Prediction of Patients Severity at Emergency Department Using NARX and Ensemble Learning

Naemi, Amin; Mansourvar, Marjan; Schmidt, Thomas; Wiil, Uffe Kock

Published in:
2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

DOI:
10.1109/BIBM49941.2020.9313462

Publication date:
2020

Prediction of Patients Severity at Emergency Department Using NARX and Ensemble Learning

Amin Naemi
SDU Health Informatics and Technology,
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
amin@mmmi.sdu.dk

Thomas Schmidt
SDU Health Informatics and Technology
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
schmidt@mmmi.sdu.dk

Marjan Mansourvar
SDU Health Informatics and Technology
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
marm@mmmi.sdu.dk

Uffe Kock Wiil
SDU Health Informatics and Technology
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
ukwiil@mmmi.sdu.dk

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313462

Abstract— **Early detection of adverse events at hospitals could be useful in terms of reducing costs, morbidity, and mortality. Therefore, in this paper, we present a personalized real-time hybrid model based on Nonlinear Autoregressive Exogenous (NARX) model and Ensemble Learning (EL) to predict patients' severity during hospitalization at Emergency Departments (ED). This model utilizes vital signs of patients, including Pulse Rate (PR), Respiratory Rate (RR), Arterial Blood Oxygen Saturation (SpO2) and Systolic Blood Pressure (SBP), which are collected automatically during the treatment to predict the illness severity of hospitalized patients at ED in the next hour based on their vital signs of the previous two hours. Two EL algorithms, including Random Forest (RF) and Adaptive Boosting (AdaBoost) are considered to build hybrid models. The performance of NARX-EL models is compared with Auto Regressive Integrated Moving Average (ARIMA), combination of NARX and Linear Regression (LR), Support Vector Regression (SVR) and K-Nearest Neighbors Regression (KNN). The results show that our proposed hybrid models can predict patients' severity with significantly higher accuracy. It is also found that NARX-RF has the best performance in the prediction of sudden changes and unexpected adverse events in patients' vital signs (R^2 score = 0.978, NRMSE = 6.16%).**

Keywords— Patient Severity, Machine Learning, Ensemble Learning, NARX, Time Series, Health Informatics.

I. Introduction

It has been shown that around 30% of admitted patients, who look stable, experience deterioration during their stay, and that leads to an increase in morbidity and mortality [1]. The probability of unexpected adverse events and deterioration, resulting in in-hospital death or Intensive Care Unit (ICU) transfer can be decreased by monitoring of patients' vital signs and utilizing predictive models to detect clinical deterioration [2,3]. Thus, many hospitals deploy costly equipment for patient monitoring. Nevertheless, most of these automatically collected data are seldom used in the process of decision-making in clinical environment, because it is difficult for clinicians to process and interpret this large amount of information. Accordingly, valuable information for detection of possible adverse events might be ignored. By conversations with clinicians and monitoring patients, it can be concluded that health status of patients is not easy to quantify [4]. Notwithstanding, dynamic changes in vital signs of patients could be useful to identify patients at the risk of clinical deterioration and adverse events. Output in such dynamic systems not only depends on the input but also depends on the past inputs and outputs values.

Vital signs of hospitalized patients registered over a period of time can be analyzed using time series techniques to predict their conditions in advance. There are various objectives to analyze a time series, but they can be categorized as description, explanation, and prediction [5]. Prediction of time series means given an observed series of values, the future values of series will be forecasted. Furthermore, Machine Learning (ML) algorithms that have been used in a variety of applications can analyze and extract information from these time series. It has been shown that ML algorithms are great tools for data mining applications, especially when a system is dynamic and there are hidden and complex patterns in the data which are difficult for humans to understand and utilize [6]. Therefore, it is essential to consider such techniques which have the ability to model dynamic systems to predict vital signs or severity trajectories of patients. To solve such these problems, the optimal solution often includes a combination of different techniques since each technique has special capabilities to solve the problem [7].

This study aims to develop a real-time model on the individual level to predict illness severity of patients during hospitalization at Emergency Departments (ED). Such a model can help clinicians to better predict patients who will experience adverse events or deterioration and provide clinicians more time to intervene. Our proposed model consists of two parts, including Nonlinear Autoregressive Exogenous (NARX) part and Ensemble Learning (EL) part. NARX is a nonlinear autoregressive model that tries to forecast the next values of a time series based on the past values of it and exogenous inputs. In other words, NARX has a memory of previous inputs and outputs that reflects the historical status of a system. Moreover, considering multi-step delay input and output gives NARX good ability to detect and follow nonlinear behaviors in a time series which makes it a great tool to describe the chaotic time series characteristics [8]. The other part of the proposed model is EL; Random Forest (RF) and Adaptive Boosting (AdaBoost) techniques are implemented as the EL part. Our proposed model utilizes each patient's vital signs and illness severity over time to predict his/her future status.

This paper's remainder is organized as follows: Section II describes related work, in Section III, data and method are explained. In Section IV the performance and results are presented. Section V is discussion, and Section VI concludes this paper.

II. RELATED WORK

ML has been used extensively in health applications [9] because a small improvement in accuracy can have significant advantages, given the high mortality rate related to clinical deterioration and the costs of resources used for false alarms [2]. Various ML algorithms have been applied to predict patients' condition, adverse events, and deterioration during hospitalization and their performances have been evaluated and compared to traditional methods. Hu et al. [10] developed a model based on multilayer perceptron neural network to investigate this model's accuracy in predicting clinical deterioration. They used clinical data such as age, gender, and vital signs to train the model and showed that their neural network model could recognize complex patterns and predict clinical deterioration more precisely than ViEWS, which is one of the best known Early Warning Scores (EWS) systems [11].

Clifton et al. [12] have used the one-class Support Vector Machine (SVM) to identify patients who experience adverse events using vital signs collected from patient monitoring. They compared discriminative and generative approaches (SVM and a probabilistic method using a mixture model, respectively). They showed that existing methods, based on generative mixture distributions had been outperformed by their SVM based method. Churpek et al. [2] have used several machine learning techniques, including Tree-based models, K-Nearest Neighbors (KNN), and SVM to build models for prediction of clinical deterioration on the wards. Different ML-based models were compared with the EWS system MEWS [13]. They found that ML algorithms can be more accurate than traditional logistic regression for predicting clinical deterioration on the wards.

Several ML techniques have been applied to predict the condition of patients in the ICU. Majority of studies in this area utilized logistic regression to predict the mortality, stability, and severity of patients in ICU [14]. However, ML-based systems provide higher prediction accuracy. Hence, researchers have tried to introduce various prediction systems for ICU. Guiza et al. [15] applied Gaussian Process (GP) which is one of the suitable tools to predict and analyze of real dynamic systems, with time series techniques to predict the status and stability of patients. Moreover, Decision Tree (DT) and SVM were deployed to predict cardiac arrest in ICU [16]. Li et al. [17] proposed long short term memory networks to predict the severity of sepsis of patients hospitalized in ICU. Ge et al. [14] used logistic regression and recurrent neural network for mortality prediction in ICU.

Although different ML techniques have been applied to build predictive models in clinical settings, various aspects of these models need further research. First, population-based models have been used widely in the literature [10,12,14,15]. However, current studies in medicine indicated that the patients population is heterogeneous, meaning that each patient has unique and specific characteristics, and these characteristics must be taken into account to build predictive models [18]. Therefore, introducing personalized predictive models that are trained based on the information from a specific patient or similar patients is essential. These personalized predictive models produce more accurate risk scores and use more relevant risk factors for each patient based on his/her characteristics [18]. Second, the human body is a complex dynamic system, and therefore, the binary classification conventionally used in many studies for prediction, cannot represent such dynamics properly [19]. Prediction of any event for this system needs continuous monitoring and considering specific characteristics of individuals.

Therefore, this paper aims to develop individual-based models for prediction of patients' illness severity using time series and ML techniques. The proposed solution consists of NARX and EL models, and utilizes vital signs time series data for prediction.

III. MATERIAL AND METHOD

A. Data collection

Data was collected from patients admitted to ED of Odense University Hospital (OUH) (June 2018 – April 2019) and Hospital of South Western Jutland (HSWJ) (May 2018 – March 2019). Data was collected from the HL7 interface of Philips IntelliVue patient monitors. Vital signs of patients were registered in 60-second intervals and were stored in a database. Messages from HL7 included information about Pulse Rate (PR) and Arterial Blood Oxygen Saturation (SpO2), calculated by pulse oximetry, Respiration Rate (RR), measured using 3-lead electrocardiography. A Noninvasive Blood Pressure (NBP) cuff was used to measure mean, systolic, and diastolic blood pressure. In this study four vital signs including PR, RR, SpO2, and systolic blood pressure (SBP) were utilized. Other information including admission date, gender, age, length of stay, and arrival and departure diagnoses was added to the dataset [20].

The age and sex of patients were extracted using their personal identification numbers. Since this database is a clinical database which is categorized as sensitive data, the personal identification numbers were removed from the database. Moreover, all data was stored on secure MySQL databases hosted on secure virtual servers operated by RegionalIT in the Region of Southern Denmark. Data was stored on Open Patient data Explorative Network[1] (OPEN).

The Danish Data Protection Agency authorized vital signs registration under journal nr. 17/14630 and registered at ClinicalTrials.gov with number: NCT03375658. According to Danish legislation on privacy concerns, the data was stored in a database with restricted access.

B. Preprocessing

Among patients admitted to the EDs of OUH and HSWJ, we excluded four types of patients including: 1) not hospitalized

[1] https://open.rsyd.dk

patients, 2) patients without triage data, 3) patients under 18 years old, and 4) patients with at least two unmonitored vital signs variable. The first group was removed because these patients were not attached to the monitoring devices. The second group was removed because the illness severity condition of patients was not recorded over time. Patients of the third group were removed because the target population of this study was adults, and the patients in the fourth group were ignored because of high amount of missing values. Table 1 provides an overview of the data sets.

Table 1. Data Set Description

	OUH	HSWJ
Patients	6,027	3,582
Number of each vital sign	1,088,650	623,040
Gender, Male n (%)	3,195 (53%)	1,916 (53%)
Age	64.2 (SD: 19.9)	65.2 (SD: 18.6)
PR (minute^{-1}), Median (IQR)	83 (69 - 99)	74 (57 - 91)
RR (minute^{-1}), Median (IQR)	18 (14 - 22)	19 (15 - 23)
SpO2 (%), Median (IQR)	96 (92 - 99)	97 (94 - 99)
SBP (mmHg), Median (IQR)	129 (110 - 147)	134 (116 - 152)
Monitoring Time (minute), Median (IQR)	156 (74 - 398)	491 (234 - 1059)

The distribution of vital signs, monitoring time, and age for OUH and HSWJ is shown in Figure 1.

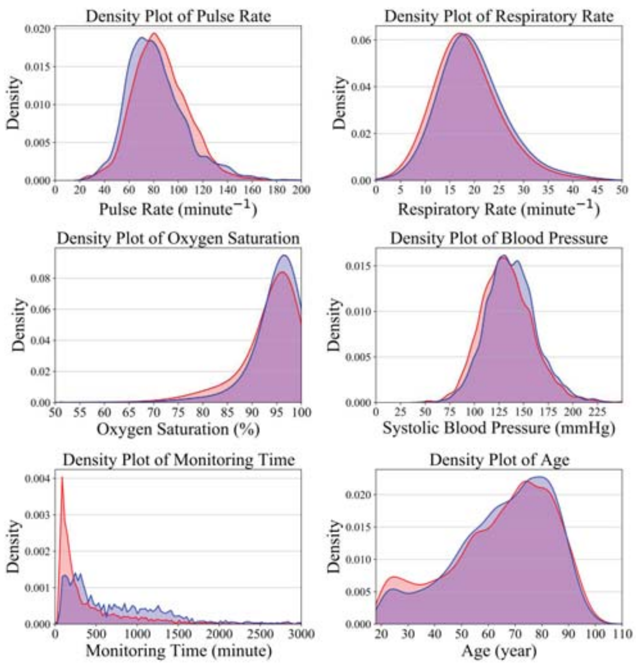

Figure 1. Vital signs, monitoring time and age distributions for patients of OUH (red) and HSWJ (blue).

Handling of missing data for secondary use of electronic health data is important [21]. Therefore, in this study, moving average technique was used to handle missing values. The size of moving average window was set to 10, which means that each missing value was replaced by the average of its ten neighbors. Moreover, since variables have different scales, normalization was applied to all independent variables (exogenous inputs) including, PR, RR, SpO2, and SBP. Normalization causes all variables to have the same scale, so different variables contribute equally to the models. Moreover, to overcome the noise effect and artifacts of equipments during collecting vital signs, the time horizon was set to 10-minutes, meaning the average illness severity of each patient in every 10 minutes interval was calculated and used to build models.

C. *Triage system*

Adaptive Process Triage (ADAPT) is a triage system introduced in Sweden and has been used since 2006. It is also utilized in several Danish hospitals. ADAPT is a track and trigger tool using vital signs according to the ABCDE-principle [22] and a brief systematic questionnaire for each main complaint. ADAPT's primary goal is to identify the seriously ill, help clinicians with patient monitoring, and function as a communication tool [23]. Table 2 shows the ADAPT triage categories [24]. ADAPT stratify patients into four color-coded groups based on their vital signs and complaints. These groups are Red, Orange, Yellow, and Green. Red signifies more critical circumstances needing a higher priority and Green refers to less urgent conditions which can be managed with a lower priority [20].

Table 2. ADAPT triage model [4]

	1 Red Resuscitation 0 min	**2 Orange** Urgent 15 min	**3 Yellow** Less urgent 60 min	**4 Green** Not urgent 180 min
Airways	Obstructed airway stridor	threatened airway		
Breathing	$SpO_2 < 80\%$ $8 > RR > 35$	$80 \leq SpO_2 \leq 89$ $31 \leq RR \leq 35$	$90 \leq SpO_2 \leq 94$ $26 \leq RR \leq 30$	$SpO_2 \geq 95$ $8 \leq RR \leq 25$
Circulation	$HR > 140$ $SBP < 80$	$121 \leq HR \leq 140$ $HR < 40$ $80 < SBP < 89$	$111 \leq HR \leq 120$ $40 \leq HR \leq 49$	$50 \leq HR \leq 110$
Disability	$GCS \leq 8$	$9 \leq GCS \leq 13$	$GCS = 14$	$GCS = 15$
Exposure	$T_p < 32$	$T_p > 40$ $32 \leq T_p \leq 34$	$38 \leq T_p \leq 40$ $34.1 \leq T_p \leq 35$	$35 \leq T_p \leq 38$

In this project, ADAPT triage system was utilized to put patients into different categories according to their vital signs at arrival. The ADAPT system can also be used for scoring the severity of vital signs, so we calculated the ADAPT scores sequences based on patients' vital signs trajectories during hospitalization. Figure 2 shows the severity trajectory of a patient over a sample time period. These severity scores were calculated based on patient vital signs and ADAPT scoring system.

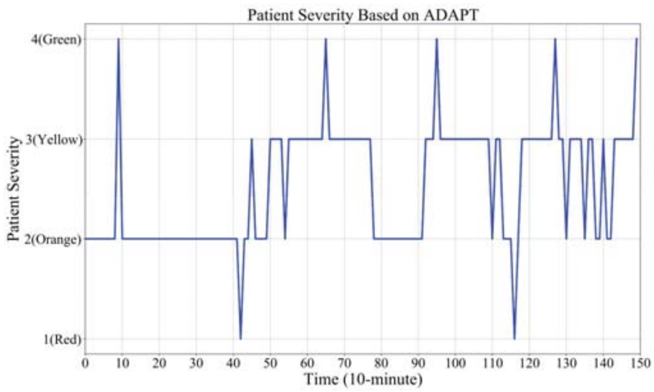

Figure 2. Illness severity of a patient based on ADAPT scoring system for a sample period.

D. *Model development*

1) *Nonlinear Autoregressive Exogenous model*

NARX [25] is a nonlinear autoregressive model with exogenous inputs meaning the current output of the model depends on 1) past and current values of input and 2) past values of output. It can be modeled as Equation 1.

$$y(t+k) = F(y(t), \ldots, y(t-p+1), \\ u_1(t), \ldots, u_1(t-q+1), \\ \ldots, u_m(t), \ldots, u_m(t-q+1)) + e(t) \quad (1)$$

Where *y(t)* is the target time series, u_1 to u_m are the exogenous input variables that are used for prediction. k is the prediction step and p and q are the autoregression order for the output and exogenous input order, respectively. $e(t)$ is the noise parameter and function F can be nonlinear function such as neural network, wavelet network, or any other type of regressors that predicts time series. In this study we used different type of regressors including RF, AdaBoost, Support Vector Regression (SVR), KNN, Linear Regression (LR), and Auto Regressive Integrated Moving Average (ARIMA) to predict the next values in a time series.

2) *Ensemble Learning*

Ensembles models are the state-of-the-art solution for various ML problems [26]. In ML and statistics, EL refers to building a model by generating and combining different individual predictive models to obtain a model with superior performance than each of the primary models. It is similar to the human decision process, where humans consider different opinions before taking important decisions, then weigh and combine these opinions to make the final decision [27].

The main idea behind using EL is the *no free lunch* theorem [28]. Based on this theorem, there is no algorithm that is the best for all tasks, so we can have a pool of models, for example, classifiers and the combination of classifiers provides the best solution. There are several advantages in using EL such as overfitting prevention, avoiding local optimal, and solving the curse of dimensionality [26].

In this study, RF and AdaBoost, were considered as ensemble learning algorithms. RF is a supervised learning algorithm for classification or regression. A RF consists of various numbers of Decision Trees (DT) where each tree uses a random portion of the original data. AdaBoost is a ML algorithm that can be used along with other ML techniques to improve the performance using an iterative process. It can be used in both classification and regression tasks, giving more focus to the patterns which are hard to detect [27]. There are some differences between RF and AdaBoost such as RF is based on bagging, but AdaBoost uses boosting techniques. In the bagging technique, several subsets of data are chosen randomly with replacement and each subset is used to train one model such as DT. The output is the average of all predictions from different single models. In case of boosting, models are trained sequentially with early models building simple models; then the error is analyzed. At each step, consecutive models are fit, and the aim is to increase the accuracy and decrease the error of the previous step. Moreover, each tree in RF utilizes all features, while AdaBoost uses one feature at a time in each tree. However, while both techniques are powerful in learning patterns; the probability of overfitting for AdaBoost is much higher than RF [27].

IV. Results

In this section, the performance of our proposed model (NARX-EL) is investigated. Six models were implemented to predict the illness severity of patients based on ADAPT scoring system over time. Various ML techniques were considered as F in Equation 1. In other words, the combination of ML algorithms as regressors with NARX was considered to predict the future severity of patients. As mentioned before, each EL model needs basic ML algorithms, so in this study, DT was considered as a basic model for learning AdaBoost. Moreover, the combinations of SVR, KNN, and LR as regression techniques and NARX were implemented. Finally, these models' performance was compared with ARIMA, which is a traditional statistical technique for time series forecasting.

There are 9,609 records in our dataset, and since we deal with time series, cross validation on a rolling basis [29] with five folds was applied. In this approach, to produce a better estimation of prediction error, many train/test splits are produced, and the average error over all splits is calculated. Moreover, every train block consists of two parts, including training subset where the model is trained on this part and a validation subset, which is used to find the best values for the model's parameters. The rolling cross validation process is shown in Figure 3.

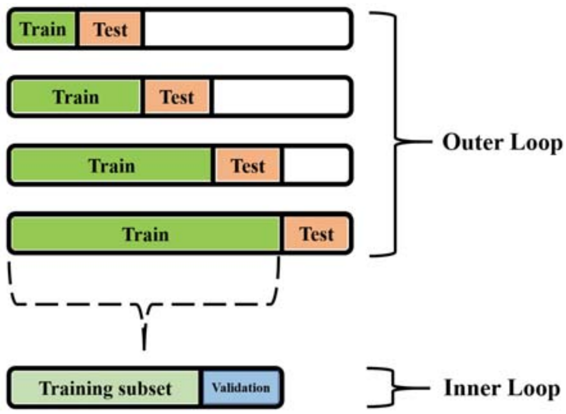

Figure 3. Cross validation on a rolling basis.

Grid search was done in the validation phase to find the optimal values for hyperparameters of models based on Mean Square Error (MSE) metric. Based on Equation 1, our models have the ability of multi-step prediction by setting variable k. In this study, we are interested in the prediction of patients' severity in the next hour based on the previous two hours. Therefore, as our time horizon is 10-minute, k, p and q should be set to 6, 12, and 12, respectively. The hyperparameters of models are presented in Table 3.

Table 3. Models hyperparameters

Model	Hyperparameters
ARIMA	*p = 12, q = 12, d = 2*
NARX-LR	*p = 12, q = 12*
NARX-RF	*p = 12, q = 12, number of DTs=50, Max depth=20*
NARX-AdaBoost	*p = 12, q = 12, Learners = DT, number of DT = 50, Max depth = 20*
NARX-SVR	*p = 12, q = 12, Kernel = RBF, C = 100, ε = 0.03 γ = 0.01*
NARX-KNN	*p = 12, q = 12, K = 4*

Based on Figure 3, the performance of all models was investigated. Two metrics, including the average R-squared (R^2) score and Normalized Root Mean Square Error

(NRMSE) were used for evaluation. R^2 score and NRMSE formula are shown in Equation 2.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$
$$NRMSE = \frac{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}}{y_{max}} \tag{2}$$

Where y_i and $\hat{y}_i$ are the actual value and the predicted value of a time series at time step i, respectively. $\bar{y}$ is the real mean value of time series and n indicates the number of samples in a timeseries. It should be noted that, R^2 score is a value between zero and one, which shows how much of the variation in a time series can be explained by the model. A closer R^2 score to one means a model is better at the estimating a time series. The performance of the models is shown in Table 4.

Table 4. Performance of models

Model	R^2	NRMSE (%)
ARIMA	0.812	19.13
NARX-LR	0.713	29.83
NARX-RF	0.978	6.16
NARX-AdaBoost	0.945	13.11
NARX-SVR	0.941	15.64
NARX-KNN	0.763	23.72

As shown in Table 4, NARX-RF had the best performance, followed by NARX-AdaBoost and NARX-SVR in terms of highest R^2 score and lowest NRMSE. Moreover, among NARX models, NARX-LR had the worst performance because this model tries to model a nonlinear system with a linear function.

To visualize the performance of models, a patient monitored for around 50h was selected. The predicted trajectories of patient's severity by different models are shown in Figure 4. The red curve in the figure shows the predicted severity of the patient based on ADAPT scores calculated using the patient's vital signs. It should be noted that the triage categories sequences in Figure 4 is categorical data, but this sequence was given to the models as a numerical time series. This explains why the predicted values are decimal and therefore must be rounded to the nearest category number however, to have a better visualization and comparison of models, the decimal outputs were plotted.

In Figure 4, three single peaks can be observed between the 100th and 250th 10-minute time intervals. Each of these peaks shows sudden changes in the condition of patients. It can be seen that NARX-RF, NARX-AdaBoost, and NARX-SVR can detect and follow these sudden changes more accurately.

V. Discussion

In this study, a hybrid model based on NARX and EL on the individual level for predicting patients' severity using vital signs (PR, RR, SpO2, and SBP) is proposed. The severity trajectories of patients indicate the level of care and monitoring they need. Such predictive systems are very crucial in the clinical environment because they help clinicians to identify patients who are more likely to deteriorate or experience adverse events in the future. Moreover, these predictive models could give clinicians enough time to intervene.

NARX have been used in medical research area such as prediction of glucose level in blood [30], forecasting daily patients arrivals in the EDs [31], hospital bed occupancy [32]. Moreover, EL also has shown great potential in building predictive models in clinical settings such as prediction of

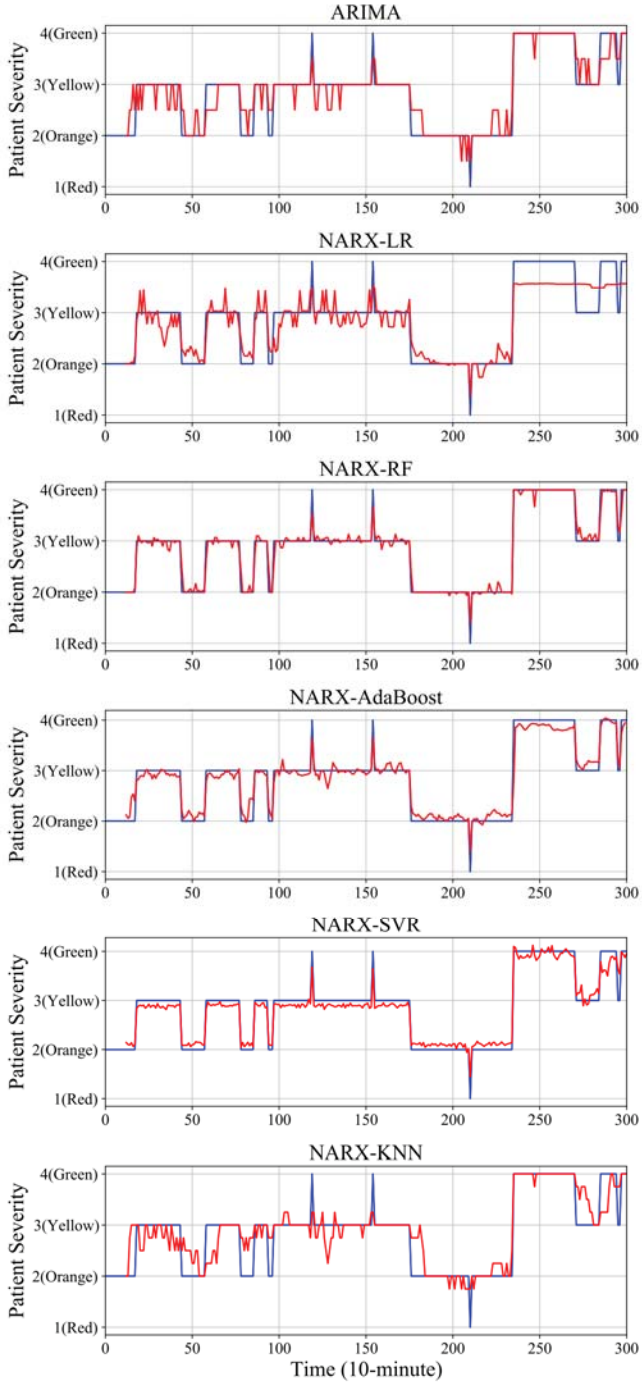

Figure 4. Prediction of patient severity trajectory (red curve) for a sample patient by different models (real trajectory: blue curve).

transfer to the pediatric ICU [33], and detection of brain diseases [34]. However, the ability of hybrid NARX-EL in prediction, especially in the clinical research area, has not been investigated. To the best of our knowledge, this is the first study that uses a combination of NARX and EL to predict the illness severity of patients during hospitalization.

Our proposed model has two different parts, including NARX and EL. NARX has a great capacity to model a wide range of nonlinear dynamic systems [35]. EL models consist of multiple models and have two significant benefits over single models: 1) better prediction and 2) more stable models. Moreover, these models are robust to noise and overfitting. The EL part of the proposed model includes RF and AdaBoost techniques. The performance of NARX-RF and NARX-AdaBoost was compared to ARIMA and combination of NARX and LR, SVR, KNN regression. The results showed that the NARX-RF and NARX-AdaBoost had superior performance in terms of R^2 score and NRMSE. These hybrid models can follow sudden changes in patients' condition, so it can be a useful tool to help clinicians identify unexpected adverse events and clinical deterioration in advance.

Since the personal models for patients' monitoring are computationally expensive, their development in real clinical settings has been limited. However, in this study, a model was proposed, which is fast in training and adjusting when new data is added to the system. In other words, the NARX-EL model, unlike neural network-based models, does not need too much time for training as well as no need for re-execution of the whole training process when new data is added to the system. Therefore, these favorable characteristics of NARX-EL model reveal its potential application in developing personal models for monitoring patients under real clinical conditions.

Scoring systems are prevalent at hospitals, and clinicians are using different scoring systems to categorize patients. These systems help clinicians to prioritize patients and stratify them into different groups based on their severity. In this study, ADAPT scoring system, which is utilized in several Danish hospitals, was used to stratify patients into four groups. Such scoring systems consider thresholds on vital signs based on which patients are divided into different categories. Sometimes deciding on putting a patient in a category is hard and clinicians are not sure which category to choose. This process is called hard decision making; so, to address this problem and consider the uncertainty in patients' health condition, soft decision-making techniques can be applied. One of the approaches is Fuzzy logic that could help us to make a better decision about the patient's condition. Fuzzy logic is based on the fact that human decisions are made based on imprecise information and uncertainties. In classical logic, one sample can belong to only one class; however, in Fuzzy logic, a sample can belong to different classes with different membership levels. Fuzzy logic has been used in various areas of medicine because of its multivariate nature [36]. Therefore, applying Fuzzy logic in our research area could lead to better performance in the prediction of patients' conditions in the future.

VI. CONCLUSION

The contribution of this paper is twofold: (1) this study is the first study on building a real-time hybrid model based on NARX and EL to predict patients' illness severity in ED using their vital signs (PR, RR, SpO2, and SBP). The proposed models outperform the commonly used ARIMA model. (2) The proposed model works at the individual level, which means each patient's characteristics are taken into account. In other words, after finding the best values of the model's parameters, for each patient, a specific model based on his/her vital signs trajectories is trained and used for prediction.

In this study, six models were considered to model and predict the nonlinear behavior of patients' vital signs and related ADAPT score trajectories during the hospitalization. The data was collected from OUH, and HSWJ and the performance of models on the prediction of patients' severity in the next hour based on the previous two hours was investigated. Moreover, cross validation on a rolling basis was applied to have a robust result and better estimation of models' performance. The results showed that NARX-RF (R^2 score = 0.978, NRMSE = 6.16%) and NARX-AdaBoost (R^2 score = 0.945, NRMSE = 13.11%) had the best performance and could follow the sudden changes and fluctuations in patients' severity more accurately.

In our future work, we will explore new ways to overcome the mentioned challenges, such as limitations of the ADAPT clinical scoring system. The developed model will then be used for seamless integration into clinical decision support systems for a more accurate prediction of patients' illness severity at hospitals.

VII. REFERENCES

1. Henriksen DP, Brabrand M, Lassen AT. Prognosis and risk factors for deterioration in patients admitted to a medical emergency department. PLoS One. 2014;9(4):e94649.
2. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med. 2016;44(2):368.
3. Kellett J, Murray A, Woodworth S, Huang W. Trends in weighted vital signs and the clinical course of 44,531 acutely ill medical patients while in hospital. Acute Med. 2015;14(1):3–9.
4. Schmidt T, Wiil UK. Identifying patients at risk of deterioration in the Joint Emergency Department. Cogn Technol Work. 2015;17(4):529–45.
5. Chatfield C, Xing H. The analysis of time series: an introduction with R. CRC press; 2019.
6. Meyfroidt G, Güiza F, Ramon J, Bruynooghe M. Machine learning techniques to examine large patient databases. Best Pract Res Clin Anaesthesiol. 2009;23(1):127–43.
7. de Schatz CHV, Schneider FK, Abatti PJ, Nievola JC. Fuzzy-NNARX based Tool for Monitoring and Predicting Patients Conditions using Selected Vital Signs. Int J Comput Sci Netw Secur. 2015;15(1):112.
8. Diaconescu E. The use of NARX neural networks to predict chaotic time series. Wseas Trans Comput Res. 2008;3(3):182–91.
9. Mansourvar M, Wiil UK, Nøhr C. Big Data Analytics in Healthcare: A Review of Opportunities and Challenges. In: International Conference for Emerging Technologies in Computing. Springer; 2020. p. 126–41.
10. Hu SB, Wong DJL, Correa A, Li N, Deng JC. Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model. PLoS One. 2016;11(8):e0161401.
11. Prytherch DR, Smith GB, Schmidt PE, Featherstone PI. ViEWS—towards a national early warning score for detecting adult inpatient deterioration. Resuscitation. 2010;81(8):932–7.
12. Clifton L, Clifton DA, Watkinson PJ, Tarassenko L. Identification of patient deterioration in vital-sign data using one-class support vector machines. In: 2011 federated conference on computer science and information systems (FedCSIS). IEEE; 2011. p. 125–31.
13. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. Qjm. 2001;94(10):521–6.
14. Ge W, Huh J-W, Park YR, Lee J-H, Kim Y-H, Turchin A. An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2018. p. 460.
15. Guiza Grandas F, Blockeel H, Bruynooghe M, Van Loon K, Aerts J-M, Berckmans D, et al. Time-series analysis techniques combined with Gaussian process classifiers for prediction of clinical stability after coronary bypass surgery. In: Proceedings of

the 6th IASTED International Conference on Biomedical Engineering. ACTA Press; 2008. p. 216–21.
16. Matam BR, Duncan H, Lowe D. Machine learning based framework to predict cardiac arrests in a paediatric intensive care unit. J Clin Monit Comput. 2018;1–12.
17. Li Q, Huang LF, Zhong J, Li L, Li Q, Hu J. Data-driven Discovery of a Sepsis Patients Severity Prediction in the ICU via Pre-training BiLSTM Networks. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2019. p. 668–73.
18. Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identification using patient similarity. AMIA Summits Transl Sci Proc. 2015;2015:132.
19. Higgins JP. Nonlinear systems in medicine. Yale J Biol Med. 2002;75(5–6):247.
20. Naemi A, Schmidt T, Mansourvar M, Wiil UK. Personalized Predictive Models for Identifying Clinical Deterioration Using LSTM in Emergency Departments. In: EFMI-STC; 2020. p. 152–156.
21. Hu Z, Melton GB, Arsoniadis EG, Wang Y, Kwaan MR, Simon GJ. Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. J Biomed Inform. 2017;68:112–20.
22. Thim T, Krarup NHV, Grove EL, Rohde CV, Løfgren B. Initial assessment and treatment with the Airway, Breathing, Circulation, Disability, Exposure (ABCDE) approach. Int J Gen Med. 2012;5:117.
23. Nordberg M, Lethvall S, Castrén M. The validity of the triage system ADAPT. Scand J Trauma Resusc Emerg Med. 2010;18(S1):P36.
24. Lauritzen, M., C. Skriver and JD. No Title. Triage-manual l. 2009.
25. Pham HT, Yang B-S. A hybrid of nonlinear autoregressive model with exogenous input and autoregressive moving average model for long-term machine state forecasting. Expert Syst Appl. 2010;37(4):3310–7.
26. Sagi O, Rokach L. Ensemble learning: A survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8(4):e1249.
27. Rokach L. Ensemble-based classifiers. Artif Intell Rev. 2010;33(1–2):1–39.
28. Wolpert DH. The supervised learning no-free-lunch theorems. In: Soft computing and industry. Springer; 2002. p. 25–42.
29. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. Inf Sci (Ny). 2012;191:192–213.
30. Assadi K, Hamdi T, Fnaiech F, Ginoux JM, Moreau E. Estimation of blood glucose levels techniques. In: 2017 International Conference on Smart, Monitored and Controlled Cities (SM2C). IEEE; 2017. p. 75–80.
31. Yucesan M, Mete S, Serin F, Celik E, Gul M. NARX Neural Networks Model for Forecasting Daily Patient Arrivals in the Emergency Department. In: Computational Intelligence and Soft Computing Applications in Healthcare Management Science. IGI Global; 2020. p. 1–18.
32. Kutafina E, Bechtold I, Kabino K, Jonas SM. Recursive neural networks in hospital bed occupancy forecasting. BMC Med Inform Decis Mak. 2019;19(1):39.
33. Rubin J, Potes C, Xu-Wilson M, Dong J, Rahman A, Nguyen H, et al. An ensemble boosting model for predicting transfer to the pediatric intensive care unit. Int J Med Inform. 2018;112:15–20.
34. Suk H-I, Lee S-W, Shen D, Initiative ADN. Deep ensemble learning of sparse regression models for brain disease diagnosis. Med Image Anal. 2017;37:101–13.
35. Chen S, Billings SA, Grant PM. Non-linear system identification using neural networks. Int J Control. 1990;51(6):1191–214.
36. Taheri F, Masoudi S, Soltani Z. Diagnosis of Cardiovascular Disease Using Fuzzy Methods in Nuclear Medicine Imaging. Arch Pharm Pract. 2019;10(4).