Artificial Intelligence Review (2022) 55:1055–1084
https://doi.org/10.1007/s10462-021-09982-2

Use of learning approaches to predict clinical deterioration in patients based on various variables: a review of the literature

Tariq Ibrahim Al-Shwaheen[1] · Mehrdad Moghbel[2] · Yuan Wen Hau[1] · Chia Yee Ooi[2]

Published online: 13 March 2021

Abstract
Machine learning can be considered as the current gold standard for predicting deterioration in Intensive Care Unit patients and is in extensive use throughout the world in different fields. As confirmed by many studies, preventing the occurrence of the onset of deterioration in a sufficient time window is a priority in healthcare centers. Also, the significance of enhancing the quality of hospital care and the reduction of adverse outcomes is of great importance. Notably, it is hypothesized that by exploiting recent technologies, models built upon dynamic variables (e.g. vital signs, lab tests, and demographic variables) could reinforce the predictive ability of models aimed at detection of in clinical deterioration with high accuracy, sensitivity and specificity. This manuscript summarises the techniques and approaches proposed in the literature for predicting deterioration and compares the performance and limitations of various approaches grouped based on their application. While several approaches can attain promising results, there is still room for additional improvement, especially in pre-processing and modeling enhancement steps where most methods do not take the necessary steps for ensuring a high-performance result. In this manuscript, the most effective machine learning models, as well as deep learning models, for predicting deterioration of patients are discussed in hopes of assisting the readers with ascertaining the best possible solutions for this problem.

Keywords Deep learning · Deterioration · Early Warning Score systems · Machine learning and prediction

1 Introduction

Several variables can be associated with the deterioration of a patient's health (Hu et al. 2016; Churpek et al. 2014, 2013; Smith et al. 2013), where there is a need to transfer the patient to an ICU or a Coronary Care Unit (CCU) (Quinten et al. 2018) due to factors such as liver or kidney injury and respiratory failure. In other instances, the patient might need to revisit the Emergency Department (ED) or be transferred to other specialized hospitals

✉ Tariq Ibrahim Al-Shwaheen
bluetareqqq@yahoo.com

Extended author information available on the last page of the article

for emergency surgical treatment (Mochizuki et al. 2017). Some studies have identified various variables resulting in the transfer of patients from a general ward to ICU, such as positive pressure ventilation, vasopressors fluid resuscitation, or any other immediate procedure two hours before or twelve hours after the transfer to represent the patient's deterioration (Wellner et al. 2017; Bonafide et al. 2014). Other studies used a deviation from an outlined standard treatment procedure within 2–24 h of arrival at a hospital (Henriksen et al. 2014) or hospital readmission within a 30-day window as variables associated with the deterioration of the patient's health (Mochizuki et al. 2017; Wellner et al. 2017). While there is a broad set of definitions on what constitutes deterioration, studies have shown the value of these variables in determining and reducing the risk of mortality (Quinten et al. 2018; Mochizuki et al. 2017).

Early Warning Score (EWS) systems are a commonly used approach for estimating the patient's deterioration and enabling the administration of pre-emptive treatments by providing early warnings (Hu et al. 2016; Quinten et al. 2018; Kivipuro et al. 2018; Singer et al. 2016; Panday et al. 2017) using variables associated with the deterioration of patient's health. EWS systems work by assigning a score to various measurements based on a predefined range and using the combination of these scores as a threshold for providing early warnings to the medical team. It should be noted that monitoring the patient's deterioration is performed using Track and Trigger (T&T) systems (Smith et al. 2008) in place of EWS in some hospitals. Similar to EWS, T&T systems depend on periodic measurements of vital signs (tracking) with a prior action (triggering) when a specified threshold is reached. Both EWS and T&T systems are based on interchangeable concepts and can be considered to have the same weaknesses and strengths (Liaw et al. 2011; Grant 2018).

A study by Prytherch et al. (2006) showed that some EWS systems used in hospitals rely on conventional pen and paper for charting scores for various measurements which not only might result in reduced accuracy; it can also reduce the speed at which predictions are determined. While this pen and paper approach is still in use in many modern health care facilities, many have moved on to a continuous monitoring solution for detecting various measures related to a patient's health with some monitoring measures extended to the daily life of patients such as wrist wearable heart rate monitors (Tilly et al. 1995). However, equipment costs prohibit the hospitals from providing continuous monitoring solutions to all patients, while the pen and paper approach consumes the time of physicians and nurses. Additionally, EWS and T&T systems often suffer from unnecessarily high alarm rates, which can cause alarm fatigue, which reduces the probability of a timely response to alarms (Gao et al. 2006; Paradiso 2003; Rothman et al. 2013). With the recent increases in data storage capacity and increased amounts of patient data gathered from various monitoring solutions, many studies have focused on developing more accurate and reliable automated prediction algorithms for EWS and T&T systems (Churpek et al. 2014; Edelson et al. 2018; Newman 2017; Schmid et al. 2013; Li and Clifford 2012; Scalzo et al. 2012; Liu et al. 2012; Mardini et al. 2012; Young et al. 2003; Mokart et al. 2013). Prediction algorithms mainly depend on a sufficient observation window and a sufficient prediction window where there is a trade-off between the accuracy and the window size. In the observation window, independent variables (predictors) are generated for a specific period, whereas the dependent variable comes from the prediction window (Wiley and Pace 2015). Figure 1 illustrates the observation window and the prediction window with respect to time.

The rest of the manuscript is organized as follows: Sect. 2 highlights different performance measures used in literature. Section 3 illustrates Medical Information Mart for Intensive Care (MIMIC) dataset and its different releases. Section 4 mentions some

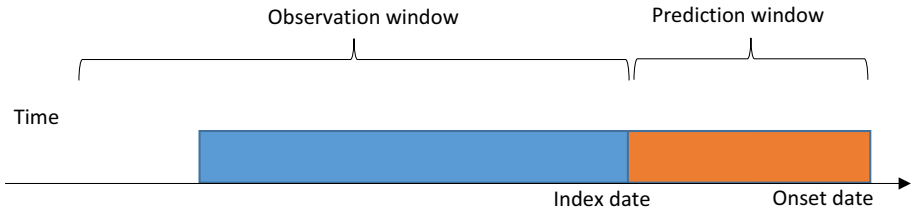

Fig. 1 The observation window and the prediction window with respect to time

EWS systems. Section 5 discusses different learning models in the prediction of patients' deterioration such as systems designed for monitoring signs of cardiac arrest, systems designed for monitoring predicting transfer of patients to ICUs or pediatric ICUs as well as unplanned readmissions. Furthermore, various systems designed for monitoring predictors, which can be associated with the deterioration of the patient's health, will also be discussed. Section 6 begins by discussing steps followed to implement a predictive model. The study then describes the data that various studies have collected and the set of clinical events that affects the proposed predictive models. Also, this section illustrates some challenges in deep learning time series classifications and widely adopted architectures. Conclusions are drawn in the last section.

2 Performance measures

Despite the utilization of different sources of datasets within the studies in literature, which trained, validated, and tested the proposed deterioration prediction models, the performance had to be compared and benchmarked with well-established deterioration modeling techniques that have been extensively used to predict. Therefore, this section describes the different metrics used in binary classification, multi-class classification, and regression in order to validate the proposed models and prove its capability of accurately predicting the deterioration of patients. Consequently, Clinical prediction models are commonly evaluated using a number of measures that quantify the model's calibration and discrimination. However, utilising these performance metrics provides some difficulties when one seeks to predict exceedingly rare outcomes (Saito and Rehmsmeier 2015). The most crucial of these is that a classifier that attempts to maximize the accuracy of its classification rule when predicting a rare outcome might achieve an accuracy of 99% simply by classifying all observations as non-events, and model enhancements (e.g. as quantified by increases in the c statistic) are overshadowed by the large true negative rate (Kipnis et al. 2016).

The performance of classification methods is commonly evaluated using their predictions about the class of the input data and the ground truth classes. Often, the performance of the classifier is assessed using various metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy using a confusion matrix. A confusion matrix representing the outcome of a binary classifier (normal/abnormal) is presented in Fig. 2. True positive (TP) indicates the number of correctly predicted abnormal classes and true negative (TN) shows the number of correctly predicted normal classes. The incorrect prediction of abnormal classes is represented as false negative (FN), and the incorrect prediction of normal objects is

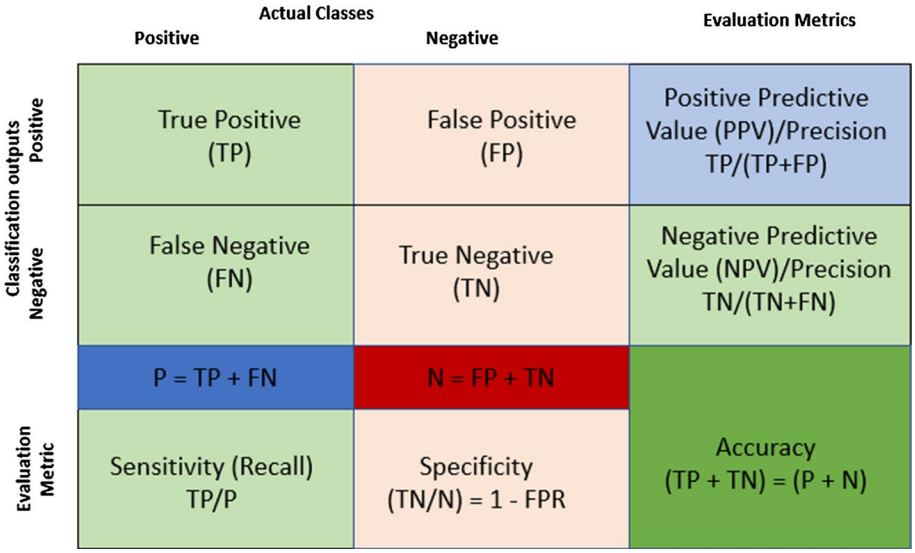

Fig. 2 Confusion matrix with evaluation metrics derived from it

presented using false positive (FP) values. Receiver operating characteristics (ROC) and area under receiver operating characteristics (AUROC) are also amongst the most popular measures for evaluating the performance of various machine learning methods. Receiver operative characteristics (ROC) is a two-dimensional graph for visualization, organization, and selection of different classifiers based on their performance. The graph axis represents relative tradeoffs between benefits (true positives) plotted on the y-axis and costs (false positives) plotted on the x-axis (Fawcett 2006) with AUROC signifying the degree or measure of separability (efficiency of discriminating between classes). Higher the AUROC, the better the model is at predicting actual classes.

F-score, a measure that combines precision and recall, is the harmonic mean of precision and recall with the traditional F-score calculated using the following formula:

$$F \times score = 2 * \frac{Precision \cdot Recall}{Precision + Recall} \tag{1}$$

The Likelihood Ratio (LR) is the likelihood that a particular test result would be anticipated in a patient with the target disorder matched to the likelihood that the same result would be anticipated in a patient without the target disorder. This performance measure has advantages over sensitivity and specificity because it is less likely to vary with the dominance of the disorder, may be considered for different levels of the symptom/sign or test, may be utilized to combine the results of several diagnostic tests and can be utilized to calculate post-test probability for a target disorder (Taylor and Creelman 1967). An LR greater than 1 results in a post-test probability that is higher than the pre-test probability, while an LR of less than 1 results in a post-test probability that is lower than the pre-test probability. If the pre-test probability lies between 30 and 70%, then test results with a very high LR rule in disease; a very low LR (say, below 0.1) virtually rules out the probability that the patient has the disease.

In Multi-class classification, it is important to consider the response variable y and the prediction variable ý as two discrete random variables. These variables presume values in range 1 to K and each number signifies a different class. The algorithm comes up with the probability that a certain unit adopts one potential class, then a classification rule is utilized to assign a single class to each individual. The rule is usually very plain andthe most widespread rule assigns a unit to the class with the highest probability. Performance metrics are very valuable when the purpose is to calculate and compare different classification models. Balanced Accuracy is a well-known metric both in binary and in multi-class classification. It gives the same weight for each class and its insensibility to class distribution helps to spot possible predictive problems also for rare and underrepresented classes. Moreover, Balanced Accuracy Weighted can be a good performance indicator when the objective is to train a classification algorithm on a wide number of classes. This metric lets to keep separate algorithm performances on the various classes, so that it allows to trace which class causes poor performance (Grandini et al. 2020).

When it comes to multi-class cases, F1-Score has to comprise all the classes. Thus, it is required a multi-class measure of Precision and Recall being included into the harmonic mean. Such metrics might have two various specifications, giving rise to two different metrics: Micro F1-Score and Macro F1-Score. To achieve Macro F1-Score, it is required to compute Macro-Precision and Macro-Recall before. They are respectively calculated by taking the average precision for each predicted class and the average recall for each actual class. Therefore, the macro approach implies all the classes as basic elements of the calculation where each class has the same weight in the average, so that there is no discrepancy between highly and poorly populated classes. On the other hand, to acquire Micro F1-Score, it is required to calculate Micro-Precision and Micro-Recall before. The idea of Micro-averaging is to consider all the units together, without taking into consideration possible differences between classes Grandini et al. (2020).

Model evaluation is very important in data science. It helps to understand the performance of the regression model and makes it easy for presentation. In fact, there are several evaluation metrics out there but only some of them are appropriate to be utilised for regression. R Square measures how much of variability in dependent variable could be clarified by a model. It is square of Correlation Coefficient (R) and that is why it is called R Square. This metric is calculated by the sum of squared of prediction error divided by the total sum of square which replace the calculated prediction with mean. R Square value ranges between 0 to 1 and bigger value implies a better fit between prediction and actual value (Legates and McCabe Jr 1999). R Square is measured using the following formula:

$$R^2 = 1 - \frac{\sum_i \left(y_i - \hat{yi}\right)^2}{\sum_i \left(y_i - \bar{yi}\right)^2} \tag{2}$$

where y_i is the real output and $\hat{y}i$ is the predicted output. Another metric used to measure the performance of regression is the Mean Square Error (MSE) which is an absolute measure of the goodness for the fit. It is calculated by the sum of square of prediction error which is real output minus predicted output and then divided by the number of data points. It gives an absolute number on how much the predicted results deviate from the actual number. It cannot clarify much insights from one single result but it supplies a real number to compare against other model results and help to select the best regression model (Nicolson and Paliwal 2019). MSE is calculated using the following formula:

$$MSE = \frac{1}{N} - \sum \left(y_i - \hat{yi}\right)^2 \tag{3}$$

where N is the number of points. Another metric and utilized more than MSE to measure regression is the Root Mean Square Error (RMSE) which is the square root of MSE. It is utilised more frequently than MSE because firstly sometimes MSE value could be too big to compare simply. Secondly, MSE is calculated by the square of error, and hence square root brings it back to the same level of prediction error and make it easier for interpretation (Wang and Lu 2018). Mean Absolute Error (MAE) is similar to MSE. Though, rather than the sum of square of error in MSE, MAE is taking the sum of absolute value of error (Qi et al. 2020). This metric is calculated using the following formula:

$$MAE = \frac{1}{N} - \sum \left|y_i - \hat{yi}\right| \tag{4}$$

In fact, MAE is a more direct representation of sum of error terms. Whereas MSE gives larger penalisation to big prediction error by square it while MAE treats all errors the same. MSE, RMSE or MAE are better to be utilised to compare performance between different regression models than R Square. However, it makes total sense to utilise MSE if value is not too big and MAE if there is a tendance to penalize large prediction error. Moreover, fireworks algorithm (FWA) is a meta-heuristic method and extensively utilised in continuous (i.e. regression). The key points of FWA are to define a proper neighborhood structure for introducing the local search procedure and to explore a metric for quantifying the disparity between solutions. It can process linear, non-linear, and multi-model test functions and is suitable to implement in parallel. Most important of all, FWA has a good convergence property and can always find the global optimal solutions (Liu et al. 2015).

Some works use statistical metrics to evaluate performance of the proposed models such as c-static value. Wellner et al. (2017) utilize this metric where it reveals the superiority of different measurements. Authors use this metric to demonstrate the time-dependent changes in c-static values (Fukushima et al. 2011). Another statistical metric is the likelihood which measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters. It is calculated by dividing the number of events by the number of possible outcomes which will generate the probability of a single event occurring (Pawitan 2001).

3 MIMIC dataset

Medical Information Mart for Intensive Care (MIMIC) (Johnson et al. 2016a, b, c) is a relational dataset of patients who stayed at critical care units at a large medical center in Boston, Massachusetts, USA. The MIMIC dataset, periodically updated with new cases added and mistakes amended, contains information about each patient starting from admission to hospital until discharge or death (in-hospital or in some cases out-hospital) along with laboratory test results and possible admittances to several medical care units such as Medical Intensive Care Units (MICU), Surgical Intensive Care Units (SICU), Coronary Care Unit (CCU), Cardiac Surgery Recovery Unit (CSRU) and Trauma Surgical Intensive Care Unit (TSCU) (Johnson et al. 2016a, b, c). There are three main releases of this open-source dataset known as MIMIC, MIMIC-II, and MIMIC-III (Johnson et al. 2018) with the latest version of MIMIC III (version 1.4) released on 2 September 2016 which includes the data of patients who were admitted between 2001 and 2012. The MIMIC-I (Saeed et al. 2011)

Dataset contains data recorded from over 90 ICU patients containing signals and periodic measurements gained from a bedside monitor and clinical data gained from the patient's medical record. It is the first attempt (1992–1999) to build a collection of multi-parameter recordings of ICU patients. The recordings differ in length; virtually all of them are at least 20 h, and many are 40 h or more. In all, the dataset covers approximately 200 patient-days of real-time signals and accompanying data. Each record normally comprises of a number of hundred individual files. The data acquired from the bedside monitors are divided into files each involving 10 min of recorded signals, which can then be assembled without gaps to form a continuous recording. For this reason, each record is kept in a separate directory, named after the record, it contains. About the only advantage of MIMIC-I over MIMIC II is that the ECG signals in MIMIC-I were recorded at 500 samples per second with 12-bit precision and negligible jitter, whereas those in MIMIC II contain 125 "peak-picked" samples per second with 8- or 10-bit precision and $\pm$ 6 ms jitter.

All data within MIMIC-I was later incorporated into the MIMIC-II and covering the period 2001–2008 (Saeed et al. 2011) dataset with the patients de-identified in compliance with Health Insurance Portability and Accountability Act (HIPAA) standards to enable public access to the dataset. MIMIC-II contains a diverse and substantial population of intensive care unit patient stays and contains comprehensive and detailed clinical data, comprising physiological waveforms and a minute-by-minute subset of records. A highly utilized version of the MIMIC-II Clinical Dataset is the v2.6, released in April 2011. It contains 32,536 subjects (with 40, 426 ICU admissions) admitted to medical, surgical, cardiovascular, and neonatal ICUs, surgical recovery units, and coronary care units at a single tertiary care hospital including approximately 7000 neonates.

MIMIC-III (Komorowski et al. 2018) is an extension of MIMIC-II by adding more data collected between 2008 and 2012. It contains 53,423 hospital admissions for adult patients and 8,100 neonates. The median age of adult patients in the dataset is 65.8 years, with 44.1% of patients being female. A mean of 4579 charted observations and about 380 laboratory measurements are obtainable with about 30 vital signs recorded once a minute with an amplitude resolution of up to 16 bits for every hospital admission (Johnson et al. 2017a, b). Also, many elements of the dataset have been regenerated from the raw data in a more robust manner to ensure the quality of the data. Additionally, the MIMIC-III dataset includes admission and discharge dates along with the location and the initial diagnosis on admission.

There are many several updates and changes between the different releases of MIMIC database. In MIMIC-III dataset, many data elements have been regenerated from the raw data in the previous releases in a more robust manner to improve the quality of the underlying data. The original Philips CareVue system (which acquired data from 2001 to 2008) was replaced with the new Metavision data management system (which continues to be utilized to the present). In MIMIC-II there were multiple tables including the same column name, ITEMID, but referring to various concepts. In attempt to alleviate confusion, MIMIC-III has merged all these tables into a single table. The ITEMID for laboratory measurements in the D_LABITEMS and LABEVENTS tables in MIMIC-II do not match the ITEMID for laboratory measurements in MIMIC-III. Hence, a mapping table was provided to facilitate the updating of queries which utilise this table. Moreover, ADMISSIONS table in MIMIC-III is sourced from the hospital database, rather than the ICU database in MIMIC-II where admission and discharge dates have the time components. The CENSUSEVENTS table was utilised in MIMIC-II to track patient hospital admissions was replaced by TRANSFERS table in MIMIC-III, hence providing greater granularity and easier tracking of a patient's hospital course. DEMOGRAPHIC_DETAIL table in

MIMIC-II was merged into ADMISSIONS table in MIMIC-III. In fact, most of the studies in this work used data extracted from an open source that can be easily benchmarked and generalising the results achieved to overcome the problem of generalisation due to data from hospitals specialized in certain diseases, or patients with certain diseases. Table 1 summarises the differences and improvements among MIMIC-II and MIMIC-III datasets.

4 EWS systems

Morgan, Williams, and Wright (Morgan et al. 1997) proposed the first EWS system based on six physiological parameters of the patient's vital signs, namely heart rate, respiratory rate, systolic blood pressure, temperature, consciousness level and saturation of oxygen in the blood for detecting early signs of deterioration. Since then, many other and more expanded EWS systems have been proposed with the Modified Early Warning Score (MEWS) (Panday et al. 2017), the VitalPAC Early Warning Score (ViEWS) (Plate et al. 2018), and the National Early Warning Score (NEWS) (Williams et al. 2012) being the most prominent. NEWS, introduced in 2012 and updated in 2017, is the most common system which has been endorsed by the National Health Service (NHS) of England and is the approved EWS system for use in hospitals in England and was adopted by 70% of hospitals by 2015 (Hogan et al. 2019). It should be noted that each EWS system has its own definition of what a pre-defined range constitutes as normal. In Table 2 are shown examples of these ranges for Respiratory Rate (RR), Saturation of Oxygen in Blood (SpO_2), Systolic Blood Pressure (SBP) (i.e. the pressure created by the beating of the heart muscle), Heart Rate (HR), Consciousness Level and Temperature (Temp). Considerable differences in normal ranges used in EWS systems, especially in respiratory rate, SpO_2, and SBP, can be observed.

As is with other fields related to medical data analysis (Kononenko 2001; Lavrač 1999), machine learning-based automatic analysis of patient data has become a common approach in EWS systems with Support Vector Machine (SVM) (Kate et al. 2016; Mao et al. 2012), Logistic Regression (Quinten et al. 2018; Kate et al. 2016; Mao et al. 2012; Spångfors et al. 2016; Churpek et al. 2016; Zhai et al. 2014; Kipnis et al. 2016), Decision Tree (AlNuaimi et al. 2015), Naive Bayes (NB) (Kate et al. 2016; Masud and Al Harahshen 2016), and Neural Network (NN) (Hu et al. 2016; Wellner et al. 2017) based methods being the most prominent. With the ability to identify intricate non-linear pattserns in the data, machine learning-based EWS and T&T systems aim to improve the predictive accuracy and reduce false alarm rate (Hu et al. 2016; Manning et al. 2014; LeCun et al. 2015; Zheng et al. 2016) with these systems shown to improve the survival rate of high-risk patients (Stanzani and Lewis 2018).

5 Learning models in the prediction of patients' deterioration

Patients in either general-surgical wards or ICUs suffer from adverse events such as sudden transfer to ICU (Wellner et al. 2017), cardiac arrests (Byrd et al. 2014) or even mortality (Rajkomar et al. 2018; Taylor et al. 2016; Hoogendoorn et al. 2016) that can show evidence of physiologic derangement prior to their deterioration. Despite the disagreement of a specific definition of deterioration, increasing availability of EHRs, and freely accessible critical care databases, predictive models that use machine learning and deep learning are

Table 1 Differences and Improvements among MIMIC-II and MIMIC-III datasets

MIMIC-II	MIMIC-III	Improvement
Data collected between 2001 and 2008	Augments data in MIMIC-III with newly collected data between 2008 and 2012	Increasing the volume of data to cover more cases and patients Regenerating many data elements from the raw data in a more robust manner to improve the quality of the underlying data
Used the original Philips CareVue system	Replaced with the new Metavision data management system (which continues to be used to the present)	Adding new data
Multiple tables containing the same column name, ITEMID, but referring to different concepts		Merging (D_CHARTITEMS, D_IOITEMS, D_MEDITEMS) tables into a single table (i.e. D_ITEMS) and deleting D_CODEDITEMS table and D_PARAM-MAP_ITEMS table
ADMISSIONS is sourced from the ICU database	ADMISSIONS is sourced from the hospital database	Presenting time component of admission and discharge dates, discharge location, diagnosis on admission, ED registration, and exit time
The CENSUSEVENTS table was used in MIMIC-II to track patient hospital admissions	This table has been removed and replaced with the TRANSFERS table	Tracking the admission, discharge, transfer (ADT) data of a patient throughout the entire hospital stay, supplying greater granularity and easier tracking of a patient's hospital course The ADT data offering information regarding ward location The ADT data has fewer erroneous admissions: frequently the ICU database involved erroneous entries based on accidental admission/discharges Providing the ADT data for all patients in the ICU database
The DEMOGRAPHIC_DETAIL table provided extra static information regarding a patient which rarely changed throughout an admission	DEMOGRAPHIC_DETAIL table merged into ADMISSIONS table	Implementing the new ADMISSION table from the hospital database which contains the same set of demographics
DRGEVENTS table	It has been renamed DRGCODES	Improving the clarity of the data
ICD9 table	It has been renamed to DIAGNOSES_ICD	Clarifying the content of the table

Table 1 (continued)

MIMIC-II	MIMIC-III	Improvement
IOEVENTS and MEDEVENTS tables	Data in the IOEVENTS and MEDEVENTS tables is involved in the OUTPUTEVENTS, INPUTEVENTS_CV and INPUTEVENTS_MV tables	Consolidating these tables to ease querying for drug deliveries
POE_MED and POE_ORDER tables	They have been merged into a single table named PRESCRIPTIONS	Clarifying the content of these tables
HADM_ID	HADM_ID have been regenerated	Differentiating the newly generated HADM_ID that range from 100,000 to 199,000 from other IDs
ICUSTAY_ID	ICUSTAY_ID have been regenerated	Preventing confusion of the newly generated ICUSTAY_ID that range from 200,000 to 299,000 with other IDs
–	A new added CALLOUT table	Providing data both on when the patient was considered ready for discharge and when the patient really left the ICU
–	A new added PROCEDURES_ICD table	Offering ICD-9 codes for procedures in the PROCEDURES_ICD table
–	New added INPUTEVENTS_CV table and INPUTEVENTS_MV table	Comprising two different monitoring systems that were operating in the hospital over the data collection period
–	A new added OUTPUTEVENTS Table2	Recording data about outputs in a consistent fashion for the Metavision and CareVue databases
COMORBIDITY_SCORES DEMOGRAPHICEVENTS, D_DEMOGRAPHICITEMS D_CAREUNITS D_CODEDITEMS D_PARAMMAP_ITEMS ICUSTAY_DETAILS PARAMETER_MAPPING WAVEFORM_*, D_WAVEFORM_SIGNALS	These tables have been removed in MIMIC-III	Providing much more efficient in terms of data transfer Clarifying that these data are not “raw” in that they are not acquired directly from the databases but rather synthesized views of this data

Table 2 Normal ranges employed in some ews systems

EWS	RR	SpO_2	SBP	HR	Consciousness	Temp (in degrees centigrade)
Morgan et al. (1997)	9–20	>92	100–199	50–99	Alert	36 to 37.9
Panday et al. (2017)	9–17	≥93	101–159	51–100	Alert	36.05 to 38
The national EWS (NEWS) Williams et al. (2012)	12–20	≥96	111–219	51–90	Alert	36 to 37.9

becoming feasible and motivated. This study tries to explore predictive models that aim to predict deterioration of patients using learning algorithms. Thus, learning algorithms proposed in the literature can be (widely) categorized into three categories. The first category includes systems designed for monitoring in-hospital patients for signs of cardiac arrest that could lead to sudden clinical deterioration. The second category contains systems designed for predicting and facilitating the unplanned transfer of patients to ICUs or PICUs as well as unplanned readmissions. The third category contains general systems designed for monitoring variables that can be associated with the deterioration of a patient's health and have no specific aim other than keeping the patient's condition under observation.

Predictive models using electronic health record (EHR) data have rapidly advanced recently. While model performance metrics have improved considerably, best practices for implementing predictive models into clinical settings for point-of-care risk stratification are still evolving. Here, we conducted a review of articles describing predictive models integrated into EHR systems and implemented to predict deterioration of patients using learning algorithms. We limited our review to peer-reviewed journal articles published in English with available full text. Our primary eligibility criteria focused on identifying articles with the following: description of a model predicting deterioration (e.g., not financial outcomes), use of EHR data for modeling, automated data extraction for modeling (e.g., not manual data entry or manual data calculation by providers), integration of the model into the EHR system, and use of learning algorithm to implement predictive models. This last criterion was critical given the emphasis of our review on implementation. We did not restrict studies to specific types of models—for instance, models using support vector machines, Baysian network models, Hidden Markov Models, and various neural network architectures were all eligible. However, our definition of "model" did require that there be some sort of mathematical calculation involving predictors based on EHR data. Therefore, this study tries to find works categorized based on the three categories adopted by this study based on the aforementioned aspects.

5.1 Systems designed for monitoring signs of cardiac arrest

Ordonez et al. (2016) proposed the use of a K-nearest neighbor (kNN) model to predict the possibility of patients developing hypotension within the next hour utilizing the patient's heart rate as their variable. The training data contained 58 patients from the 2009 PhysioNet Challenge (Moody and Lehman 2009), of which 28 experience an episode of hypotension in the hour following the data capture. In their proposed model, the heart rate time-series data was converted into a sequenced symbolic representation through the Symbolic

Aggregate Approximation (SAX) method based on a heuristically determined window size, symbols, alphabet and used as the input for the kNN mode with the expectation–maximization (EM) algorithm used for optimizing the parameters. Their proposed method was able to achieve an accuracy of 0.85, PPV of 0.82, the sensitivity of 0.87, and an F-score of 0.85. It is demonstrated that the advantages of k-NN are twofold. First, it is easy to implement as in the training phase the only step is storing the training set and their class labels. Second, it is flexible to handle diverse data via utilising particular distance metrics. Though, it is computationally intensive when training set is very large because it must recognize the k nearest neighbor for every testing data (Colque 2018).

Lee and Mark (2010a) used the ANN algorithm to predict hypotensive events (defined as a drop in systolic pressure below 100 mm Hg, or diastolic pressure below 65 mm Hg) utilizing the patient's vital signs, HR, SBP, diastolic blood pressure (DBP) (i.e. the pressure exerted by the heart muscle at rest), and Mean Blood Pressure (MBP) as variables. They have utilized the publicly available MIMIC-II dataset (Saeed et al. 2002) for designing and validating their proposed method. Depending on the label assigned to each example (control or hypotensive), a binary ANN model with a log-sigmoid activation function with 20 hidden neurons was trained to predict hypotensive events. Their proposed method was able to achieve an AUROC of 0.934. Lee and Mark (2010b) later expanded their proposed ANN model to predict hypotensive events with Mean Arterial Pressure (MAP), HR, pulse pressure and relative cardiac output as variables with the pulse pressure derived by subtracting diastolic blood pressure from systolic blood pressure and the relative cardiac output derived by multiplying pulse pressure by the heart rate value. Furthermore, the time-series data embedded in the MIMIC-II dataset was reorganized into records, each of which corresponded to a particular ICU stay. The study also used age and medication information from the MIMIC-II dataset as well as hemodynamic data series and clinical data. Each sample consisted of three-time intervals, which were an observation window of either 30 or 60 min, a target window of 1 h and a 1- or 2-h gap interval between the windows. The study utilized a minute-by-minute time series for heart rate, systolic blood pressure, diastolic blood pressure, and mean arterial blood pressure with a total of 102 features extracted from the observation window for every sample. Data applied to the observation window was used to input source for various pattern classifiers and predictions were performed at the end of the observation window with every target window labeled as either control or hypotensive. Also, independent neural networks were trained for several gaps and observation window sizes, as well as several cross-validation folds and compilation modes. An AUROC of 0.918 was achieved by their proposed method, evaluated using five-fold cross-validation.

The next two studies were verified to explore the importance of extracting data from a single medical center rather than MIMIC-II database. Lee et al. (2016a) insisted to use the same architecture as before. Thus, Lee et al. proposed an Artificial Neural Network (ANN) based method for predicting the possibility of Ventricular Tachycardia (VT) occurring within the next hour utilizing the patient's Heart Rate Variability (HRV) and the RR variability (RRV) as variables. However, variations within heart rhythms and respiratory rhythms are labeled as HRV and RRV, respectively. The ANN model was generated using 14 parameters obtained from HRV and RRV analysis with the ANN model containing 13 hidden neurons in the hidden layer. The dataset was gathered from patients admitted to the cardiovascular intensive care unit (CICU) at Asan Medical Center between September 2013 and April 2015. The dataset consisted of 52 recordings acquired one hour before ventricular events and 52 control readings. The model had an AUROC of 0.93. The study showed the possible contribution of combining HRV and RRV in detecting ventricular tachycardia

one hour before its onset as the sole use of HRV did not lead to acceptable performance. While the study is promising, the limited number of samples (due to limited access to ventricular tachycardia patients) is the main drawback of the study. Ong et al. (2012) utilized an SVM based method to predict the possibility of cardiac arrest occurring within a 72-h window using variables like HRV, age, sex, medical history, heart rate, BP, SpO_2, RR, and Glasgow Coma Scale (GCS). The study categorized patients into low, intermediate and high-risk groups according to the prediction provided by the SVM. The study was carried out in a single healthcare center with the results compared to the Modified Early Warning System (MEWS) (Subbe et al. 2001) used in the health center. Their proposed method was able to achieve an AUROC of 0.781 with a sensitivity of 0.814 and specificity of 0.723, substantially higher than the performance of the MEWS system with an AUROC of 0.680, sensitivity of 0.744 and specificity of 0.542. While the study shows that several other variables such as vital signs and age could be used alongside HRV to forecast the chances of cardiac arrest and mortality, their dataset is the limiting factor as they have gathered data from emergency department patients in a single-center with the patient conditions not being followed after their examination in the emergency department. Additionally, patients from various diagnosis groupings were not separated, which might decrease the accuracy as the variability in the heart rate of non-cardiovascular patients might differ from cardiovascular patients. To conclude, as two different algorithms, SVM and ANN share the same concept utilising linear learning model for pattern recognition. The difference is primarily on how non-linear data is classified. Fundamentally, SVM employs nonlinear mapping to produce the data linear separable, therefore the kernel function is the key. Though, ANN utilizes multi-layer connection as well as several activation functions to deal with nonlinear issues. Furthermore, single layer ANN can only create linear boundary, and the 2nd layer can combine the linear boundary together; while at least three layers are needed to generate boundary of arbitrary shapes. Subsequently, the training results from SVM have better generalization capability than those from ANN. Thus, SVM and ANN are two typical classifiers which are utilised to validate balanced learning strategy (Ren 2012).

Chen et al. (2017) utilized a random forest classification model to predict Cardio-Respiratory Insufficiency (CRI), a term referring to the symptoms associated with the loss of normal cardio-respiratory reserve, which is often life-threatening. CRI frequently arises in hospitalized patients, but its risk factors have not been studied in detail. Their study aimed at defining the risk factors along with a monitoring system for providing early warning for CRI events. Using standard bedside monitors, they recorded heart rate, respiratory rate, SpO_2, systolic blood pressure and diastolic blood pressure of 1,880 unique patients (1971 admissions) admitted to an adult surgical trauma step-down unit at an urban teaching hospital in Pittsburgh, USA for a window of 4 h before the CRI event. Validated using tenfold cross-validation, their proposed method was able to achieve an AUROC of 0.94. More work is crucial to integrate insights into a completely operational predictive algorithm and to assess it in a prospective framework. The model was restricted to noninvasive vital sign data; using different data types such as lab measurements as well as demographic data might further enhance prediction accuracy.

Donald et al. (2012) used the Bayesian Artificial Neural Network (BANN) to predict hypotensive events utilizing the Edinburgh University Secondary Insult Grades (EUSIG) definitions for hypotension (systolic arterial pressure $<$ 90 mmHg OR mean arterial pressure $<$ 70 mmHg) using systolic and mean arterial pressure, heart rate, age, and gender. The study collected around 2000 events from 22 hospitals in Europe by analyzing the Brain-IT dataset utilizing 15-min sub-windows starting at 15 and 30 min before an event. Initial results from the clinical study illustrate a model sensitivity of 0.4095 and specificity

of 0.8646. Despite the low performance, it can be considered clinically useful, provided the false positives remain low as to be practical in an intensive care environment. It was demonstrated that ANN and random forest can be used to implement predictive models. It was also found that different vital signs are very important to predict deterioration of patients especially when it is associated with the heart like cardiac arrest and ventricular tachycardia.

5.2 Systems designed for monitoring predicting transfer of patients to icus or picus as well as unplanned readmissions

Wickramasinghe et al. Wickramasinghe (2017) proposed a Convolutional Neural Network (CNN) and logistic regression-based deep learning method called Deepr (short for Deep record) for predicting unplanned readmissions after the discharge of hospitalized patients. Based on concepts used in natural language processing, their method converts 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 utilizing a validation dataset containing 300,000 patient records divided into three subsets based on the unplanned readmission of patients within different periods. Measures on three and six-month unplanned readmission prediction following a random index discharge with and without time-gaps show that the proposed model obtained AUROC of 0.795 and 0.800 for the three months for without and with time, respectively. The model obtained AUROC of 0.809 and 0.819 for without time and with time for the six months, respectively.

To signify the importance of merging different types of data. The next two studies use vital signs, lab tests to implement the predictive models. These studies were also used the same model to predict deterioration of patients which is the neural networks. Wellner et al. (2017) proposed a Feed-Forward Neural Network (FFNN) based ANN for predicting the unplanned transfer of patients to ICU six hours before the transfer. Their dataset consisted of patient information gathered from 3 children's hospitals, along with notes from the medical team. However, they did not follow any standardization procedures, which resulted in noticeable variations in their dataset, such as using three different EHR systems to collect data and differences in vocabulary between medical team assessments. Additionally, their study suffered from overfitting as they utilized 4000 variables, which included vital signs, lab tests, nurse notes and acuity, which constitutes the need for a feature selection step in their proposed method. Validated using c-static values across three hospitals, their proposed method was deemed to have an acceptable performance with c-static values for the three hospitals being 0.892, 0.902, and 0.899.

Hu et al. (2016) proposed an ANN with 24 hidden neurons to predict clinical deterioration (i.e., unplanned transfer to ICU or cardiac arrest) utilizing a dataset containing 522 normal and 43 abnormal cases along with EMRs collected from various centers in United States with 15 variables used including vital signs and standard lab tests. While the study was aimed at providing a warning to the medical team 4 h before the onset of the patient deterioration, the developed ANN was able to provide predictions 8–12 h before deterioration, thus giving sufficient time for clinical teams to intervene with their proposed method achieving a PPV of 0.7758.

To identify early clinical deterioration by combining physiologic and/or laboratory measures to generate a quantified score as well as developing and evaluating a machine

learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children; Zhai et al. (2014) proposed a logistic regression method using 29 variables for predicting unplanned transfer of pediatric patients from general wards to pediatric ICU (PICU) 24 h before the transfer. Data used in the study was collected from a single PICU health center with 29 variables computed using the patient's vital signs, having a total of 6772 normal and 526 abnormal cases. Their dataset can be considered lacking concerning the reliability of data as they considered abnormal cases that were in need of transfer to PICU as normal in cases where clinical limitation prohibited the transfer such as a lack of beds along with the use of clinical physicians to compensate for missing values in the data. While their proposed method achieved an AUROC of 0.912, having a less than ideal dataset might have had a negative impact on the accuracy of their proposed method.

5.3 Systems designed for monitoring variables that can be associated with the deterioration of patient's health

AlNuaimi et al. (2015) proposed a decision tree-based method for predicting the chance of death in patients warded at the ICU. Experimenting using the MIMIC-II dataset, they determined that lab tests could be used as variables based on a maximum relevance selection approach. They further used feature selection and dimensionality reduction techniques to obtain the best set of features extracted from lab tests. They implemented and compared four different classifiers based on their performance using a ranking system and determined that a C4.5 Decision Tree showed the best potential with a mean accuracy of 0.7868. While accurate, they recommended that a more extensive dataset coupled with more advanced classification and feature selection could drastically increase the accuracy and helping the clinical staff in better management of patients. Mochizuki et al. (2017) designed a study to validate the applicability of Respiratory Rate (RR) as a vital sign for monitoring deterioration using statistical analysis. Based on a dataset of 340 cases divided equally to healthy and unhealthy cases, their study concluded that the most important vital signs are systolic blood pressure, heart rate, respiratory rate, SpO_2, body temperature, and Glasgow Coma Scale. The study used respiratory rate as a significant factor while deciding whether to discharge patients in emergency departments. While their data collection could be considered less than ideal, as there was a probability that the patient's vital signs were not fully recorded, patients with an increased respiratory rate in the deteriorated group were not adequately identified and separated.

Quinten et al. (2018) designed a study to illustrate the importance and effects of frequent measurement of patient's vital signs at the emergency departments in identifying patients in danger of clinical deterioration than singular readings of vital signs at the admission to the emergency department using Logistic Regression. Their dataset consisted of HR, RR, and BP information gathered from a tertiary care teaching hospital with a total of 253 normal and 106 abnormal cases with a prediction window of 72 h. As was the case with the study done by Mochizuki et al. (2017), respiratory rate measurements at triage in the emergency department did not adequately record their dataset resulting in low performance. Patients who deteriorated had a lower mean arterial pressure (MAP) and a higher respiratory frequency at admission to the emergency department. The base model, extended with the patient's vital signs at emergency department admission, showed that both a higher heart rate and a higher respiratory rate were associated with a higher risk of deterioration. A higher MAP at ED admission was associated with a lower risk of deterioration. The body temperature at ED admission was not independently associated with deterioration.

Although the performance of their method can be considered low, it showed the importance of periodical measurements of patient's vital signs in ensuring prompt response in case of clinical deterioration.

To compare the effectiveness of MIMIC-II database in implementing new predictive models to predict deterioration of patients; the next two studies were reviewed. Reyes-García et al. (2018) designed a study to illustrate the impact of missing data on prediction models where a General Regression Neural Network (GRNN) algorithm was used for modeling forecasted vital signs along with an SVM used for identifying abnormal cases. They extracted MAP and HR of patients 25–60 years old as variables from the MIMIC-II dataset with a prediction window of 1 h with MAP and HR signals stored as minute-by-minute numeric measurements computationally converted to MAP and HR time series. Each time series in the dataset was randomly split into 70% training and 30% testing sets with an additional reduced validation set obtained from the whole training dataset that corresponded to 10% of the training subset using the k-medoid algorithm (Han et al. 2001). Moreover, the study used two methods to label the training dataset. The first method used a "fixed ranges-based labeling" that defined a normality threshold with a lower and upper limit for every vital sign where measurements of 60–110 for HR and 50–120 for MAP with measurements outside these ranges considered abnormal. The second method used a clustering algorithm-based labeling that labeled samples in the reduced training dataset. The study also included pre-processing steps within the observation window to correct for incorrect values consisting of several steps such as interval selection, null values interpolation, peak suppression, and normalization. Achieving a sensitivity of 0.982 and specificity of 0.641, their study showed that it might be possible to increase the prediction accuracy of EWS systems in cases where some measurements are missing from the data.

Masud and Al Harahshen (2016) designed a study aimed at selecting the most representative features computed using variables from laboratory test results for predicting the chance of mortality in patients coupled with an ensemble-based classification approach. There extracted lab results of 2173 control cases and 1449 abnormal cases (diseased patients) from the MIMIC-II dataset with a total of 287 features for each patient. The extracted patients with the same set of lab tests were grouped together and further categorized into five categories, namely new-borns consisting of patients that were younger than three months old, infants consisting of patients between four months to two years old, children consisting of patients three to seventeen years old, adults consisting of patients between eighteen to sixty-four years old and seniors consisting of patients older than 65 years old with each category being handled as a separate prediction problem. For each category, Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (J48), and Bayes Net (BN) classifiers were trained and combined in an ensemble based on weighted majority voting. The authors assessed the supposed technique on a real ICU patients' data and accomplished distinguished success in minimizing 50% or more of the feature vector while refining the prediction accuracy by 2–5% and achieving more than 200% speedup in most cases. For the J48 learner applied on the Adult dataset, the weighted average f-measures (Fwa) value of Feature Vector Compaction and Ensemble Training (FCE) = 0.78, whereas that of the baselines 1 and 2 are 0.72 and 0.71, respectively.

The next two studies demonstrate the importance of International Statistical Classification of Diseases and Related Health Problems (ICD codes), and clinical notes to implement predictive models for deterioration of patients. Rafiq et al. (2018) designed a study aimed at developing deep learning-based methods for identifying the factors contributing to hospital readmissions of patients (within 30-days) with multiple chronic concurrent conditions (diabetes, cardiovascular and kidney diseases) known as MCC using sequential Electronic

Health Records (EHR) gathered from 610 patient undergoing treatment at Danderyd Hospital in Stockholm, Sweden. 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 the later assessment of these EHR data tedious and time-consuming. Based on the illness, the patients were categorized as having chronic obstructive pulmonary disease, kidney transplant, and paroxysmal ventricular tachycardia. In their study, Word2Vec approach (Chien et al. 2018) was used to convert the non-sequential records in EHR data into a vector form and then a CNN was used to reordered and make the EHR sequential. Word2Vec uses procedural codes and diagnoses in the form of International Statistical Classification of Diseases and Related Health Problems (ICD codes) and converts them to words in the form of output vectors. This sequenced EHR data is then used in an RNN deep learning architecture for predicting the hospital readmissions with their proposed method achieving an accuracy of 0.794 in predicting the possibility of hospital readmissions within a 30-day timeframe. It can be concluded that the main advantage of CNN compared to its predecessors is that it automatically detects the crucial features without any human supervision. CNN is also computationally efficient due to the fact that it utilises special convolution and pooling operations and performs parameter sharing (He et al. 2018).

Chien et al. (2018) designed a study aimed at developing a deep learning-based method for identifying and organizing the patient's information stored in EHR records for use in screening methods for ensuring that patients receive care that is consistent with the medical institute's goals and standards. They extracted EHR records, and clinical notes of 58,000 ICU warded patients from the MIMIC-III dataset divided into 70% training and 30% testing sets. Using the NeuroNER deep learning framework Dernoncourt et al. (2017) which is specially designed for recognizing entities of interest in the text (such as location and name), their proposed method was able to achieve a sensitivity of 0.935, PPV of 0.905, specificity 0.91 and an F-score of 0.92. The Vita Sign Big Data (ViSiBiD) model aimed at identifying critical clinical events of home-monitored patients in advance was proposed by Forkan et al. (2017) where the model uses heart rate, blood pressure, respiratory rate, blood oxygen saturation, and body temperature extracted from the MIMIC-II dataset as the vital signs. Their proposed model uses a Random Forest classifier with a simple feature selection method called "forward feature subset selection" for selecting the most representative features. While their proposed method can be considered as one of the better approaches as they have used an appropriate set of features, a more complex feature selection step combined with a better classifier should further increase the accuracy of their proposed method. Aimed at decreasing the workload of hospital staff load, especially in emergency conditions, by providing early notifications when there is a sudden and unexpected deterioration of the patient's health via remote patient monitoring, their proposed method was able to achieve an accuracy of 0.9585.

Some studies use different endpoints to measure deterioration. The next two studies use the onset of septic shock as the definition of patient's deterioration. Ghosh et al. (2017) proposed a predictive model to prevent dangerous complications such as sepsis or septic shock in ICUs that may produce multiple organ failures and subsequently result in fatalities. The study merges highly informative sequential models obtained from multiple physiological variables and describes the interfaces amongst these models using Coupled Hidden Markov Models (CHMM). Models are obtained from three non-invasive waveform measurements (minimum of 1 h of measurement) of MAP, HR, and RR of 1310 adult septic shock patients from the MIMIC-II dataset. The likelihood of the discrete multivariate test sequence was estimated at the level of 0.71 by CHMM. Desautels et al. (2016) applied

InSight, a machine learning classification system that utilizes multivariable combinations of SBP, pulse pressure (PP), HR, Temp, SpO_2, age and GCS obtained from the MIMIC-III dataset, restricted to intensive care unit (ICU) patients aged 15 years or more. Their proposed InSight system was able to achieve an AUROC of 0.8799. While accurate, the study was developed using patient data recorded using the Metavision system, which results in only a subset of patients from the MIMIC-III dataset to be included in the study, which can severely impact the generalizability of the study.

Tang et al. (2010) proposed an SVM based approach for separating sepsis continuum patients into severe sepsis and systemic inflammatory response syndrome (SIRS) groups using a dataset comprised of patients at the risk of SIRS/sepsis continuum who visited the Emergency Department of the Prince of Wales Hospital from August 2006 to January 2007. Variables used in the study were the ECG, index finger PPG (Fin-PPG), and ear-lobe PPG (Ear-PPG) signals that were measured from the patients in a supine position. Their proposed model was able to achieve a sensitivity of 0.94, a specificity of 0.62, PPV of 0.85, NPV of 0.83, and an accuracy of 0.84 with the critical limitation of this study being the utilization of a small sample size. Due to the importance of static data (i.e. data that is measured only once for a patient) such as age, gender and admission type; Crump et al. (2009) proposed a Multivariate Bayesian model in conjunction with rule-based time-series statistical techniques for monitoring of ICU patients using clinical data collected at Virginia Commonwealth University (VCU) at Level One Trauma facility containing age, gender, admission diagnosis, Temp, HR, and arterial oxygen saturation (SpO_2) from 52 patients in the ICU. Their proposed Bayesian model was able to achieve an AUROC of 0.70 in detecting patients at risk of deterioration. The low performance of the study can be attributed to the small dataset used, which limits the generalizability of their proposed method.

Eshelman et al. (2008) proposed a rule-based method using RIPPER (Repeated Incremental Pruning to Produce Error Reduction) for identifying ICU patients who are at risk of becoming hemodynamically unstable using MIMIC-II dataset. The study used 12,695 adult patient records collected between 2001 and 2005 for improving and evaluating predictive alerts to indicate impending physiologic instability. The proposed model comprises of a set of 15 rules with every rule having a list of conditions that have to be fulfilled based on the measurements of BUN (blood urea nitrogen), WBC (white blood cell count), PTT (partial thromboplastin time), hematocrit, HR, SBP (arterial if available, otherwise non-invasive) and Oxygenation Index (OxI) based on the last recorded value for each measurement. Their proposed method was able to achieve a sensitivity of 0.6067, a specificity of 0.9285, and a PPV of 0.7970. Table 3 summarizes the studies mentioned earlier.

6 Discussion

This research investigated differences in several proposed models to predict deterioration of patients in terms of study goal, variables used, machine-learning techniques, performance, and data source. A review of various deterioration prediction techniques from the literature makes it clear that most methods are planned around similar fundamental concepts. First, case and control groups are identified. Then, predictor variables are determined where vital signs and/or lab tests and/or demographic data are often used. Then, the observation window and prediction window are determined based on the study requirements. Additionally, pre-processing steps are often performed for intervals selection, missing values interpolation, peak suppression and normalization. Finally, the model for prediction of patients'

Table 3 Summary of previous works related to prediction models

Author	Study goal	Variables used	Machine-learning techniques	Performance	Data source
Ordoñez et al. (2016)	Hypotension scenario within an hour	SpO_2, SBP, DBP	k-NN	Accuracy: 0.852PPV: 0.82 Sensitivity: 0.87 F-score: 0.86	The 2009 physio-net challenge
Lee and Mark (2010a)	Hypotensive events	HR, SBP, DBP, MAP	ANN	AUROC: 0.934 Accuracy: 0.861 Sensitivity: 0.851 Specificity: 0.862 NPV: 0.995	MIMIC-II
Lee and Mark (2010b)	Hypotensive events	MAP, HR, pulse pressure, relative cardiac output	ANN	AUROC: 0.918 Sensitivity: 0.826 Specificity: 0.859	MIMIC-II
Lee et al. (2016b)	Occurrence of Ventricular Tachycardia (VT) within the next hour	HRV, RRV	ANN	Accuracy: 0.853 Sensitivity: 0.882 Specificity: 0.824 PPV: 0.833 NPV: 0.875 AUROC: 0.93	Private data from a single medical center
Ong et al. (2012)	Possibility of cardiac arrest within the next 72 h	HRV, age, gender, medical history, HR, BP, SpO_2, RR, GCS	SVM	AUROC: 0.781	Private data from a single medical center
Chen et al. (2017)	Cardiorespiratory insufficiency	HR, RR, SpO_2, SBP, DBP	Random forest classification model	AUROC: 0.94	Private data from a single medical center
Donald et al. (2012)	Hypotensive events	Age, sex, SBP, BP (mean), HR	Bayesian ANN	Sensitivity: 0.40 Specificity: 0.86	Private data
Wickramasinghe (2017)	Unplanned readmissions after the discharge of hospitalized patients	medical records	Convolutional neural network (CNN) and logistic regression	AUROC: 0.819	Electronic medical records
Wellner et al. (2017)	Unplanned transfer of patients to ICU	Vital signs, lab tests, nurse notes, and acuity	Feed-Forward Neural Network	Maximum c-static value: 0.902	Private data from 3 children hospitals

Table 3 (continued)

Author	Study goal	Variables used	Machine-learning techniques	Performance	Data source
Hu et al. (2016)	Unplanned transfer to ICU or onset of cardiac arrest	Vital signs and lab tests	NN	PPV: 0.7758	Private data from a single medical center
Zhai et al. (2014)	Unplanned transfer of pediatric patients from general wards to pediatric ICU (PICU)	Vital signs	Logistic regression	AUROC: 0.912	Private data from a single PICU
AlNuaimi et al. (2015)	Mortality	Lab tests	Decision Tree	Accuracy: 0.7868	MIMIC-II
Quinten et al. (2018)	Importance of frequent measurement of patient's vital signs in emergency departments	HR, RR, and BP	Logistic Regression	AUROC: 0.679	Private data from a single medical center
Reyes-García et al. (2018)	Predict the impact of missing data on prediction models	MAP + HR	GRNN + SVM	Sensitivity: 0.982 Specificity: 0.641	MIMIC-II
Masud et al. (2016)	Importance of the most representative features	Lab tests	weighted majority voting	Fwa = 0.78	MIMIC-II
Rafiq et al. (2018)	Identifying the factors contributing to hospital readmissions of patients (within 30-days)	International statistical classification of diseases and related health problems (ICD codes)	CNN	Accuracy: 0.794	Danderyd hospital in Stockholm, Sweden
Chien et al. (2018)	Identifying and organizing the patient's information stored in EHR records	EHR records and clinical notes	NeuroNER deep learning framework	Sensitivity: 0.935 Specificity: 0.91 PPV: 0.905 F-score: 0.92	MIMIC-III
Forkan et al. (2017)	Identify dangerous clinical events of a home-monitoring patient in advance	Vital signs	Hidden Markov model	Accuracy: 0.9585	MIMIC-II
Ghosh et al. (2017)	The onset of septic shock	MAP, HR, RR	Coupled hidden Markov models	Likelihood: 0.71	MIMIC-II

Table 3 (continued)

Author	Study goal	Variables used	Machine-learning techniques	Performance	Data source
Desautels et al. (2016)	Sepsis onset	SBP, pulse pressure, HR, RR, T, SpO_2, age, GCS	Continuous nonlinear function approximations	AUROC: 0.8799	MIMIC-III
Tang et al. (2010)	Discriminate severe sepsis patients from SIRS patients	ECG signal, PPG waveform	SVM	Sensitivity: 0.94 Specificity: 0.62 PPV: 0.85 NPV: 0.8 Accuracy: 0.84	Emergency department of the Prince of Wales hospital
Crump et al. (2009)	Setting alerts from personal baselines	Age, gender, T, HR, SpO_2, admission diagnosis	Bayesian network models	AUROC: 0.91	Virginia Commonwealth University
Eshelman et al. (2008)	Identifying hemodynamically unstable patients	BUN, WBC, PTT, Ht, HR, SBP, oxygenation index	Rule-based method algorithm	Sensitivity: 0.60 Specificity: 0.9285 PPV: 0.7970	MIMIC-II

deterioration is implemented and validated. An overview of various techniques proposed for the prediction of patients' deterioration shows that most methods have chosen to use features derived from the original set of features proposed by Morgan et al. (1997) for the original EWS system, while the MIMIC dataset contains many more features. Some studies use different types of predictor variables such as vital signs (Zhai et al. 2014; Forkan et al. 2017; Chen et al. 2017; Ordoñez et al. 2016; Mochizuki et al. 2017) and/or lab tests (Masud and Al Harahsheh 2016; AlNuaimi et al. 2015) and/or demographic data (Wellner et al. 2017; Desautels et al. 2016; Donald et al. 2012; Crump et al. 2009). To improve the predictive model's performance and/or to provide more prediction tasks during the implementation of a model, the combination of several predictors has been proposed by many researchers and is adopted by many approaches as studies have shown that the performance of the model can be improved by utilizing different types of predictors (Wellner et al. 2017; Donald et al. 2012; Crump et al. 2009). Although some studies have used feature selection for increasing the accuracy and optimize their proposed approach (Huang and Wang 2006; Capan et al. 2015; AlNuaimi et al. 2015; Forkan et al. 2017), majority of methods are focused on a specific task such as mortality prediction (Polley and Van Der Laan 2010), condition monitoring (Tlegenov et al. 2018) and length of stay (Strzelczyk et al. 2017; Morris et al. 2016). While this approach is beneficial in predicting the possible health complications for the admitted patients, they do not utilize all the information gathered, such as the results of laboratory testing.

Recently, there is a revolution taking effect in the medical analysis field fuelled by the availability of more data combined with more advanced machine learning techniques, with the most important challenge being the extraction of beneficial information embedded in the data. Furthermore, the adoption of electronic health record (EHR) systems (Wellner et al. 2017; Hu et al. 2016; Quinten et al. 2018; Zhai et al. 2014; Lee et al. 2016a; Chen et al. 2017; Ong et al. 2012; Donald et al. 2012; Tang et al. 2010; Crump et al. 2009); Wickramasinghe 2017; Mochizuki et al. 2017; Rafiq et al. 2018) has been increasing as more medical centers are transitioning to digital record keeping (Johnson et al. 2017a, b). This trove of digital clinical data offers a substantial prospect for data mining, machine learning and deep learning researchers to solve pressing health care issues, like primary triage and risk valuation, prediction of physiologic decompensation, and identification of high-cost patients. While the MIMIC dataset provides a large and diverse set of features, a few approaches have used deep learning for constructing a predictive model. Pirracchio (2016) proposed a Super Learner algorithm (Polley and Van Der Laan 2010), which is an ensemble of machine learning algorithms for predicting hospital mortality in ICU patients using the data provided by the MIMIC-II dataset. Johnson et al. (2017a, b) compared different published studies versus gradient boosting and logistic regression algorithms utilizing a plain set of characteristics obtained from the MIMIC-III dataset Johnson et al. (2016a, b, c) for ICU mortality prediction. Harutyunyan et al. (2017) experimentally authenticated four clinical prediction benchmarking tasks using the MIMIC-III dataset and deep learning algorithms. Purushotham et al. (2017) has shown that deep learning-based models can achieve competitive results on 'raw' features without any feature selection. Table 4 summarizes some methods which have utilized a more significant set of features provided by the MIMIC-II and MIMIC-III datasets. Interestingly, regular ANN (Hu et al. 2016; Wellner et al. 2017; Lee et al. 2016a; Lee and Mark 2010a, 2010b; Donald et al. 2012) is the most common neural network technique to predict deterioration of patients and is often used with one hidden layer.

One of the main motivations that ANNs with multiple fully connected layers have not achieved popularity in various real-world applications for decades is their computation

Table 4 Methods that have utilized a greater set of features provided by the mimic-ii and mimic-iii datasets

Author	Dataset	Prediction model	Feature type	No. of features	Prediction task	AUROC
Purushotham et al. (2017)	MIMIC-III (v1.4)	Gated recurrent unit (GRU)	Hour-by-hour	17	Mortality Length of stay	0.86
Pirracchio (2016)	MIMIC-II	Super learner	Hour-by-hour	15	Mortality	0.88
Harutyunyan et al. (2017)	MIMIC-III (v1.4)	Long short-term memory (LSTM)	Non-time-series	17	Mortality Length of stay Decompensation Phenotyping	0.87
Johnson et al. (2017a, b)	MIMIC-III (v1.4)	Gradient boosting (GB)	Non time-series	37	Mortality	0.84

complexity. The idea of deep learning, also inspired by biological processes, powered by high performance computing hardware, has made very deep models computationally feasible for real-world applications (Chen et al. 2019). Even though many models could attain promising results, there is still room for more improvement, especially considering the new advancements in deep learning and availability of more data such as vital signs, laboratory measurements and demographic data. Patient mortality and sudden transfer to ICU are considered the most crucial outcome for ICU admission. Accurately predicting mortality and sudden transfer to ICU could help with the assessment of the severity of illness and determining the value of novel treatments, interventions and health care policies. Various predictive variables from different patients can be assembled for an extended period to obtain a large dataset and utilized to implement prognostic models through deep learning techniques that can precisely distinguish clinical events. Hence, big data and deep learning are potential approaches to build predictive models for different clinical events. With the current innovations in deep learning approaches, there is expanding importance in employing these approaches in healthcare (Che et al. 2015; Oellrich et al. 2016; Lasko et al. 2013). The crucial distinction between machine learning and deep learning stems from the way data is presented to the network (Chen and Lin 2014). Machine learning algorithms (Hu et al. 2016; Quinten et al. 2018; Zhai et al. 2014; AlNuaimi et al. 2015; Lee et al. 2016a; Lee and Mark 2010a, 2010b; Forkan et al. 2017; Chen et al. 2017; Ghosh et al. 2017; Ordoñez et al. 2016; Ong et al. 2012; Tang et al. 2010; Crump et al. 2009; Mochizuki et al. 2017) quite often require structured data and are not appropriate to work out complex set of features that consist of a considerable amount of data while deep learning systems (Rafiq et al. 2018; Chien et al. 2018; Purushotham et al. 2017; Pirracchio 2016; Harutyunyan et al. 2017; Johnson et al. 2017a, b) can provide better performance.

There is no agreed upon definition for deterioration. However, most definitions agree that goal of prediction of deterioration is to prevent adverse events before its occurrence in a prediction window that gives sufficient time to the medical team to save patients' lives. There are many learning algorithms available for prediction of deterioration. Most of the machine learning algorithms discussed here are categorized as supervised machine learning. That is where an algorithm (classifier) tries to map inputs to desired outputs utilising a specific function. In classification problems a classifier tries to learn several features (variables or inputs) to predict an output (response). In the case of prediction of patient's deterioration, a classifier will try to classify patients to be deteriorated or not by learning certain characteristics (features) in the training process. Figure 3 below illustrates a flow chart to demonstrate the pipelines and indicate the major differences (colored ones) between the different adopted proposed methods.

7 Conclusion

In this paper, various methods were proposed for predicting deterioration in patients using machine learning and deep learning. While traditional EWS systems are still commonly used for the early identification of deterioration, the availability of large datasets and more advanced classification systems have made accurate detection and continuous monitoring of patients a possibility. Deep learning methods offer high potentials as the data needed for a proper system is already available publicly. More research is needed to identify the clinical condition of a patient using the present and past data of several parameters and measurements (i.e. periodic data) to discover relationships between various variables

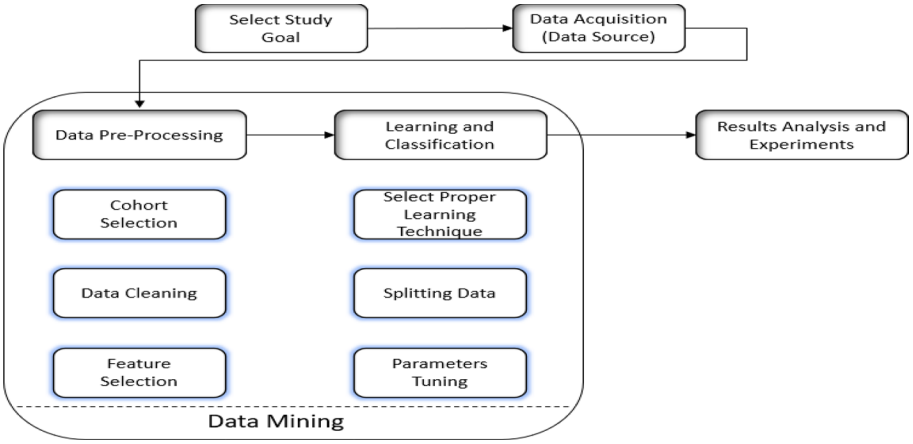

Fig. 3 Flow chart to demonstrate the pipelines and indicate the major differences (colored ones) between the different adopted proposed methods

(predictors) which might lead to valuable diagnostic or prognostic visions, clarifying the consequences of the deterioration for patients. Moreover, the definitions of deterioration, where its endpoint measure different tasks such as mortality and/or sudden transfer to ICU and/or length of stay and/or decompensation and/or phenotyping, need to be standardized for obtaining a more consistent performance among different methods.

Acknowledgement This work was supported by the Ministry of Higher Education under Prototype Research Grant Scheme (PRGS/1/2019/TK04/UTM/02/12), and in part by the UTM International Doctoral Fellowship.

References

AlNuaimi, Noura, Mohammad M Masud, and Farhan Mohammed (2015) ICU patient deterioration prediction: a data-mining approach, arXiv preprint arXiv:1511.06910

Bonafide CP, Russell Localio A, Song L, Roberts KE, Nadkarni VM, Priestley M, Paine CW, Zander M, Lutts M, Brady PW (2014) Cost-benefit analysis of a medical emergency team in a children's hospital. Pediatrics 134:235–241

Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF (2014) Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform 83:983–992

Capan M, Ivy JS, Rohleder T, Hickman J, Huddleston JM (2015) Individualizing and optimizing the use of early warning scores in acute medical care for deteriorating hospitalized patients. Resuscitation 93:107–112

Che Z, Sanjay P, Robinder K, Yan Liu. 2015. Distilling knowledge from deep networks with applications to healthcare domain, arXiv preprint arXiv:1512.03542

Chen L, Ogundele O, Clermont G, Hravnak M, Pinsky MR, Dubrawski AW (2017) Dynamic and personalized risk forecast in step-down units Implications for monitoring paradigms. Ann Am Thorac Soc 14:384–391

Chen Qi, Wang W, Fangyu Wu, De S, Wang R, Zhang B, Huang X (2019) A survey on an emerging area: deep learning for smart city data. IEEE Trans Emerg Top Comput IntelL 3:392–410

Chen X-W, Lin X (2014) Big data deep learning: challenges and perspectives. IEEE Access 2:514–525

Chien I, Alvin S, Alex C, Charlotta L (2018) Identification of serious illness conversations in unstructured clinical notes using deep neural networks. International workshop on artificial intelligence in health. Springer

Churpek MM, Yuen TC, Edelson DP (2013) Predicting clinical deterioration in the hospital: the impact of outcome selection. Resuscitation 84:564–568
Churpek MM, Yuen TC, Park SY, Gibbons R, Edelson DP (2014) Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards. Crit Care Med 42:841
Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP (2016) Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med 44:368
Colque RVHM (2018) Robust approaches for anomaly detection applied to video surveillance
Crump C, Sunil S, Bruce W, Patrick F, Azhar R, Christine TS (2009) Using Bayesian networks and rule-based trending to predict patient status in the intensive care unit. In: AMIA Annual Symposium Proceedings, 124, American Medical Informatics Association
Dernoncourt F, Ji YL, Peter S (2017) NeuroNER: an easy-to-use program for named-entity recognition based on neural networks, arXiv preprint arXiv:1705.05487
Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C (2016) Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. JMIR Med Inform 4:e28
Donald R, Tim H, Ian P, Chambers I, Citerio G, Enblad P, Gregson B, Kiening K, Mattern J, Nilsson P (2012) Early warning of EUSIG-defined hypotensive events using a Bayesian artificial neural network. Intracranial pressure and brain monitoring. Springer
Edelson DP, Carey K, Winslow CJ, Churpek MM (2018) Less is more: detecting clinical deterioration in the hospital with machine learning using only age, heart rate and respiratory rate. C15. Critical care: big data and artificial intelligence in critical illness. American Thoracic Society
Eshelman LJ, Lee KP, Joseph JF, Wei Z, Larry N, Mohammed S (2008) Development and evaluation of predictive alerts for hemodynamic instability in ICU patients. In: AMIA annual symposium proceedings, American Medical Informatics Association. 379
Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874
Forkan AR, Mohammad IK, Atiquzzaman M (2017) ViSiBiD: a learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. Comput Netw 113:244–257
Fukushima K, Ueno Y, Kawagishi N, Kondo Y, Inoue J, Kakazu E, Ninomiya M, Wakui Y, Saito N, Satomi S (2011) The nutritional index 'CONUT'is useful for predicting long-term prognosis of patients with end-stage liver diseases. Tohoku J Exp Med 224:215–219
Gao T, Dan G, Matt W, Radford RJ, Alex A (2006) Vital signs monitoring and patient tracking over a wireless network. In: 2005 IEEE engineering in medicine and biology 27th annual conference, IEEE, 102–05
Ghosh S, Li J, Cao L, Ramamohanarao K (2017) Septic shock prediction for ICU patients via coupled HMM walking on sequential contrast patterns. J Biomed Inform 66:19–31
Grandini M, Enrico B, Giorgio V (2020) Metrics for multi-class classification: an overview, arXiv preprint arXiv:2008.05756
Grant S (2018) Limitations of track and trigger systems and the national early warning score. Part 1: areas of contention. Br J Nurs 27:624–631
Han J, Micheline K, Anthony KHT (2001) Spatial clustering methods in data mining. Geographic data mining and knowledge discovery. Taylor & Francis
Harutyunyan H, Hrant K, David CK, Greg VS, Aram G (2017) Multitask learning and benchmarking with clinical time series data, arXiv preprint arXiv:1703.07771
He N, Fang L, Li S, Plaza A, Plaza J (2018) Remote sensing scene classification using multilayer stacked covariance pooling. IEEE Trans Geosci Remote Sens 56:6899–6910
Henriksen DP, Mikkel B, Annmarie TL (2014) Prognosis and risk factors for deterioration in patients admitted to a medical emergency department, PloS one 9
Hogan H, Hutchings A, Wulff J, Carver C, Holdsworth E, Welch J, Harrison D, Black N (2019) Interventions to reduce mortality from in-hospital cardiac arrest: a mixed-methods study. Health Serv Deliv Res 7:1–110
Hoogendoorn M, Ali El H, Kwongyen M, Marzyeh G, Peter S (2016) Prediction using patient comparison vs. modeling: a case study for mortality prediction. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, 2464–67
Hu SB, Wong DJL, Correa A, Li N, Deng JC (2016) Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model. PLoS ONE 11:e0161401
Huang C-L, Wang C-J (2006) A GA-based feature selection and parameters optimizationfor support vector machines. Expert Syst Appl 31:231–240

Johnson A, Pollard T, Mark R III (2016a) MIMIC-III clinical database. PhysioNet 10:C2XW26
Johnson AEW, Tom JP, Roger GM (2017) Reproducibility in critical care: a mortality prediction case study. In: Machine learning for healthcare conference, 361–76
Johnson AEW, Pollard TJ, Lu Shen H, Li-wei L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) MIMIC-III, a freely accessible critical care database. Sci data 3:160035
Johnson AEW, Stone DJ, Celi LA, Pollard TJ (2017) The MIMIC code repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 25:32–39
Johnson AEW, Stone DJ, Celi LA, Pollard TJ (2018) The MIMIC code repository: enabling reproducibility in critical care research. J Am Med Inform Assoc 25:32–39
Johnson L, Zheng M, Vorobyeva Y, Gabriel A, Qi H, Velásquez N (2016) NMC Horizon report: 2016 higher education edition
Kate RJ, Perez RM, Mazumdar D, Pasupathy KS, Nilakantan V (2016) Prediction and detection models for acute kidney injury in hospitalized older adults. BMC Med Inform Decis Mak 16:39
Kipnis P, Turk BJ, Wulf DA, LaGuardia JC, Liu V, Churpek MM, Romero-Brufau S, Escobar GJ (2016) Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. J Biomed Inform 64:10–19
Kivipuro M, Tirkkonen J, Kontula T, Solin J, Kalliomäki J, Pauniaho S-L, Huhtala H, Yli-Hankala A, Hoppu S (2018) National early warning score (NEWS) in a Finnish multidisciplinary emergency department and direct vs. late admission to intensive care. Resuscitation 128:164–169
Komorowski M, Leo AC, Omar B, Anthony CG, Aldo AF (2018) The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 24:1716
Kononenko I (2001) Machine learning for medical diagnosis: history, state of the art and perspective. Artif Intell Med 23:89–109
Lasko TA, Joshua CD, Mia AL (2013) Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data, PloS one 8
Lavrač N (1999) Machine learning for data mining in medicine. In: Joint European conference on artificial intelligence in medicine and medical decision making, Springer, 47–62
LeCun Y, Yoshua B, Geoffrey H (2015) Deep learning. Nature 521:436–444
Lee H, Shin S-Y, Seo M, Nam G-B, Joo S (2016) Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks. Sci Rep 6:32390
Lee H, Shin S-Y, Seo M, Nam G-B, Joo S (2016) Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks. Sci Rep 6:1–7
Lee J, Mark RG (2010a) A hypotensive episode predictor for intensive care based on heart rate and blood pressure time series. In: 2010 computing in cardiology, IEEE, 81–84
Lee J, Mark RG (2010) An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. Biomed Eng Online 9:62
Legates DR, Mccabe GJ Jr (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 35:233–241
Li Q, Clifford GD (2012) Signal quality and data fusion for false alarm reduction in the intensive care unit. J Electrocardiol 45:596–603
Liaw SY, Scherpbier A, Klainin-Yobas P, Rethans J-J (2011) A review of educational strategies to improve nurses' roles in recognizing and responding to deteriorating patients. Int Nurs Rev 58:296–303
Liu V, Kipnis P, Rizk NW, Escobar GJ (2012) Adverse outcomes associated with delayed intensive care unit transfers in an integrated healthcare system. J Hosp Med 7:224–230
Liu Z, Zuren F, Liangjun K (2015) Fireworks algorithm for the multi-satellite control resource scheduling problem. In: 2015 IEEE congress on evolutionary computation (CEC), IEEE, 1280–86
Manning T, Sleator RD, Walsh P (2014) Biologically inspired intelligent decision making: a commentary on the use of artificial neural networks in bioinformatics. Bioengineered 5:80–95
Mao Y, Wenlin C, Yixin C, Chenyang L, Marin K, Thomas B (2012) An integrated data mining approach to real-time clinical monitoring and deterioration warning. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, 1140–48
Mardini L, Lipes J, Jayaraman D (2012) Adverse outcomes associated with delayed intensive care consultation in medical and surgical inpatients. J Crit Care 27:688–693
Masud MM, Al Harahsheh AR (2016) Mortality prediction of ICU patients using lab test data by feature vector compaction and classification. In: 2016 IEEE international conference on big data (big data), IEEE, 3404–11
Mochizuki K, Shintani R, Mori K, Sato T, Sakaguchi O, Takeshige K, Nitta K, Imamura H (2017) Importance of respiratory rate for the prediction of clinical deterioration after emergency department discharge: a single-center, case–control study. Acute Med Surg 4:172–178

Mokart D, Lambert J, Schnell D, Fouché L, Rabbat A, Kouatchet A, Lemiale V, Vincent F, Lengliné E, Bruneel F (2013) Delayed intensive care unit admission is associated with increased mortality in patients with cancer with acute respiratory failure. Leuk Lymphoma 54:1724–1729

Moody GB, Lehman L-WH (2009) Predicting acute hypotensive episodes: the 10th annual physionet/computers in cardiology challenge. In: 2009 36th annual computers in cardiology conference (CinC), IEEE, 541-44

Morgan, RJMWF, Lloyd-Williams F, Wright MM, Morgan-Warren RJ (1997) An early warning scoring system for detecting developing critical illness

Morris PE, Berry MJ, Clark Files D, Clifton Thompson J, Hauser J, Flores L, Dhar S, Chmelo E, Lovato J, Douglas L, Case. (2016) Standardized rehabilitation and hospital length of stay among patients with acute respiratory failure: a randomized clinical trial. JAMA 315:2694–2702

Newman S (2017) Do not disturb: vital sign monitoring as a predictor of clinical deterioration in monitored patients, Kentucky Nurs, 65

Nicolson A, Paliwal KK (2019) Deep learning for minimum mean-square error approaches to speech enhancement. Speech Commun 111:44–55

Oellrich A, Collier N, Groza T, Rebholz-Schuhmann D, Shah N, Bodenreider O, Boland MR, Georgiev I, Liu H, Livingston K (2016) The digital revolution in phenotyping. Brief Bioinform 17:819–830

Ong ME, Hock CH, Ng L, Goh K, Liu N, Koh ZX, Shahidah N, Zhang TT, Fook-Chong S, Lin Z (2012) Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. Crit Care 16:R108

Ordoñez P, Schwarz N, Figueroa-Jiménez A, Garcia-Lebron LA, Roche-Lima A (2016) Learning stochastic finite-state transducer to predict individual patient outcomes. Health Technol 6:239–245

Panday RSN, Minderhoud TC, Alam N, Nanayakkara PWB (2017) Prognostic value of early warning scores in the emergency department (ED) and acute medical unit (AMU): a narrative review. Eur J Intern Med 45:20–31

Paradiso R (2003) Wearable health care system for vital signs monitoring. In: 4th international IEEE EMBS special topic conference on information technology applications in biomedicine, IEEE, 283–86.

Pawitan Y (2001) In all likelihood: statistical modelling and inference using likelihood. Oxford University Press

Pirracchio R (2016) Mortality prediction in the icu based on mimic-ii results from the super icu learner algorithm (sicula) project. Secondary analysis of electronic health records. Springer

Plate JDJ, Peelen LM, Leenen LPH, Hietbrink F (2018) Validation of the VitalPAC early warning score at the intermediate care unit. World J Critical Care Med 7:39

Polley EC, van der Laan MJ (2010) Super learner in prediction. Springer

Prytherch DR, Smith GB, Schmidt P, Featherstone PI, Stewart K, Knight D, Higgins B (2006) Calculating early warning scores—a classroom comparison of pen and paper and hand-held computer methods. Resuscitation 70:173–178

Purushotham S, Chuizheng M, Zhengping C, Yan L (2017) Benchmark of deep learning models on large healthcare mimic datasets, arXiv preprint arXiv:1710.08531

Qi J, Jun Du, Siniscalchi SM, Ma X, Lee C-H (2020) On mean absolute error for deep neural network based vector-to-vector regression. IEEE Signal Process Lett 27:1485–1489

Quinten VM, van Meurs M, Olgers TJ, Vonk JM, Ligtenberg JJM, ter Maaten JC (2018) Repeated vital sign measurements in the emergency department predict patient deterioration within 72 hours: a prospective observational study. Scand J Trauma Resusc Emerg Med 26:57

Rafiq M, George K, Pamela M, Jonas S, Carl S, and Christian G (2018) Deep learning architectures for vector representations of patients and exploring predictors of 30-day hospital readmissions in patients with multiple chronic conditions. In: International workshop on artificial intelligence in health, Springer, 228–44

Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M (2018) Scalable and accurate deep learning with electronic health records. NPJ Digit Med 1:18

Ren J (2012) ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging. Knowl-Based Syst 26:144–153

Reyes-García J, Galeana-Zapién H, Galaviz-Mosqueda A, Torres-Huitzil C (2018) Evaluation of the impact of data uncertainty on the prediction of physiological patient deterioration. IEEE Access 6:38595–38606

Rothman MJ, Rothman SI, Joseph Beals IV (2013) Development and validation of a continuous measure of patient condition using the electronic medical record. J Biomed Inform 46:837–848

Saeed M, Christine L, Greg R, Roger GM (2002) MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In: Computers in cardiology, IEEE, 641–44

Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L-W, Moody G, Heldt T, Kyaw TH, Moody B, Mark RG (2011) Multiparameter intelligent monitoring in intensive care II (MIMIC-II): a public-access intensive care unit database. Crit Care Med 39:952

Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10:e0118432

Scalzo F, Liebeskind D, Xiao Hu (2012) Reducing false intracranial pressure alarms using morphological waveform features. IEEE Trans Biomed Eng 60:235–239

Schmid F, Goepfert MS, Reuter DA (2013) Patient monitoring alarms in the ICU and in the operating room. Crit Care 17:216

Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM (2016) The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA 315:801–810

Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI (2013) 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 84:465–470

Smith GB, Prytherch DR, Schmidt PE, Featherstone PI (2008) Review and performance evaluation of aggregate weighted 'track and trigger'systems. Resuscitation 77:170–179

Spångfors M, Arvidsson L, Karlsson V, Samuelson K (2016) The national early warning score: translation, testing and prediction in a Swedish setting. Intensive Crit Care Nurs 37:62–67

Stanzani M, Lewis RE (2018) Development and applications of prognostic risk models in the management of invasive mold disease. J Fungi 4:141

Strzelczyk A, Ansorge S, Hapfelmeier J, Vijayveer Bonthapally M, Erder H, Rosenow F (2017) Costs, length of stay, and mortality of super-refractory status epilepticus: a population-based study from Germany. Epilepsia 58:1533–1541

Subbe CP, Kruger M, Rutherford P, Gemmel L (2001) Validation of a modified early warning score in medical admissions. QJM 94:521–526

Tang CHH, Middleton PM, Savkin AV, Chan GSH, Bishop S, Lovell NH (2010) Non-invasive classification of severe sepsis and systemic inflammatory response syndrome using a nonlinear support vector machine: a preliminary study. Physiol Meas 31:775

Taylor MM, Douglas-Creelman C (1967) PEST: efficient estimates on probability functions. J Acoust Soc Am 41:782–787

Taylor RA, Joseph RP, Arjun KV, Hani M, Edward RM, William F, Kennedy-Hall M (2016) Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach. Acad Emerg Med 23:269–278

Tilly KF, Belton AB, McLachlan JFC (1995) Continuous monitoring of health status outcomes: experience with a diabetes education program. Diabet Educ 21:413–419

Tlegenov Y, Hong GS, Wen Feng Lu (2018) Nozzle condition monitoring in 3D printing. Robot Comput-Integr Manuf 54:45–55

Wang W, Yanmin L (2018) Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model. In: IOP conference series materials science and engineering, 012049

Wellner B, Grand J, Canzone E, Coarr M, Brady PW, Simmons J, Kirkendall E, Dean N, Kleinman M, Sylvester P (2017) Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements. JMIR Med Inform 5:e45

Wickramasinghe N (2017) Deepr: a convolutional net for medical records

Wiley JF, Pace LA (2015) Multiple regression beginning R. Springer

Williams B, Alberti G, Ball C, Ball D, Binks R, Durham L (2012) Royal college of physicians, national early warning score (NEWS), standardising the assessment of acute-illness severity in the NHS, London

Young MP, Gooder VJ, Bride KM, James B, Fisher ES (2003) Inpatient transfers to the intensive care unit: delays are associated with increased mortality and morbidity. J Gen Intern Med 18:77–83

Zhai H, Brady P, Li Qi, Lingren T, Ni Y, Wheeler DS, Solti I (2014) Developing and evaluating a machine learning based algorithm to predict the need of pediatric intensive care unit transfer for newly hospitalized children. Resuscitation 85:1065–1071

Zheng Y, Qi L, Enhong C, Yong G, Leon-Zhao J (2016) Exploiting multi-channels deep convolutional neural networks for multivariate time series classification. Front Comput Sci 10:96–112

Authors and Affiliations

Tariq Ibrahim Al-Shwaheen[1] · Mehrdad Moghbel[2] · Yuan Wen Hau[1] · Chia Yee Ooi[2]

Mehrdad Moghbel
mehrdad2275@gmail.com

Yuan Wen Hau
hauyuanwen@biomedical.utm.my

Chia Yee Ooi
ooichiayee@utm.my

[1] Faculty of Engineering, UTM-IJN Cardiovascular Engineering Center, School of Biomedical Engineering and Health Sciences, Universiti Teknologi Malaysia, UTM, 81310 Johor Bahru, Johor, Malaysia

[2] Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia