python Library is used to implement the GA model. After that, some parameters are initialised as Bernoulli random variables. Chromosomes are reordered by shuffling mutation. Then, a roulette wheel is used for selecting algorithms. Finally, the function used for evaluating the fitness of individual solutions is trained.

In an LSTM model, the determination of the optimal number of lags and number of hidden layers is a non-deterministic polynomial (NP) hard problem. GA is a meta-heuristic algorithm, so there is no guarantee that a global optimum solution can be found. Nevertheless, meta-heuristic algorithms tend to have suboptimal good solutions that sometimes can be near the global optimal solutions. GA algorithms have shown in past research that they can be used effectively to find a near-optimal set of time lags [120], [121]. This problem is solved by GA using an evolutionary process inspired by natural selection and genetic science mechanisms.

GA is used in this section to obtain the proposed model's optimum hyperparameters, the size of the observation window, how many neurons are present in the first hidden layer, and the size of the prediction window. This proposed optimisation model is used to find the hyperparameters that reduce the fitness function of the model (i.e., RMSE) [122]. At the beginning of the GA algorithm, the required modules such as *random* module and *scipy.stats* module are imported [123]. Aside from the DEAP module, *base, creator, tools*, and *algorithms* are also imported [124]. In DEAP, a class is created that is inherited from the *deap.base* module. Then, using the weight parameters, the function is maximised. An individual class is defined, which is used to inherit the class list. It also informs the DEAP creator module to assign *FitnessMax* as its fitness attribute [125]. Now, DEAP toolbox is used to define the gene pool and create a population. All the objects used are stored in a container called *toolbox*. However, contents can be added to the container using the *register* method [126].

After creating an *individual* and a *population* by repeatedly utilising the *individual class*, a class is passed to the toolbox for creating a gene of length $N$ [127]. Then, the fitness function is defined, which is returned in DEAP library as a tuple for permitting multi-objective fitness function. Now, the fitness function, crossover operator, mutation operator, and parent selector operator are added to the container. Then, the fitness function defined previously is registered. The crossover operator is also registered. In this proposed algorithm, a *cxOrdered* operator is used. The mutation operator is registered, and the *mutShuffleIndexes* option is selected [128], which shuffles the attributes of the individual input with a probability $indpb = 0.6$. The selection operator is registered, which defines the method that the parents are selected, and the *selRoulette* technique is used.

The hyperparameters are encoded in a binary string with 10 bits for OW size, 5 bits for number of neurons in the first hidden layer, and 10 bits for PW size. Therefore, the complete encoded chromosome involves 25 bits.

## VI. RESULTS AND DISCUSSION

The features chosen by the ''Selecting a Percentile'' feature selection approach in this work were heart rate, systolic blood pressure, mean blood pressure, diastolic blood pressure, respiratory rate, $SpO_2$, glucose, GCS, PTT, PT, age, and admission type. The proposed model's operation can be summarised as the following. Initially, the observation

**TABLE 3. Prediction results for the proposed LSTM-RNN predictive model**

| No. | OW (minutes) | PW (minutes) | Test Processing Time (s) | Test Loss | Accuracy | AUROC |
|---|---|---|---|---|---|---|
| 1 | 60 | 60 | 5.22 | 0.713 | 0.853 | 0.871 |
| 2 | 120 | 60 | 13.17 | 0.661 | 0.869 | 0.881 |
| 3 | 180 | 60 | 9.194 | 0.689 | 0.877 | 0.892 |
| **4** | **240** | **60** | **17.149** | **0.539** | **0.918** | **0.90** |
| 5 | 240 | 120 | 15.91 | 0.556 | 0.884 | 0.891 |
| 6 | 240 | 180 | 16.774 | 0.833 | 0.875 | 0.877 |
| 7 | 240 | 240 | 64.226 | 0.975 | 0.853 | 0.876 |
| 8 | 240 | 300 | 63.531 | 0.621 | 0.852 | 0.874 |
| 9 | 240 | 360 | 15.198 | 0.780 | 0.848 | 0.836 |

window and the prediction window are determined. After that, the number of epochs involved is determined. Furthermore, the batch size is determined. Then, the features are defined and imported. The dataset is then normalised. The dataset is divided afterwards into three datasets, which are training dataset, validation dataset, and testing dataset. The hidden layers are implemented after that. Furthermore, the output layer is defined. Then, the model is compiled to make a prediction. Finally, the performance is calculated.

Table 3 shows the outcomes resulting from metrics that involve different observation window and prediction window sizes. There are four sizes in the observation window: 1 hour, 2 hours, 3 hours, and 4 hours. Previous works like [129, 130] utilized a 24-hour observation window to achieve acceptable prediction performance level (i.e., accuracy and/or AUROC $> 0.80$). This study tends to reduce the observation window from 24 hours to only 4 hours (i.e., a reduction of 83%) while achieving an acceptable prediction accuracy rate. The prediction window involves six sizes with an increment of 1 hour each time starting from 1 hour until 6 hours.

The results show that as the observation window increases, the performance becomes better. Thus, at each time step in the input sequence, the model learns to predict the value of the next time step. It also shows that as the prediction window increases, the performance becomes a little bit worse. The best results were obtained when the observation window was 4 hours, and the prediction window was 1 hour. The results also demonstrate that the proposed model is able to perform with high results up to 6 hours before the onset of ICU patient deterioration. The results of LSTM-RNN showed that best solutions had an accuracy of 0.90 and 0.918 for AUROC, when the OW was 4 hours, the PW was 1 hour, and there were 128 neurons in the first hidden layer.

The proposed model's performance is superior to the SVM model and logistic regression model using the same dataset. The SVM model had an accuracy of 0.821 and the logistic regression model scored an accuracy of 0.792. The proposed model achieved an accuracy of 0.84 when using all 19 features. To get a higher prediction accuracy, the manual feature engineering still need to work with deep learning model for high prediction accuracy.

Table 4 below demonstrates the results that verified the performance difference between the proposed optimal solutions obtained from the proposed optimisation algorithm based on GA and the best solution from the proposed predictive model based on LSTM-RNN. This comparison is based on accuracy and AUROC.

The improvements can be summarised as follows:

1) The accuracy is improved by 0.327%. This small enhancement shows that, based on accuracy, the proposed optimisation algorithm can improve the proposed predictive model's performance in ICU patient deterioration.
2) The AUROC is improved by 3.67%. This good improvement indicates that the proposed GA algorithm can build the proposed deep learning model's performance based on LSTM-RNN.
3) The prediction window is improved by 4.67%. This significant improvement confirms that the proposed models are efficient in their ICU patient deterioration prediction.
4) The observation window is reduced by 40% compared to the size of observation windows used by most of the studies that tend to predict deterioration of patients (i.e., 24 hours).

The quantification of patient health and prediction of future outcomes are vital issues in studies into critical care. The death of patients and their sudden transfer to the ICU are regarded as the most crucial outcomes in ICU admission. The accurate prediction of deaths and sudden ICU transfer might assist in illness severity assessments and as a potential parameter in determining the value of new treatments, interventions, and health care policies. The goal of this study is the accurate prediction of clinical outcomes and benchmarking the performance against several other recent research [112], [113], [131], [132]. The results of the benchmarking against the research mentioned in terms of several criteria like the database source, prediction model, features type (time series or non-time series), ages of patients involved in each work, number of features, prediction task, and the area under the receiver operating curve result are presented in Table 5. Referring to Table 5, it verifies that the proposed models in this study outperform previous works that utilized deep learning approaches to implement predictive models. Johnson *et al.* [113] used an observation window of 24 hours and they demonstrated the lowest AUROC among previous studies at 84%. Pirrachio [131] employed a 24-hour observation window and it achieved the highest AUROC among other works with 88%. Harutyuyan *et al.* [84] achieved an AUROC of 87% despite using an observation window of 48 hours. All these previous studies did not divide the dataset into training, validation, and testing datasets in evaluating performance, but only divided them into two. Harutyuyan *et al.* [84] divided their dataset into training and testing datasets at 85% and 15%, respectively. Such an approach cannot capture the right prediction performance on the unseen testing dataset. The optimized work in this study acquired a significant reduction of the observation window of about 40% and still outperforms the other works.

**TABLE 4. Comparison between the best results obtained from the proposed predictive model based on LSTM-RNN and the optimal solutions given by the proposed optimisation algorithm**

| No. | OW (minutes) | No. of Neurons in the First Hidden Layer | PW (minutes) | Accuracy | AUROC | Change in Accuracy Percentage (%) | Change in AUROC Percentage (%) |
|---|---|---|---|---|---|---|---|
| 1 | 312 | 51 | 41 | 0.840 | 0.762 | -8 | -15.33 |
| 2 | 578 | 51 | 32 | 0.916 | 0.890 | -0.21 | -1.11 |
| 3 | 578 | 128 | 32 | 0.825 | 0.834 | -10.13 | -7.33 |
| 4 | 297 | 5 | 82 | 0.891 | 0.843 | -2.94 | -6.33 |
| **5** | **578** | **56** | **286** | **0.921** | **0.933** | **+0.327** | **+3.67** |
| 6 | 105 | 5 | 80 | 0.800 | 0.831 | -12.85 | -7.67 |

**TABLE 5. Benchmarking results against previous works that used deep learning models**

| | Prediction Model | Sequence Type | Ages Involved | No. of Features | Prediction Task | Splitting | Observation Window | Data Source | Metric | AUROC |
|---|---|---|---|---|---|---|---|---|---|---|
| Purushotham et al. [112] | Gated Recurrent Unit (GRU) | Hour-by-hour | > 15 years old | 17 | Mortality, length of stay, ICD-9 coding | Training and testing datasets | 24 hours | MIMIC-III database | AUROC | 0.86 |
| Pirrachio et al. [131] | Super Learner | Hour-by-hour | > 16 years old | 15 | Mortality | Training and testing datasets | 24 hours | MIMIC-II database | AUROC | 0.88 |
| Harutyuyan et al. [84] | LSTM | Non-time series | > 18 years old | 17 | Mortality, length of stay, decompensation, phenotyping | Training (85%) and testing (15%) datasets | 48 hours | MIMIC-III database | AUROC | 0.87 |
| Johnson et al. [113] | Gradient Boosting (GB) | Non-time series | > 15 years old | 37 | Mortality | Training and testing datasets | 24 hours | MIMIC database | AUROC | 0.8273 |
| Che et al. [80] | GRU | Hour-by-hour | PhysioNet 2012 Challenge dataset | 33 | Mortality and ICD-9 diagnosis categories | Training and testing datasets | 48 hours | MIMIC-III database | AUROC | 0.8461 |
| Optimised work | LSTM | Minute-by-minute | > 15 years old | 19 | Mortality, sudden transfer to ICU | Training (70%), validation (15%), and testing (15%) datasets | 9.6 hours | MIMIC-III database | AUROC | 0.933 |

Table 6 illustrates the results of benchmarking with previous works regarding several criteria, which are the size of observation window, prediction task, inclusion criteria, models used to perform prediction, and AUROC achieved, where the observation window used was 24 hours. Referring to Table 6, it shows that the proposed models in this paper outperform previous works that utilized different linear and non-linear models to implement predictive models. The AUROC of previous works in Table 6 ranged between 0.762 and 0.89. All the previous works in this table utilized an observation window of 24 hours. These works used different prediction tasks. Joshi and Szolovits [138] performed a new unsupervised learning approach, radial domain folding, to predict in-hospital mortality and achieved the highest AUROC with 89%. The results obtained by this study outperform all previous works in Table 6 despite using a shorter observation window. All studies in this table performed one prediction task, while this study performs two prediction tasks (i.e., mortality and sudden transfer to ICU). It can be noticed that all studies use MIMIC open-source database which makes the benchmarking task easier.

Table 7 shows the results of benchmarking with previous works regarding several criteria, where the observation window used was 48 hours or more. The table shows that the AUROC ranged between 0.72 and 0.8602. Johnson *et al.* [145] used a varied range of feature types from the original time-series signals, including standard statistical descriptors such as the minimum, maximum, median, first, last, and the number of values. A new Bayesian ensemble scheme comprising 500 weak learners was used in the proposed model to group the data samples. This model could achieve the highest AUROC among the previous works that used 48 hours or more as the observation window. The table shows that the results achieved by this study outperform other works. It also shows that the observation window was reduced about 80%.

Experiments in this research are performed in a virtual machine provided by Google Colaboratory using an advanced

**TABLE 6. Benchmarking results against previous works that employed observation window of 24 hours**

| Method | Observation Window (hours) | Prediction Task | Inclusion Criteria | Model | Data Source | AUROC |
|---|---|---|---|---|---|---|
| Caballero Barajas and Akella [133] | 24 hours | In-hospital mortality | Age > 18, random fixed size subsample | Generalised linear dynamic model | MIMIC-II database | 0.8657 |
| Celi et al. [134] | 24 hours | In-hospital mortality | ICD-9 code with parameters 430 or 852 | Logistic regression, Bayesian network, artificial neural network | MIMIC database | 0.875 |
| Ghassemi et al.[135] | 24 hours | In-hospital mortality | Age > 18 | Latent Dirichlet Allocation | MIMIC-II database | 0.841 |
| Ghassemi et al.[136] | 24 hours | In-hospital mortality | Age >18 | Multi-task Gaussian process (MTGP) | MIMIC-II database | 0.812 |
| Hoogendoorn et al.[137] | 24 hours | In-hospital mortality | Age > 18, length of stay ≥ 24 hours | Logistic regression, Cox model | MIMIC-II database | 0.841 |
| Joshi and Szolovits [138] | 24 hours | In-hospital mortality | Not NSICU or CSICU, 1st ICU stay, full code, no eventual brain death | Radial domain folding | MIMIC-II database | 0.890 |
| Lee et al.[139] | 24 hours | In-hospital mortality | Patients with full data | Customised severity of illness scores | MIMIC-II database | 0.775 |
| Lehman et al. [140] | 24 hours | In-hospital mortality | Length of stay ≥ 24 hours, 1st ICU stay, have SAPS-I | Hierarchical Dirichlet Processes (HDP) | MIMIC-II database | 0.82 |
| Ripoll et al. [141] | 24 hours | In-hospital mortality | Only septic patients with full data | Quotient Basis Kernel | MIMIC-II database | 0.8223 |
| Hug et al. [142] | 24 hours | Post ICU discharge mortality within 30 days | Not NSICU or CSICU, 1st ICU stay, full code, and no eventual brain death | Logistic regression | MIMIC-II database | 0.8527 |
| Lee et al. [143] | 24 hours | Post hospital discharge mortality within 30 days | ICU stays with complete SAPS data | Cosine-similarity-based patient similarity metric (PSM) | MIMIC-II database | 0.784 |
| Lee and Maslove [139] | 24 hours | Post hospital discharge mortality within 30 days | ICU stays with complete SAPS data | Customised severity of illness scores | MIMIC-II database | 0.762 |
| Lee [144] | 24 hours | Post hospital discharge mortality within 30 days | ICU stays with complete SAPS data | Random forest | MIMIC-II database | 0.815 |
| Ghassemi et al.[136] | 24 hours | Post hospital discharge mortality within a year | Age >18 | MTGP | MIMIC-II database | 0.812 |
| Lee and Maslove [139] | 24 hours | Post hospital discharge mortality within two years | ICU stays with complete SAPS data | Customised severity of illness scores | MIMIC-II database | 0.830 |
| Optimised work | 9.6 hours | Mortality, sudden transfer to ICU | Age > 15 | LSTM based on GA | MIMIC-III database | 0.933 |

GPU of NVIDIA-SMI 440.82, driver version of 418.67, and CUDA version of 10.1. Table 8 shows the results of performing different predictive models utilising various hardware for experiments performed. The processor used here is Intel ®Core $^{TM}$i7-3770 CPU @ 3.40 GHz, 3.90 GHz, and the installed memory is 16.0 GB. The first row in Table 8 (i.e., Gain) compares training time on CPU and GPU using LSTM and GRU predictive models. The execution time of the training uses the default parameters (e.g., epochs = 20 and batch size = 64) as they provide a good baseline in many cases. The time to fit the LSTM model on top of CPU is 432 minutes, the time to fit the GRU model on top of GPU is

**TABLE 7.** Benchmarking results against previous works that employed observation window of 48 hours

| Method | Observation Window (hours) | Prediction Task | Inclusion Criteria | Model | Data Source | AUROC |
|---|---|---|---|---|---|---|
| Caballero Barajas and Akella [133] | 48 hours | In-hospital mortality | Age > 18, random fixed size subsample | Generalised linear dynamic model | MIMIC-II database | 0.7985 |
| Caballero Barajas and Akella [133] | 72 hours | In-hospital mortality | Age > 18, random fixed size subsample | Generalised linear dynamic model | MIMIC-II database | 0.7385 |
| Ding et al. [146] | 48 hours | In-hospital mortality | PhysioNet 2012 Challenge dataset | Just-in-time learning-extreme learning machine (JUST-ELM) | PhysioNet database | 0.8177 |
| Johnson et al. [145] | 48 hours | In-hospital mortality | PhysioNet 2012 Challenge dataset | Bayesian ensemble learning algorithm | MIMIC-II database | 0.8602 |
| Johnson et al. [147] | 48 hours | In-hospital mortality | PhysioNet 2012 Challenge dataset | Regularised logistic regression (RLR), regularised logistic regression with the addition of each covariate squared to the design matrix SVM, random forest | MIMIC-II database | 0.8457 |
| Joshi et al [148] | 48 hours | In-hospital mortality within 30 days | Patients who have stayed in the ICU for at least 48 hours | Constrained non-negative matrix factorisation (CNMF) | MIMIC-III database | 0.72 |
| Optimised work | 9.6 hours | Mortality, sudden transfer to ICU | Age > 15 | LSTM based on GA | MIMIC-III database | 0.933 |

**TABLE 8.** Comparison between different predictive models using various hardware for previous experiments

| CRITERIA | LSTM (CPU) | GRU (GPU) CUDNNGRU | LSTM (GPU) CUDNNLSTM |
|---|---|---|---|
| GAIN (MINUTES) | 432 | 36 | 29 |
| ESTIMATION TIME (S/SAMPLE) | 8 M | 677 µ | 538 µ |
| TESTING PROCESSING TIME (SECONDS) | 9.568 | 2.338 | 1.010 |
| VALIDATION ACCURACY | 0.8804 | 0.8427 | 0.918 |

36 minutes, and the time to fit the LSTM model on top of GPU is 29 minutes. Thus, GPU speedup over CPU of about 14.89. Furthermore, a reduction of 9.47 times in testing processing time between CPU and GPU is recorded. The performance of LSTM model using GPU is better than that of using CPU (i.e., 0.9180 vs. 0.8804).

Generally, the training of LSTM models involves high computational power and utilisation of GPUs, which results from backpropagation through time training model. Moreover, the vanishing and exploding gradient problem must be tackled to accomplish adequate findings. Assigning proper settings and parameters can alleviate these problems. The computational complexity of the proposed LSTM model for the inference requires significantly fewer computing operations in comparison to the training phase. The proposed model utilizes only CPU processing in the Google cloud submission system. Moreover, Purushotham *et al.* [112] used Python implementation of the Super Learner algorithm to predict the in-hospital mortality. This predictive model took about 25 – 30 minutes for evaluating the prediction task using a feature set consists of the 17 features used in the calculation of the SAPS-II score (i.e., feature set A) and it took about

**TABLE 9. Comparison between different predictive models using various hardware units for magnetic properties**

| Method | Model | Prediction Task | No. of Features | Programming Language | Gain | AUROC |
|---|---|---|---|---|---|---|
| Purushotham et al. [112] | Super Learner I | Mortality | 17 | R version | 36 hours | 0.8402 |
| Purushotham et al. [112] | Super Learner I | Mortality | 17 | Python version | 30 minutes | 0.8448 |
| Purushotham et al. [112] | Super Learner II | Mortality | 17 | R version | 28 hours | 0.8646 |
| Purushotham et al. [112] | Super Learner II | Mortality | 17 | Python version | 25 minutes | 0.8701 |
| Purushotham et al. [112] | Deep Learning (FFN) | Mortality | 17 | Python version | 90 minutes | 0.8496 |
| Purushotham et al. [112] | Deep Learning (RNN) | Mortality | 135 | Python version | 1 hour | 0.8544 |
| Purushotham et al. [112] | Deep Learning (Multi Modal Deep Learning (MMDL) | Mortality | 135 | Python version | 1 hour | 0.8664 |
| This work | LSTM | Mortality sudden transfer to ICU | 19 | Python version | 29 minutes | 0.918 |

3 hours for evaluating the prediction task using a feature set consists of 135 raw features (feature set C).

This predictive model took about 25 – 30 minutes for evaluating the prediction task using a feature set consists of the 17 features used in the calculation of the SAPS-II score (i.e., feature set A) and it took about 3 hours for evaluating the prediction task using a feature set consists of 135 raw features (feature set C). A deep Feed forward neural (FFN) network implemented using Keras took around 90 and 100 minutes for evaluating the same mortality task using Feature sets A and C respectively, while the Multi Modal Deep Learning (MMDL) model took around 30 minutes and 1 hour for Feature sets A and C respectively. All the experiments involved in [112] were run on a 32-core Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz machine with NVIDIA TITAN-X GPU processor. Table 9 summarises the benchmarking results between Purushotham *et al.* [112] and this study. The table shows the efficiency of implementing Python in the top of GPU and using LSTM-RNN deep learning prediction model to predict deterioration of patients in ICU since this work outperforms the previous works in terms of computation time and AUROC.

Chen *et al.* [56] proposed a new CNN-based multimodal disease risk prediction algorithm utilising structured and unstructured data from hospital. The study performed experiments on a regional chronic disease of cerebral infraction. It illustrated the running time of CNN- unimodal disease risk prediction (UDRP) using text data (T-data) and CNN-multimodal disease risk prediction (MDRP) using structured and text data (S & T- data) in personal computer (2core CPU, 8.00G RAM) and data center (6core∗2∗7D84core CPU, 48∗7D336G RAM). For CNN-UDRP (T-data) algorithm, the running time in data center is 178.5s while the time in personal computer is 1646.4s. For CNN-MDRP (S&T - data) algorithm, its running time in data center is 178.2s while the time in personal computer is 1637.2s. thus, the running speed of the data center is 9.18 times on the personal computer. On the other hand, the running time for the proposed predictive model based on LSTM-RNN to predict deterioration is 64.39s, thus, it is faster of about 2.77 times than that for CNN-UDRP (T-data) in data center. It is also faster of about 25.57 than that for CNN-UDRP (T- data) in personal computer. Moreover, the running time for the proposed predictive model based on LSTM-RNN to predict deterioration is faster than that for CNN-MDRP (S&T- data) of about 2.76 times in data center. Besides, it is also faster of about 25.56 than that for CNN-UDRP (T- data) in personal computer. There are a significant difference in favor of the proposed model performed by this study in the running time compared to personal computer as well as the data center performed by Chen *et al.* [56] despite the slightly difference in accuracy between the 2 studies. Table 9 shows the efficiency of the proposed model based on LSTM-RNN in terms of running time over the CNN model.

Tables demonstrate that open-source data, such as MIMIC database in all its versions, makes the benchmarking task easier. The tables show that the results obtained in this study outperform studies that used linear and non-linear models. The tables also confirm that open-source datasets public dissemination is essential in the facilitation of iterative improvement in predictive models. Furthermore, the tables show that some studies [134], [149] focused on specific patient groups, while others required clinical notes [135], [140]. Moreover, the tables show wide inter-study heterogeneity in inclusion criteria, model adopted, and performance. In fact, the AUROC obtained in this study outperforms previous works that used different sizes of observation windows and different predictive models, either linear or non-linear.

## VII. CONCLUSION

In the quest to improve the performance of ICU patient deterioration prediction which uses dynamic and static data, a novel and reliable prediction framework is proposed in this study. The framework consists mainly of a predictive

model based on deep learning (i.e., LSTM-RNN) and an optimisation model based on GA. These models are used to predict deterioration before its onset.

The proposed predictive model overcomes the shortcomings that exist in current prediction processes of research based on deterioration of patients in ICU. Deep learning model based on LSTM that uses dynamic and static data implemented on GPU and an optimisation model based on GA are proposed to achieve the research objectives. Different benchmarking metrics are adopted by this research, which are sequence type, inclusion criteria, number of features, prediction task, splitting ratios, observation window size, data source, AUROC, model type, gain, estimation time, testing processing time, and validation accuracy.

The observation window used by the proposed predictive model is minimised during the training, validation, and testing steps. Different advanced techniques are proposed to achieve better performance and a reduction in the observation window. These techniques are normalisation, learning rate, dropout, and early stopping. In fact, computing hardware advances in the last decade, particularly GPUs, have enabled larger, deeper networks to be trained. These more sophisticated networks have demonstrated remarkable success in wide ranging applications such as prediction of deterioration for patients in ICU. Furthermore, the experimental work illustrates that using GPU can reduce gain (execution time), estimation time, and testing processing time. The validation accuracy has also been shown to be better compared with the LSTM model implemented using a CPU or another predictive model (i.e., GRU) using a GPU. This research also proposed an optimisation algorithm based on GA-based multi-objective optimisation algorithm. In the proposed algorithm, the data is trained using the LSTM model. Optimisation of the observation window size, the number of hidden units in the first hidden layer, and the prediction window size are carried out to enhance the accuracy and AUROC and lower the proposed predictive model's test loss. Chart event features are considerably sensitive to time series among the data utilized in this research, and they cannot be properly obtained by conventional machine learning models (e.g., logistic regression and/or Support Vector Machine). An LSTM-RNN deep learning framework for ICU patient deterioration prediction is proposed in this research, as it can incorporate time-series data properly with no information loss. Moreover, advances in learning algorithms (i.e., machine learning and deep learning) are being driven by three technical forces which are hardware, datasets and benchmarks, and algorithms advances.

## REFERENCES

[1] H. Wunsch, C. Guerra, A. E. Barnato, D. C. Angus, G. Li, and W. T. Linde-Zwirble, "Three-year outcomes for Medicare beneficiaries who survive intensive care," *Jama*, vol. 303, no. 9, pp. 849–856, 2010.

[2] E. E. Vasilevskis, M. W. Kuzniewicz, M. L. Dean, T. Clay, E. Vittinghoff, D. J. Rennie, and R. A. Dudley, "Relationship between discharge practices and intensive care unit in-hospital mortality performance: Evidence of a discharge bias," *Med. Care*, vol. 47, no. 7, pp. 803–812, Jul. 2009.

[3] W. B. Hall, L. E. Willis, S. Medvedev, and S. S. Carson, "The implications of long-term acute care hospital transfer practices for measures of in-hospital mortality and length of stay," *Amer. J. Respiratory Crit. Care Med.*, vol. 185, no. 1, pp. 53–57, Jan. 2012.

[4] B. Wellner, J. Grand, E. Canzone, M. Coarr, P. W. Brady, J. Simmons, E. Kirkendall, N. Dean, M. Kleinman, and P. Sylvester, "Predicting unplanned transfers to the intensive care unit: A machine learning approach leveraging diverse clinical elements," *JMIR Med. Informat.*, vol. 5, no. 4, p. e45, Nov. 2017.

[5] D. Kolte, S. Khera, W. S. Aronow, C. Palaniswamy, M. Mujib, C. Ahn, S. Iwai, D. Jain, S. Sule, A. Ahmed, and H. A. Cooper, "Regional variation in the incidence and outcomes of in-hospital cardiac arrest in the United States," *Circulation*, vol. 131, no. 16, pp. 1415–1425, 2015.

[6] M. Spångfors, L. Arvidsson, V. Karlsson, and K. Samuelson, "The national early warning score: Translation, testing and prediction in a Swedish setting," *Intensive Crit. Care Nursing*, vol. 37, pp. 62–67, Dec. 2016.

[7] D. R. Levinson and I. General, "Adverse events in hospitals: National incidence among Medicare beneficiaries," Dept. Health Hum. Services Office Inspector Gen., Washington, DC, USA, Tech. Rep., 2010.

[8] P. Davis, R. Lay-Yee, R. Briant, W. Ali, A. Scott, and S. Schug, "Adverse events in New Zealand public hospitals II: Preventability and clinical context," *New Zealand Med. J.*, vol. 116, no. 1183, pp. 1–12, 2003.

[9] C. Vincent, G. Neale, and M. Woloshynowych, "Adverse events in British hospitals: Preliminary retrospective record review," *BMJ*, vol. 322, no. 7285, pp. 517–519, Mar. 2001.

[10] G. R. Baker, "The Canadian adverse events study: The incidence of adverse events among hospital patients in Canada," *Can. Med. Assoc. J.*, vol. 170, no. 11, pp. 1678–1686, May 2004.

[11] S. B. Hu, D. J. L. Wong, A. Correa, N. Li, and J. C. Deng, "Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model," *PLoS ONE*, vol. 11, no. 8, Aug. 2016, Art. no. e0161401.

[12] M. M. Churpek, T. C. Yuen, S. Y. Park, R. Gibbons, and D. P. Edelson, "Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards," *Crit. Care Med.*, vol. 42, no. 4, p. 841, 2014.

[13] M. M. Churpek, T. C. Yuen, and D. P. Edelson, "Predicting clinical deterioration in the hospital: The impact of outcome selection," *Resuscitation*, vol. 84, no. 5, pp. 564–568, May 2013.

[14] G. B. Smith, D. R. Prytherch, P. Meredith, P. E. Schmidt, and P. I. Featherstone, "The ability of the national early warning score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death," *Resuscitation*, vol. 84, no. 4, pp. 465–470, 2013.

[15] K. Mochizuki, R. Shintani, K. Mori, T. Sato, O. Sakaguchi, K. Takeshige, K. Nitta, and H. Imamura, "Importance of respiratory rate for the prediction of clinical deterioration after emergency department discharge: A single-center, case-control study," *Acute Med. Surgery*, vol. 4, no. 2, pp. 172–178, Apr. 2017.

[16] V. M. Quinten, M. van Meurs, T. J. Olgers, J. M. Vonk, J. J. M. Ligtenberg, and J. C. T. Maaten, "Repeated vital sign measurements in the emergency department predict patient deterioration within 72 hours: A prospective observational study," *Scandin. J. Trauma, Resuscitation Emergency Med.*, vol. 26, no. 1, p. 57, Dec. 2018.

[17] C. P. Bonafide, A. R. Localio, L. Song, K. E. Roberts, V. M. Nadkarni, M. Priestley, C. W. Paine, M. Zander, M. Lutts, P. W. Brady, and R. Keren, "Cost-benefit analysis of a medical emergency team in a children's hospital," *Pediatrics*, vol. 134, no. 2, pp. 235–241, 2014.

[18] D. P. Henriksen, M. Brabrand, and A. T. Lassen, "Prognosis and risk factors for deterioration in patients admitted to a medical emergency department," *PLoS ONE*, vol. 9, no. 4, Apr. 2014, Art. no. e94649.

[19] H. Zheng and D. Shi, "Using a LSTM-RNN based deep learning framework for ICU mortality prediction," in *Proc. Int. Conf. Web Inf. Syst. Appl.* Springer, 2018, pp. 60–67.

[20] T. Bonnici, L. Tarassenko, D. A. Clifton, and P. Watkinson, "The digital patient," *Clin. Med.*, vol. 13, no. 3, p. 252, 2013.

[21] National Confidential Enquiry into Patient Outcome and Death, Findley, "Knowing the risk: A review of the peri-operative care of surgical patients: Summary," in *Proc. NCEPOD*, 2011.

[22] National Confidential Enquiry into Patient Outcome and Death and Stewart, "Adding insult to injury: A review of the care of patients who died in hospital with a primary diagnosis of acute kidney injury (acute renal failure): A report by the national confidential enquiry into patient outcome and death," in *Proc. NCEPOD*, 2009.

[23] G. Findlay, "Time to intervene? A review of patients who underwent cardiopulmonary resuscitation as a result of an in-hospital cardiorespiratory arrest. A report by the national confidential enquiry into patient outcome and death," Tech. Rep., 2012.

[24] M. Santamaria Ariza, I. Zambon, H. S. Sousa, J. A. Campos e Matos, and A. Strauss, "Comparison of forecasting models to predict concrete bridge decks performance," *Struct. Concrete*, vol. 21, no. 4, pp. 1240–1253, Aug. 2020.

[25] Y. Abebe and S. Tesfamariam, "Storm sewer pipe renewal planning considering deterioration, climate change, and urbanization: A dynamic Bayesian network and GIS framework," *Sustain. Resilient Infrastruct.*, pp. 1–16, Mar. 2020.

[26] D. P. Edelson, K. Carey, C. J. Winslow, and M. M. Churpek, "Less is more: Detecting clinical deterioration in the hospital with machine learning using only age, heart rate and respiratory rate," in *Proc. C15. Crit. CARE, BIG DATA Artif. Intell. Crit. ILLNESS*. American Thoracic Society, 2018, p. A4444.

[27] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson, "Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards," *Crit. Care Med.*, vol. 44, no. 2, p. 368, 2016.

[28] D. R. Goldhill and A. Sumner, "Outcome of intensive care patients in a group of British intensive care units," *Crit. Care Med.*, vol. 26, no. 8, pp. 1337–1345, Aug. 1998.

[29] J. S. Lundberg, T. M. Perl, T. Wiblin, M. D. Costigan, J. Dawson, M. D. Nettleman, and R. P. Wenzel, "Septic shock: An analysis of outcomes for patients with onset on hospital wards versus intensive care units," *Crit. Care Med.*, vol. 26, no. 6, pp. 1020–1024, Jun. 1998.

[30] F. J. R. Catling and A. H. Wolff, "Temporal convolutional networks allow early prediction of events in critical care," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 3, pp. 355–365, Mar. 2020.

[31] C. Stergiou, K. E. Psannis, B.-G. Kim, and B. Gupta, "Secure integration of IoT and cloud computing," *Future Gener. Comput. Syst.*, vol. 78, pp. 964–975, Jan. 2018.

[32] F. Douglis *et al.*, "Unified Web hosting and content distribution," 2018.

[33] J. Zhang and D. Centola, "Social networks and health: New developments in diffusion, online and offline," *Annu. Rev. Sociol.*, vol. 45, no. 1, pp. 91–109, Jul. 2019.

[34] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings Bioinf.*, vol. 18, no. 5, pp. 851–869, 2017.

[35] I.W.I.B Data, "Bring big data to the enterprise," Tech. Rep., 2012.

[36] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[37] Y. Wang, L. Kung, and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technol. Forecasting Social Change*, vol. 126, pp. 3–13, Jan. 2018.

[38] Y. Mao, W. Chen, Y. Chen, C. Lu, M. Kollef, and T. Bailey, "An integrated data mining approach to real-time clinical monitoring and deterioration warning," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2012, pp. 1140–1148.

[39] J. E. Johnson, R. Blanes, D. Sheng, and A. Narayanan, "Face recognition for fast information retrieval and record lookup," Tech. Rep., 2019.

[40] Z. Batmaz, A. Yurekli, A. Bilge, and C. Kaleli, "A review on deep learning for recommender systems: Challenges and remedies," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 1–37, Jun. 2019.

[41] F. Nargesian, H. Samulowitz, U. Khurana, E. B. Khalil, and D. S. Turaga, "Learning feature engineering for classification," in *Proc. IJCAI*, 2017, pp. 2529–2535.

[42] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, Feb. 2018, Art. no. 7068349.

[43] N. Kumar, V. N. Lolla, E. Keogh, S. Lonardi, C. A. Ratanamahatana, and L. Wei, "Time-series bitmaps: A practical visualization tool for working with large time series databases," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2005, pp. 531–535.

[44] M. Längkvist, L. Karlsson, and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," *Pattern Recognit. Lett.*, vol. 42, pp. 11–24, Jun. 2014.

[45] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[46] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," 2017, *arXiv:1708.02182*. [Online]. Available: http://arxiv.org/abs/1708.02182

[47] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "HybridSN: Exploring 3-D–2-D CNN feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, Feb. 2020.

[48] H. Ramchoun, M. A. Idrissi, Y. Ghanou, and M. Ettaouil, "Multilayer perceptron: Architecture optimization and training," *IJIMAI*, vol. 4, no. 1, pp. 26–30, 2016.

[49] J. Adnan, N. N. Daud, A. S. Mokhtar, F. R. Hashim, S. Ahmad, A. F. Rashidi, and Z. I. Rizman, "Multilayer perceptron based activation function on heart abnormality activity," *J. Fundam. Appl. Sci.*, vol. 9, no. 3, pp. 417–432, 2017.

[50] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. S. Tan, "A deep convolutional neural network model to classify heartbeats," *Comput. Biol. Med.*, vol. 89, pp. 389–396, Oct. 2017.

[51] J. Tobias Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014, *arXiv:1412.6806*. [Online]. Available: http://arxiv.org/abs/1412.6806

[52] M. Rafiq, G. Keel, P. Mazzocato, J. Spaak, C. Savage, and C. Guttmann, "Deep learning architectures for vector representations of patients and exploring predictors of 30-day hospital readmissions in patients with multiple chronic conditions," in *Proc. Int. Workshop Artif. Intell. Health*. Springer, 2018, pp. 228–244.

[53] S. Ghosh, P. Chakraborty, E. Cohn, J. S. Brownstein, and N. Ramakrishnan, "Characterizing diseases from unstructured text: A vocabulary driven Word2vec approach," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 1129–1138.

[54] N. Wickramasinghe, "Deepr: A convolutional net for medical records," Tech. Rep., 2017.

[55] L. Brand, A. Patel, I. Singh, and C. Brand, "Real time mortality risk prediction: A convolutional neural network approach," in *Proc. HEALTHINF*, 2018, pp. 463–470.

[56] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[57] A. Yigit and Z. Isik, "Applying deep learning models to structural MRI for stage prediction of Alzheimer's disease," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, pp. 196–210, Jan. 2020.

[58] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice-Hal, 2007.

[59] Y. Bengio, *Learning Deep Architectures for AI*. Now Publishers, 2009.

[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.

[61] C. Dong, C. Change Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[62] F. Zhang, N. Cai, J. Wu, G. Cen, H. Wang, and X. Chen, "Image denoising method based on a deep convolution neural network," *IET Image Process.*, vol. 12, no. 4, pp. 485–493, Apr. 2018.

[63] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[64] M. Aczon, D. Ledbetter, L. Ho, A. Gunny, A. Flynn, J. Williams, and R. Wetzel, "Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks," 2017, *arXiv:1701.06675*. [Online]. Available: http://arxiv.org/abs/1701.06675

[65] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4945–4949.

[66] X. Wang, S. Takaki, and J. Yamagishi, "A simple RNN-plus-highway network for statistical parametric speech synthesis," Nat. Inst. Inform., Tokyo, Japan, Tech. Rep. NII-2017-003E, 2017.

[67] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (GRU) neural networks," in *Proc. IEEE 60th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2017, pp. 1597–1600.
[68] Y. Zhang, G. Chen, D. Yu, K. Yao, S. Khudanpur, and J. Glass, "Highway long short-term memory RNNS for distant speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5755–5759.
[69] Y. Dai, C. Wang, J. Dong, and C. Sun, "Visual relationship detection based on bidirectional recurrent neural network," *Multimedia Tools Appl.*, vol. 79, pp. 35297–35313, May 2019.
[70] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," 2014, *arXiv:1409.2329*. [Online]. Available: http://arxiv.org/abs/1409.2329
[71] S. Zhang, M. Abdel-Aty, Y. Wu, and O. Zheng, "Modeling pedestrians' near-accident events at signalized intersections using gated recurrent unit (GRU)," *Accident Anal. Prevention*, vol. 148, Dec. 2020, Art. no. 105844.
[72] A. N. Shewalkar, "Comparison of RNN, LSTM and GRU on speech recognition data," Tech. Rep., 2018.
[73] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.
[74] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[75] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ. San Diego La Jolla Inst. Cogn. Sci., Tech. Rep., 1985.
[76] J. L. Elman, "Finding structure in time," *Cognit. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990.
[77] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," Tech. Rep., 1999.
[78] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proc. IEEE-INNS-ENNS Int. Joint Conf. Neural Networks. IJCNN. Neural Comput., New Challenges Perspect. New Millennium*, Jul. 2000, pp. 189–194.
[79] N. K. Manaswi, "Deep learning with applications using Python: Chatbots and face, object, and speech recognition with tensorflow and Keras," Tech. Rep., 2018.
[80] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, Dec. 2018.
[81] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Deepcare: A deep dynamic memory model for predictive medicine," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Springer, 2016, pp. 30–41.
[82] Y. Zhang, Y. Lin, M. Chi, J. Ivy, M. Capan, and J. M. Huddleston, "LSTM for septic shock: Adding unreliable labels to reliable predictions," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 1233–1242.
[83] C. Lin, Y. Zhang, J. Ivy, M. Capan, R. Arnold, J. M. Huddleston, and M. Chi, "Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM," in *Proc. IEEE Int. Conf. Healthcare Inform. (ICHI)*, Jun. 2018, pp. 219–228.
[84] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. V. Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," 2017, *arXiv:1703.07771*. [Online]. Available: http://arxiv.org/abs/1703.07771
[85] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, Jul. 2014.
[86] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients," *Crit. Care Med.*, vol. 34, no. 5, pp. 1297–1310, 2006.
[87] B. Williams, G. Albert, C. Ball, D. Bell, R. Binks, L. Durham, J. Eddleston, N. Edwards, D. Evans, and M. Jones, "National early warning score (NEWS): Standardizing the assessment of acute illness severity in the NHS," Roy. College Physicians, London, U.K., Tech. Rep., 2012.
[88] D. Dahl, G. G. Wojtal, M. J. Breslow, D. Huguez, D. Stone, and G. Korpi, "The high cost of low-acuity ICU outliers," *J. Healthcare Manage.*, vol. 57, no. 6, pp. 421–433, 2012.
[89] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," 2015, *arXiv:1511.03677*. [Online]. Available: http://arxiv.org/abs/1511.03677
[90] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, no. 1, 2016, Art. no. 160035.
[91] Y. W. Lin, Y. Zhou, F. Faghri, M. J. Shaw, and R. H. Campbell, "Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory," *PloS ONE*, vol. 14, no. 7, 2019, Art. no. e0218942.
[92] K. Junwei, H. Yang, L. Junjiang, and Y. Zhijun, "Dynamic prediction of cardiovascular disease using improved LSTM," *Int. J. Crowd Sci.*, vol. 3, no. 1, pp. 14–25, May 2019.
[93] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 1, pp. 198–208, 2017.
[94] J. Plate, F. Hietbrink, L. P. H. Leenen, M. C. J. Eijkemans, and L. M. Peelen, "Predicting clinical deterioration at the intermediate care unit: Comparing a joint modelling approach with a long short-term recurrent neural network," in *Optimizing Care for the Critically Ill Surgical Patient: The Role of the IMCU*. 2018, p. 207.
[95] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Healthcare Conf.*, 2016, pp. 301–318.
[96] T. Peiffer, J. Ruyssinck, J. Decruyenaere, F. D. Turck, F. Ongenae, and T. Dhaene, "Early detection of positive blood cultures using recurrent neural networks on time series data," in *Proc. 25th Belgian-Dutch Conf. Mach. Learn. (BeneLearn)*, 2016, pp. 1–3.
[97] Q. Pan, S. Wang, and J. Zhang, "Prediction of Alzheimer's disease based on bidirectional LSTM," *J. Phys., Conf. Ser.*, vol. 1187, no. 5, 2019, Art. no. 052030.
[98] B. Holt, *Writing and Querying MapReduce Views in CouchDB: Tools for Data Analysts*. Newton, MA, USA: O'Reilly Media, Inc, 2011.
[99] M. Ross, W. Wei, and L. Ohno-Machado, "'Big data' and the electronic health record," *Yearbook Med. Inform.*, vol. 23, no. 1, pp. 97–104, 2014.
[100] A. E. W. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "The MIMIC code repository: Enabling reproducibility in critical care research," *J. Amer. Med. Inform. Assoc.*, vol. 25, no. 1, pp. 32–39, Jan. 2018.
[101] J. D. Plate, L. M. Peelen, L. P. Leenen, and F. Hietbrink, "Validation of the VitalPAC early warning score at the intermediate care unit," *World J. Crit. Care Med.*, vol. 7, no. 3, p. 39, 2018.
[102] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database," *Crit. Care Med.*, vol. 39, no. 5, p. 952, 2011.
[103] B. K. Hua, "An optimization method based on genetic algorithm for heart rate variability analysis in the prediction of the onset of cardiac arrhythmia," Tech. Rep., 2017.
[104] K. Ng, S. R. Steinhubl C. D. Filippi, S. Dey, and W. F. Stewart, "Early detection of heart failure using electronic health records: Practical implications for time before diagnosis, data diversity, data quantity, and data density," *Circulat., Cardiovascular Qual. Outcomes*, vol. 9, no. 6, pp. 649–658, 2016.
[105] J. Reyes-Garcia, H. Galeana-Zapien, A. Galaviz-Mosqueda, and C. Torres-Huitzil, "Evaluation of the impact of data uncertainty on the prediction of physiological patient deterioration," *IEEE Access*, vol. 6, pp. 38595–38606, 2018.
[106] D. Chicco, "Ten quick tips for machine learning in computational biology," *BioData Mining*, vol. 10, no. 1, p. 35, Dec. 2017.
[107] M. Tan, C. D. Santos, B. Xiang, and B. Zhou, "LSTM-based deep learning models for non-factoid answer selection," 2015, *arXiv:1511.04108*. [Online]. Available: http://arxiv.org/abs/1511.04108
[108] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," 2019, *arXiv:1909.09586*. [Online]. Available: http://arxiv.org/abs/1909.09586
[109] V. Kyurkchiev and N. Kyurkchiev, "A family of recurrence generated functions based on the 'half-hyperbolic tangent activation function,'" *Biomed. Statist. Inform.*, vol. 2, no. 3, pp. 87–94, 2017.
[110] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An Introduction to Statistical Learning: With Applications in R*. Spinger, 2013.
[111] S. Russell and P. Norvig, "Artificial intelligence: A modern approach," Tech. Rep., 2002.

[112] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmark of deep learning models on large healthcare MIMIC datasets," 2017, *arXiv:1710.08531*. [Online]. Available: http://arxiv.org/abs/1710.08531
[113] A. E. Johnson, T. J. Pollard, and R. G. Mark, "Reproducibility in critical care: A mortality prediction case study," in *Proc. Mach. Learn. Healthcare Conf.*, 2017, pp. 361–376.
[114] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for ICU outcome prediction," in *Proc. AMIA Annual Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2016, p. 371.
[115] N. AlNuaimi, M. M Masud, and F. Mohammed, "ICU patient deterioration prediction: A data-mining approach," 2015, *arXiv:1511.06910*. [Online]. Available: http://arxiv.org/abs/1511.06910
[116] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 11, pp. 2825–2830, 2011.
[117] H. Chiroma, S. Abdulkareem, A. Abubakar, and T. Herawan, "Neural networks optimization through genetic algorithm searches: A review," *Appl. Math. Inf. Sci.*, vol. 11, no. 6, pp. 1543–1564, 2017.
[118] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis With Applications to Biology, Control, and Artificial Intelligence*. Cambridge, MA, USA: MIT Press, 1992.
[119] C.-L. Huang and C.-J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Syst. Appl.*, vol. 31, no. 2, pp. 231–240, Aug. 2006.
[120] K. Lukoseviciute and M. Ragulskis, "Evolutionary algorithms for the selection of time lags for time series forecasting by fuzzy inference systems," *Neurocomputing*, vol. 73, nos. 10–12, pp. 2077–2088, Jun. 2010.
[121] Z.-L. Sun, D.-S. Huang, C.-H. Zheng, and L. Shang, "Optimal selection of time lags for TDSEP based on genetic algorithm," *Neurocomputing*, vol. 69, nos. 7–9, pp. 884–887, Mar. 2006.
[122] D. Chhachhiya, A. Sharma, and M. Gupta, "Designing optimal architecture of recurrent neural network (LSTM) with particle swarm optimization technique specifically for educational dataset," *Int. J. Inf. Technol.*, vol. 11, no. 1, pp. 159–163, 2019.
[123] V. Porcu, "SciPy and NumPy," in *Python for Data Mining Quick Syntax Reference*. Springer, 2018, pp. 177–200.
[124] S. Aeeni, "Development of an open-source multi-objective optimization toolbox," Tech. Rep., 2019.
[125] Z. Wang, Y. Sun, X. Yang, and S. Li, "Hybrid optimisation method of improved genetic algorithm and IFT for linear thinned array," *J. Eng.*, vol. 2019, no. 20, pp. 6457–6460, Oct. 2019.
[126] N. Andrade, F. A. Faria, and F. A. M. Cappabianco, "A practical review on medical image registration: From rigid to deep learning based approaches," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, Oct. 2018, pp. 463–470.
[127] J. H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, and J. Brink, "Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success," *J. Amer. College Radiol.*, vol. 15, no. 3, pp. 504–508, Mar. 2018.
[128] X. Wang and R. Miikkulainen, "MDEA: Malware detection with evolutionary adversarial learning," 2020, *arXiv:2002.03331*. [Online]. Available: http://arxiv.org/abs/2002.03331
[129] C. Potes, B. Conroy, M. Xu-Wilson, C. Newth, D. Inwald, and J. Frassica, "A clinical prediction model to identify patients at high risk of hemodynamic instability in the pediatric intensive care unit," *Crit. Care*, vol. 21, no. 1, p. 282, Dec. 2017.
[130] J.-R. L. Gall, S. Lemeshow, and F. Saulnier, "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study," *Jama*, vol. 270, no. 24, pp. 2957–2963, 1993.
[131] R. Pirracchio, "Mortality prediction in the ICU based on mimic-ii results from the super ICU learner algorithm (SICULA) project," in *Secondary Analysis of Electronic Health Records*. Springer, 2016, pp. 295–313.
[132] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. V. Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Sci. Data*, vol. 6, no. 1, p. 96, Dec. 2019.
[133] K. L. C. Barajas and R. Akella, "Dynamically modeling Patient's health state from electronic medical records: A time series approach," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 69–78.
[134] L. A. Celi, S. Galvin, G. Davidzon, J. Lee, D. Scott, and R. Mark, "A database-driven decision support system: Customized mortality prediction," *J. Personalized Med.*, vol. 2, no. 4, pp. 138–148, Sep. 2012.
[135] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits, "Unfolding physiological state: Mortality modelling in intensive care units," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2014, pp. 75–84.
[136] M. Ghassemi, M. A. Pimentel, T. Naumann, T. Brennan, D. A. Clifton, P. Szolovits, and M. Feng, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, p. 446.
[137] M. Hoogendoorn, A. El Hassouni, K. Mok, M. Ghassemi, and P. Szolovits, "Prediction using patient comparison vs. Modeling: A case study for mortality prediction," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 2464–2467.
[138] R. Joshi and P. Szolovits, "Prognostic physiology: Modeling patient severity in intensive care units using radial domain folding," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2012, p. 1276.
[139] J. Lee and D. M. Maslove, "Customization of a severity of illness score using local electronic medical record data," *J. Intensive Care Med.*, vol. 32, no. 1, pp. 38–47, 2017.
[140] L. W. Lehman, M. Saeed, W. Long, J. Lee, and R. Mark, "Risk stratification of ICU patients using topic models inferred from unstructured progress notes," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2012, p. 505.
[141] V. J. R. Ripoll, A. Vellido, E. Romero, and J. C. Ruiz-Rodríguez, "Sepsis mortality prediction with the quotient basis kernel," *Artif. Intell. Med.*, vol. 61, no. 1, pp. 45–52, May 2014.
[142] C. W. Hug and P. Szolovits, "ICU acuity: Real-time models versus daily models," in *Proc. AMIA Annu. Symp.* Bethesda, MD, USA: American Medical Informatics Association, 2009, p. 260.
[143] J. Lee, D. M. Maslove, and J. A. Dubin, "Personalized mortality prediction driven by electronic medical data and a patient similarity metric," *PLoS ONE*, vol. 10, no. 5, May 2015, Art. no. e0127428.
[144] J. Lee, "Patient-specific predictive modeling using random forests: An observational study for the critically ill," *JMIR Med. Informat.*, vol. 5, no. 1, p. e3, Jan. 2017.
[145] A. E. Johnson, N. Dunkley, L. Mayaud, A. Tsanas, A. A. Kramer, and G. D. Clifford, "Patient specific predictions in the intensive care unit using a Bayesian ensemble," in *Proc. Comput. Cardiol.*, Sep. 2012, pp. 249–252.
[146] Y. Ding, X. Li, and Y. Wang, "Mortality prediction for ICU patients using just-in-time learning and extreme learning machine," in *Proc. 12th World Congr. Intell. Control Autom. (WCICA)*, Jun. 2016, pp. 939–944.
[147] A. E. Johnson, A. A. Kramer, and G. D. Clifford, "Data preprocessing and mortality prediction: The Physionet/CinC 2012 challenge revisited," in *Proc. Comput. Cardiol.*, Sep. 2014, pp. 157–160.
[148] S. Joshi, S. Gunasekar, D. Sontag, and J. Ghosh, "Identifiable phenotyping using constrained non-negative matrix factorization," 2016, *arXiv:1608.00704*. [Online]. Available: http://arxiv.org/abs/1608.00704
[149] J. Calvert, Q. Mao, J. L. Hoffman, M. Jay, T. Desautels, H. Mohamadlou, U. Chettipally, and R. Das, "Using electronic health record collected clinical variables to predict medical intensive care unit mortality," *Ann. Med. Surg.*, vol. 11, pp. 52–57, Nov. 2016.

**TARIQ I. ALSHWAHEEN** received the B.S. degree in biomedical engineering from The Hashemite University, Zarqa, Jordan, in 2003, and the M.S. degree in embedded engineering systems from Al-Yarmouk University, Irbid, Jordan, in 2017. He is currently pursuing the Ph.D. degree in biomedical engineering with Universiti Teknologi Malaysia. From 2003 to 2015, he was a Maintenance Engineer and a Lecturer with the Medical Device Technology Institute, Jordanian Royal Medical Services. His research interest includes the prediction of deterioration of intensive care units' patients using deep learning.

**YUAN WEN HAU** received the master's and Ph.D. degrees in electrical engineering from Universiti Teknologi Malaysia in 2005 and 2009, respectively. She is currently a Research Fellow and the Head of cardiac informatics cluster with the UTM-IJN Cardiovascular Engineering Centre, Universiti Teknologi Malaysia, where she is also a Senior Lecturer with the School of Biomedical Engineering and Health Sciences. Her research interests are in the design of multi-processor systems-on-chip (MPSoC) embedded system architecture for biomedical instrumentation, health care applications, cryptographic and data security applications, ECG signal processing, arrhythmia detection and prediction using machine learning, vital sign monitor design, telemedicine or telecardiology development, and ADHD/autism assessment. She has authored or coauthored more than 50 articles in peer-reviewed international journals and conferences and several book chapters.

**NIZAR ASS'AD** is currently pursuing the Ph.D. degree with The University of Newcastle, Australia. He is considered one of the cancer care specialists in the Great Sydney area. He is also one of the leaders in education and training (a consultant). He is currently a Clinical Supervisor. His main interests are centered on topics related to patients in intensive care unit especially their sudden adverse events.

**MAHMOUD M. ABUALSAMEN** is currently pursuing the degree with the Department of Family and Community Medicine, Faculty of Medicine, The University of Jordan, Amman, Jordan. He is currently a Researcher in public health with The University of Jordan. He is a clinical pharmacist by training and his research interests focus on data-driven health interventions for refugees in Jordan.