MASTER'S THESIS

Ali El Adi

Deep neural networks to forecast cardiac and respiratory deterioration of intensive care patients

UNIVERSITY OF TAMPERE
Faculty of Natural Sciences
Computational Big Data Analytics
Supervisor: Jaakko Peltonen

December 2018

University of Tampere

Faculty of Natural Sciences

El Adi, Ali: Deep neural networks to forecast cardiac and respiratory deterioration of intensive care patient

Master's thesis, 60 p.

Computational Big Data Analytics

January 2017

Abstract

Deep neural networks have proven valuable in several applications. The availability of electronic health records at high frequency has made it possible to provide real-time prediction to stay relevant to the user's immediate and changing context. This thesis implements deep neural networks for the prediction of short term cardiac and respiratory deterioration. It is based on the cardiac and respiratory SOFA sub-scores to define the event of deterioration, and it uses convolutional neural networks, long short-term memory and multitask learning to construct models that alert if the patient is prone to deterioration. Data from the FINNAKI study was used in training the predictive models, and a subset of the MIMIC III clinical database was used to investigate the applicability of those models in intensive care units from different locations. In terms of area under the ROC curve, the predictive models could achieve an area under score of 0.7812 from the FINNAKI data and 0.6816 for a subset of MIMIC III. Those results confirm that short-term deterioration is predictable which could help caregivers in focusing more on the patients at risk of deterioration in the short term.

Key words: Deep learning, convolutional neural networks, long short-term memory, multitask learning, sequential organ failure assessment score, cardiac and respiratory deterioration, intensive care unit.

Acknowledgements

I would like first to thank my teachers and colleagues for the valuable knowledge I acquired during my studies at the University of Tampere, especially my supervisor Jaakko Peltonen for his insightful comments all along the period of this thesis.

I would also like to thank my colleagues at GE Healthcare for sharing their experience and information, and making the workplace very welcoming. Particularly, I am extremely grateful to my company supervisor Mika Sarkela for his everyday support and helpful adivices. I also address special thanks to Rene Coffeng and Hanna Viertio-Oja.

Finally, I can't thank enough my parents and friends for their unlimited support that made me who I am today.

Contents

List of Figures

List of Tables

List of abbreviations

ICU :	Intensive Care Unit
MIMIC:	Medical Information Mart for Intensive Care
FINNAKI:	Finnish Acute Kidney Injury
SOFA:	Sequential Organ Failure Assessment
LODS:	Logistic Organ Dysfunction System
CNN :	Convolutional Neural Network
RNN :	Recurrent neural network
LSTM :	Long Short-Term Memory
k-NN :	k-Nearest Neighbors algorithm
RF :	Random Forest
ROC :	Receiver Operating Characteristic
AUC :	Area Under the ROC Curve
NPV :	Negative Predictive Value
PPV :	Positive Predictive Value

Chapter 1

Introduction

A lot of data is collected about patients during their hospital stay for clinicians to make decisions on appropriate treatment. As more parameters are added to the medical records and measurements are made more frequently, caregivers can gain more insight into the patient situation. On the other hand, it becomes more difficult for health care professionals to process big amounts of data in a limited amount of time. This is why clinical decision support systems are useful for health caregivers in many decision making tasks such as diagnosis, therapy planning and monitoring. Based on statistics and machine learning, several clinical decision support systems emerged as early as in the 1970s [2]. Particularly in the intensive care unit (ICU), ICU scoring systems give an assessment of the patient's health status in the form of a score [3], and their relationship with mortality has been investigated by many researchers [4]. Usually, the worst score within 24 hours after admission is considered in estimating the probability of survival (prognosis), thereby predicting the patient outcome [3] [5].

The research on ICU scoring systems has tended to focus on mortality prediction rather than on their potential in monitoring and predicting the evolution of a patient's health status over time. Yet, as many ICU scores refer to the clinical severity of the intensive care patient, the outcome of forecasting ICU scores is important in that it can help caregivers to be aware of patients at risk so that they take appropriate measures in advance to prevent these patients from deteriorating. Patients in the

ICU require constant attention and close watch. Indeed, studies show an association of the ratio of nurse staffing in the ICU with patient outcomes [6]. By restricting the focus to the patients who are likely to be unstable imminently, it is possible to improve services in the ICU and prevent some adverse patient outcomes. Furthermore, predicting which patients are deteriorating not only helps doctors in their decision-making, but also in their decision towards stable patients in discharging them, or transferring them to the ward which requires less attention from the staff. For these reasons, a model that could indicate an ICU score of the patient in advance seems beneficial in the ICU.

The availability of electronic health records has made this kind of study possible. In this thesis, the general objective is to forecast the medical state of adult patients in the intensive care unit. This study focuses on the prediction of cardiac and respiratory organ dysfunctions. Using the sequential organ failure assessment score (SOFA), thresholds are set for defining cardiac and respiratory deterioration so as to simplify the problem into a binary classification. This research investigates different parameters on which the target may depend, and explores different deep neural networks. The datasets involved in training and testing the predictive models are also described and compared to answer questions related to the applicability of predictive models in different locations.

Chapter 2

Background

2.1 ICU scoring systems

There are several ICU scoring systems which aim to give an insight into the medical state of intensive care patients. [3] Although these scoring systems share a general objective, they differ in the choice of input parameters and the method for computing their scores so as to show specific patient outcomes. For instance, Logistic Organ Dysfunction System (LODS) and Sequential Organ Failure Assessment (SOFA) are designed to detect organ dysfunctions, whereas Simplified Acute Physiology Score (SAPS) and Acute Physiology And Chronic Health Evaluation (APACHE) are severity of disease classification systems [3]. In general, the scores are ordinal: higher scores correspond to higher patient severity, and they are computed based on several parameters such as measurements of vital signs and laboratory test results.

2.2 Sequential Organ Failure Assessment (SOFA) score

Sequential organ failure assessment score, previously known as Sepsis-related organ failure assessment score, indicates the degree of an intensive care patient's organ dysfunction [7]. According to previous studies [5] [8], SOFA has a good predictive power

for hospital mortality prediction and shows a high correlation with patient outcome. The overall score is a sum of six different scores representing the cardiovascular, respiratory, renal, coagulation, hepatic and neurological systems. Hence, it is possible to track each of these biological systems on their own, which enables selectively considering certain organ dysfunctions. Further in this thesis, the cardiovascular and respiratory sub-scores of SOFA are used to determine cardiac and respiratory deterioration. The tables 2.1 and 2.2 list conditions and their corresponding sub-score for the cardiovascular and respiratory systems respectively. The cardiac SOFA sub-score depends on the value of the mean arterial pressure and the administrated doses of vasopressors (dopamine, dobutamine, epinephrine and norepinephrine), whereas the respiratory SOFA sub-score depends on whether the patient is under mechanical ventilation, and on the ratio of the partial pressure of oxygen (PaO2) measurement to the fraction of inspired oxygen (FiO2).

Conditions on the mean arterial pressure (MAP) and doses of vasopressors	**Cardiovascular SOFA sub-score**
MAP $\geq$ 70 mmHg	0
MAP $<$ 70 mmHg	1
dopamine $\leq$ 5 μg/kg/min OR dobutamine $>$ 0 μg/kg/min	2
dopamine $>$ 5 μg/kg/min OR norepinephrine $\leq$ 0.1 μg/kg/min OR epinephrine $\leq$ 0.1 μg/kg/min	3
dopamine $>$ 15 μg/kg/min OR norepinephrine $>$ 0.1 μg/kg/min OR epinephrine $>$ 0.1 μg/kg/min	4

Table 2.1: Cardiovascular SOFA sub-score table

Conditions on the $PaO2/FiO2$ (mmHg) and mechanical ventilation	**Respiratory SOFA sub-score**
$PaO2/FiO2 \geq 400$	0
$PaO2/FiO2 < 400$	1
$PaO2/FiO2 < 300$	2
$PaO2/FiO2 < 200$ AND mechanically ventilated	3
$PaO2/FiO2 < 100$ AND mechanically ventilated	4

Table 2.2: Respiratory SOFA sub-score table

2.3 Machine learning

As early as in 1959, Arthur Samuel defined Machine learning as a "field of study that gives computers the ability to learn without being explicitly programmed" [9]. Today, machine learning is vastly applied in teaching computer systems from data to perform a specific task without human intervention. There are several approaches in machine learning, and the choice over the machine learning methods mainly depends on the task and the data-set. Generally, machine learning tasks are categorized into unsupervised, supervised, and semi-supervised learning [10]. Unsupervised learning algorithms draw inferences from unlabeled data. Tasks in this category include clustering which divides a set of objects into groups of similar objects. On the other hand, supervised learning algorithms learn functions from labeled data-sets that map an observation to its desired output. In this category, classification tasks relate to categorical output values whereas regression tasks concern continuous responses. The prediction problem in this thesis falls into the category of supervised learning, and is a classification problem since the output of the prediction is a class label. Semi-supervised learning involves both unlabeled and labeled data in training the model. Since our prediction task involves deep neural networks in the construction of the models, particularly convolutional neural networks (CNN) and Long Short Term Memory networks (LSTM), this subsection presents the principles of neural network and deep learning, then explains how CNNs and LSTMs specifically process temporal data, and finally introduces multitask learning.

2.3.1 Artificial neural networks

Artificial neural networks are computational models of similar structure to biological neural networks [11]. They consist of a collection of connected nodes called artificial neurons, and can be depicted by a directed graph. As illustrated in the figure 2-1, a neuron is a computational unit that receives inputs from other neurons or data and maps them to an output that is then conveyed to its connected neurons. It uses a function of a linear combination of the inputs to calculate the output, as expressed in the following equation:

$$o_w(x) = \phi(w.x) = \phi(\sum_i w_i * x_i) \tag{2.1}$$

where ϕ is called the activation function. It is usually a sigmoid, a hyperbolic tangent, or a rectified linear unit.

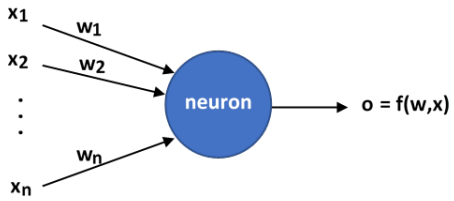


Figure 2-1: An artificial neuron

Typically, neural networks are feedforward, i.e. the connections between neurons do not form cycles. A basic form of feedforward neural networks is the multilayer perceptron (MLP) or fully feedforward neural network in which neurons form a sequence of layers that each have neurons connected to the neurons in the subsequent layer. A multilayer perceptron is composed of an input layer and an output layer related by an arbitrary number of hidden layers in between. The input layer only transmits the input data to the network without transformation, while each one of the other layers receives the outputs of their prior layer and applies a transformation at the level of each neuron. The figure 2-2 illustrates an example of a multilayer perceptron with 3 layers.

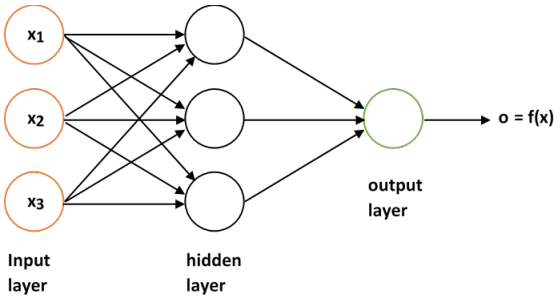


Figure 2-2: Example of a multilayer perceptron

Training a neural network consists of adjusting the weights in every neuron so as to minimize a cost function, also called loss function, that expresses the errors between the predictions and the desired outputs. One of the most common methods for training is the backpropagation algorithm, also called backward propagation of errors. It uses a gradient descent procedure to search for the solution along the direction given by the partial derivatives of the error with respect to the weights.

2.3.2 Convolutional neural networks

A convolutional neural network (CNNs) is a class of deep neural networks used for processing data of temporal or spatial structure in which the arrangement of inputs is important. In practice, CNNs are popular for their successful applications in video and image recognition, but they have also been applied in other application areas such as recommender systems [12], natural language processing [13] and healthcare science [14], many of which are examples of using CNNs on time series data. A CNN is typically characterized by 2 layer types: Convolutional layers and Max-pool layers.

Convolutional layers

For temporal input features, a convolutional layer creates a set of filters (kernels) that are convolved with the layer input over time to extract new features and, as a result, produce feature maps. Mathematically, a convolution is the integral of the product of 2 functions after reversing and shifting one of them, as shown in the formula for 2

functions x and w:

$$s(t) = (x * w)(t) = \int_{-\infty}^{+\infty} x(\tau)w(t-\tau)d\tau \tag{2.2}$$

In machine learning, the actual operation is implemented as a (discrete) matrix operation. For an input layer W with N features and a kernel X of size M, the applied form of the convolution over the time axis is:

$$S(t) = (W * X)(t) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(t+m, n)W(m, n) \tag{2.3}$$

The figure 2-3 illustrates an example of the operation applied to an input layer with features a, b, c and d, and one kernel of size 3.

Following the convolution, a bias b is added to the output of the convolution, then an activation function is applied. In convolutional layers, a commonly used activation function is the rectified linear unit (ReLu): $\phi(x) = max(0, x)$. Finally, the learning process in a convolutional layer consists of adjusting the weights and the biases so as to optimize an error function.

Defining a convolutional layer requires setting some hyperparameters, most importantly the kernel size, depth, stride, and padding.

- The kernel size specifying the length of the 1D convolution window
- The depth indicating the number of filters, each of which is assumed to extract a different feature map.
- The stride specifying the stride length of the convolution, that is the number of steps with which we slide the filters.
- The choice of the padding, that is whether to pad the input with zeros beyond the borders such that the output of the convolution has the same length as the original input.

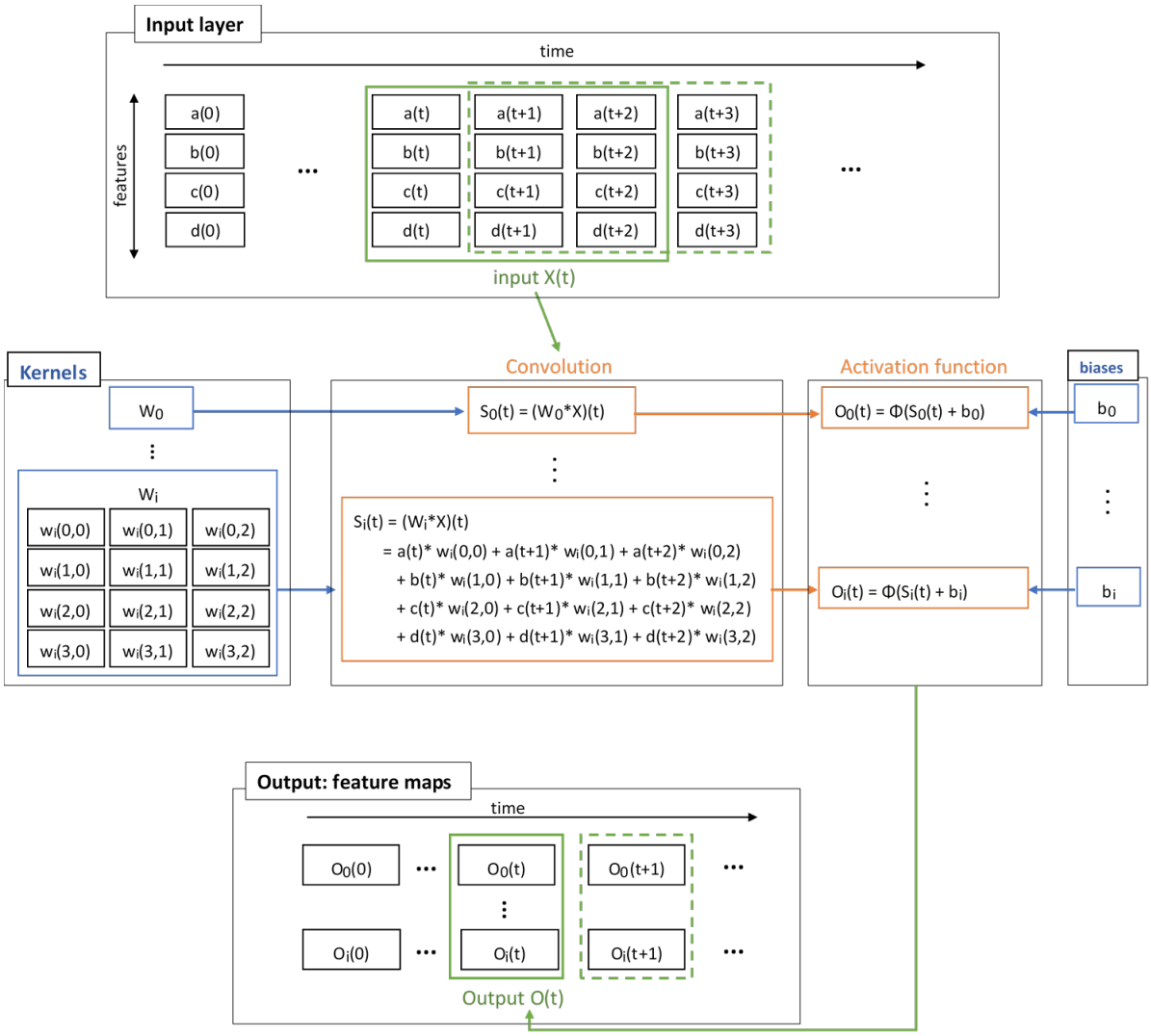


Figure 2-3: Example of CNN

Max-pool layers

Max-pool layers are often inserted between two successive convolutional layers. Their function is to decrease the size of feature maps to reduce the number of parameters in the network, thereby reducing computation and controlling overfitting [15]. A max-pool layer simply downsamples the output from the prior layer dividing the outputs into regions and by taking the maximum in each region, and therefore does not require training. Despite their benefit to the CNN, some research suggests other alternatives to max-pool layers, such as using larger strides in convolutional layers to control overfitting [16].

Dropout layers (in training only)

As a large amount of weights is trained in a deep neural network, there is a considerable risk of overfitting. One simple technique to prevent this problem is to use dropout layers during the training process [17]. This layer drops a random set of units from the output of previous layer during training, by setting them to zero.

2.3.3 Long Short Term Memory networks (LSTM)

A recurrent neural network (RNN) is another class of deep neural networks whose particularity is that its output in the past affects its output in the present. That is, for a time sequence, an RNN takes as input not only new data, but also a feedback of its previous outputs, called the hidden state. It can represented by a network with a loop or a chain of repeating units, as illustrated in the figure 2-4. A standard RNN unit has a simple structure such as a tanh activation layer.

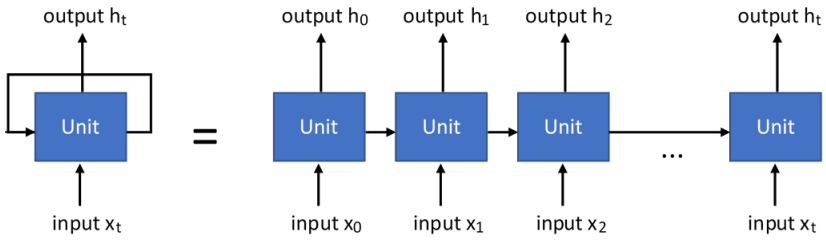


Figure 2-4: Recurrent neural network structure [1]

RNNs have particularly shown to be successful in natural language processing [18]. However, the standard RNN seem hardly capable of learning long-term dependencies due to the vanishing gradient problem [19]. That is why Hochreiter and Schmidhuber presented the long short term memory network (LSTM) [20], a variant of RNN that solves the problem by introducing another input to the RNN unit, called the cell state or memory. Besides, a typical LSTM unit is composed of three structures, called gates: an input gate, an output gate and a forget gate. The figure 2-5 illustrates the gates in an LSTM. Given a hidden state h_t and a cell state c_t at index t-1, and an input x_t at index t, the operations inside the gates are non-linear transformations of

vectors by tanh activation functions $tanh$ and sigmoid functions σ, and resulting from matrix multiplications. The output vectors from the gates are then used in pointwise multiplications in order to update of the cell state and the hidden state.

- From the input gate: $i_t = \sigma(W^i x_t + U^i h_{t-1} + b^i)$
- From the forget gate: $f_t = \sigma(W^f x_t + U^f h_{t-1} + b^f)$
- From the output gate: $o_t = \sigma(W^o x_t + U^o h_{t-1} + b^o)$
- A new memory cell is created: $\tilde{c}_t = tanh(W^c x_t + U^c h_{t-1} + b^c)$
- As a result:
  - The cell state at t is: $c_t = f_t.c_{t-1} + i_t.\tilde{c}_t$
  - The hidden state at t is: $h_t = o_t.tanh(c_t)$

  where . is the pointwise multiplication

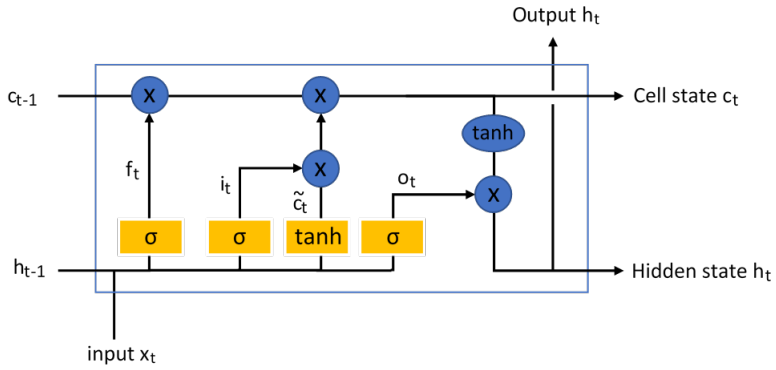


Figure 2-5: LSTM unit [1]

Finally, the learning process in an LSTM layer consists of adjusting the weights and the biases in the gates so as to optimize an error function.

Chapter 3

Literature Review

This chapter gives an overview of the studies related to the prediction of the patient future state based on available clinical data, ranging from detecting life-threatening conditions (e.g. sepsis, arrhythmia) to estimating the risk of death.

As for mortality prediction, researchers aimed to predict whether a patient dies during their stay at the ICU, and expanded research on predicting mortality over a specific time after ICU discharge. That is, given the medical record of a patient (generally during the first day of stay in the ICU), literature suggests various algorithms and aggregate functions which output a probability that the patient dies within a specified period of time. For instance, ICU scoring systems have demonstrated good predictive powers in mortality prediction [5]. Moreover, they have provided a basis of comparison for other mortality prediction models such as the Super ICU Learner Algorithm (SICULA): an ensemble machine learning technique combining regression models, classification trees and neural networks [21]. Those algorithms and scoring systems describe well how critical the patient's health is. Yet, they do not indicate changes in the patient state and most of them ignore the temporal characteristics of their input features. Besides, mortality prediction models seem to express unconditional fatality in that their estimated probability of death is independent of any potential medical treatment following the prediction. In contrast, the thesis outcome provides actionable prediction that can be updated.

One life-threatening condition that is closely related to the topic of research is sepsis. According to the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3), it is a "life-threatening organ dysfunction caused by a dysregulated host response to infection" [22] in which the organ dysfunction is indicated by an increase in the SOFA score by 2 points. Insight is a machine learning classifier that extracts changes and correlations of vital sign measurements to predict sepsis with better performance than existing methods such as the SIRS criteria which is of poor specificity [23]. Different deep neural networks such as LSTMs have also been useful at the early detection of sepsis. [24].

Monitoring and forecasting patient deterioration in the ICU is of interest in research. For instance, M. Wu et al. dealt with predicting the onset of vasopressor intervention in the ICU [25] which corresponds to a cardiac SOFA sub-score of at least 2. Their method consists of predicting the current values of the observed variables (vital signs and laboratory measurements) from their past values using a switching-state autoregressive model (SSAM). This model also learns latent variables informing of the physiological state of the patient. Then given these variables, a classifier (RF or Gaussian Naive Bayes (NB)) predicts the need for vasopressor administration and the weaning. This achieved an AUC of 0.92 for imminent vasopressor need predictions (i.e vasopressor need within 2 hours) and an AUC of 0.88 for short-term need prediction (i.e vasopressor need within a 2 hour window after a 4 hour gap) using the last four hours of patient data. Another example is a recent study [26] of Harini et al. that implemented LSTMs and convolutional neural networks to forecast the need for different clinical interventions including the administration of vasopressors and mechanical ventilation . A new representation of physiological data in the form of categories was also tested to address the problem of class imbalance. Using a subset of MIMIC III patients, the predictions were made for a window of 4 hours after a gap time of 6 hours. The study resulted in an AUC of 0.75 for the prediction of invasive ventilation intervention using LSTM and categorized physiological data, and 0.77 for the prediction of vasopressor intervention using LSTM or CNN.

In more restricted cohort studies, Fialho et al. [27] predicted vasopressor administra-

tion within the 2 following hours for patients receiving fluid resuscitation using fuzzy rule-based models. A model applicable to all that population achieved an AUC of 0.79 while disease-based models had AUCs of 0.82 for patients with pneumonia and 0.83 for patients with pancreatitis. Similarly, Salgado et al. [28] built an ensemble fuzzy model that predicts the need for administrating vasopressors in septic shock patients with an AUC of 0.85.

On the other hand, Crump et al [29] worked on predicting decline in the patient's condition by detecting abnormal deviations of vital signs from population norms or personal baselines. They used bayesian networks and rule-based trending.

Most of the above-mentioned works in monitoring cardiac and respiratory deterioration consider clinical intervention as a sign of deterioration in electronic health records. In conparison, this thesis formulates the detection of such types of deterioration from thresholds in the SOFA scoring system which also considers clinical interventions. Vasopressor administration corresponds to cardiac SOFA sub-scores of more than 1, and mechanical ventilation leads to respiratory SOFA sub-scores of more than 2 when the $PaO2/FiO2$ ratio is less than 200 mmHg. In contrast, this thesis detects deterioration when the cardiac SOFA sub-score is more than 0 or the respiratory SOFA sub-score is more than 2. Hence, cardiac deterioration here is additionally referred to the condition when the mean arterial pressure is less than 70 mmHg, and the respiratory deterioration no longer includes the condition when $PaO2/FiO2$ is more than 200 mmHg. Based on experts opinion, this difference relies on the fact that clinicians decide on the starting time of vasopressor administration that can be later than the optimal time. Then incorporating an additional condition on mean arterial pressure aims to detect cardiac deterioration closer to the optimal time for intervention. Since early intervention and weaning could influence the patient outcome [30] [31], the additional condition for cardiac deterioration and the further restriction in respiratory deterioration should mitigate the risk of late intervention and weaning in the data.

In parallel with this thesis, another study aimed at predicting short term patient state in the ICU characterized by the change in the cardiac and pulmonary LODS score [32].

It developed a convolutional neural network (CNN) that outputs the probability of a high cardiac or pulmonary LODS sub-score in the next 3 hours with 77% sensitivity and 64% positive predictive value.

The Logistic Organ Dysfunction system (LODS) is also a scoring system for assessing severity levels for organ dysfunction in the ICU. Unlike SOFA, LODS does not incorporate vasopressor administration in the calculation of its cardiac sub-score, but rather relies on the heart rate and the systolic blood pressure. The outcome does not differentiate patients under vasopressor administration by the cardiac and hemodynamic signs and does not take advantage of the physician's decision-making in predicting cardiac deterioration. Similarly, this thesis implements CNNs. Yet it explores further other architectures of deep learning with recurrent networks and multitask learning, and it compares to other machine learning methods. The prediction performance is primarily measured by the AUC instead of sensitivity and positive predictive value as in the literature, and an analysis of the feature importance is eventually made. In sum, this thesis presents a different prediction task, tests various machine learning techniques and offers a deeper analysis of the performance results.

It is worth noting that most of the aforementioned studies predicting deterioration in the ICU rely on hourly sampled data to make predictions and at a history window of at least 5 hours. At this sample rate, their predictive models may overlook the frequency characteristics of the vital signs, which could pose a limitation. In comparison, our work investigates the prediction task using more frequently sampled physiological data and shorter history window. The time scope of the prediction is another difference as each research paper attempts to predict patient decline during different time periods in the future from different time periods of the past. That is why the related studies are difficult to compare.

Chapter 4

Material and Methods

This section presents the data-sets used in this study. Then it defines the prediction task, and describes the neural networks that were tested along with the preprocessing, learning and evaluation steps.

4.1 Data

This thesis work relies on two sources of data: MIMIC III critical care database and FINNAKI. Further in this chapter, they are introduced, described and compared with each other with respect to time frequency and SOFA distribution.

4.1.1 FINNAKI

The main data-set originates from the prospective observational Finnish Acute Kidney Injury (FINNAKI) study [33]. It comprises deidentified health-related data associated with over 2900 adult patients ($>$ 18 years) and is collected from 17 intensive care units in Finland between September 2011 and February 2012. It includes demographic details, vital sign measurements, and laboratory test results.

4.1.2 MIMIC III

MIMIC III (Medical Information Mart for Intensive Care III) is a freely available database collected from the critical care units of the Beth Israel Deaconess Medical Center (in Boston, Massachusetts, USA) between 2001 and 2012 [34]. It contains deidentified data associated with over 40 000 patients in 2 forms of data: a relational database (MIMIC III Clinical database) and a waveform database. As for this thesis, the MIMIC III Clinical database will be used as it includes the information needed for the calculation SOFA scores.

4.2 Features

From the FINNAKI and the MIMIC III clinical databases the following classes of data were extracted:

Vital signs

Measurements made at the bedside include the heart rate, respiratory rate, temperature, blood oxygen saturation (SpO2), fraction of inspired oxygen (FiO2), and arterial blood pressure - systolic (SAPS), diastolic (SAPD) and mean (SAPM).

Vasopressors

Vasopressors are a class of drugs that induce the constriction of blood vessels, and as a result elevate arterial pressure. In particular, the computation of the Cardiac SOFA sub-score requires to know the dose of dopamine, dobutamine, epinephrine and norepinephrine.

Blood gas data

An arterial-blood gas test measures the amounts of arterial gases, such as partial pressure of oxygen (PaO2) and carbon dioxide (PaCO2). Blood sample data also includes pH and lactate.

Descriptive information

Demographic information such as the age and gender are among the input parameters of the predictive models.

Severity scoring systems

Other ICU scoring systems are considered such as the APACHE II morbidity score or the Glascow coma scale that provides the status for the central nervous system.

4.3 Predictive task

The overall objective is to predict cardiac and respiratory deterioration for intensive care patients. Specifically for a patient at the time of prediction, the task is to predict if any deterioration occurs within the next 3 hours, given a history of its health record during the past 2 hours. Along the patient's stay, the prediction is updated every 3 minutes.

In terms of the SOFA scoring system, thresholds are set to formulate cardiac and respiratory deterioration. Deterioration is said to occur when the cardiac SOFA sub-score turns to a value that is greater than 0, or when the respiratory SOFA sub-score becomes greater than 2. The problem is then reduced to a binary classification, wherein the class 0 refers to the category of cases in which the patient incurs no deterioration for 3 hours after the time of prediction, whereas the class 1 refers to the opposite category (i.e. deterioration occurs within the prediction window). This is equivalent to label samples as 1 or 0 depending on whether the maximum cardiac SOFA sub-score over the prediction time window is greater than 0, or the maximum respiratory SOFA sub-score over that time window is greater than 2.

No deterioration (0)	**Deterioration (1)**
cardiac SOFA $= 0$ AND respiratory SOFA ≤ 2	cardiac SOFA > 0 OR respiratory SOFA > 2

Table 4.1: Thresholds on SOFA for deterioration detection

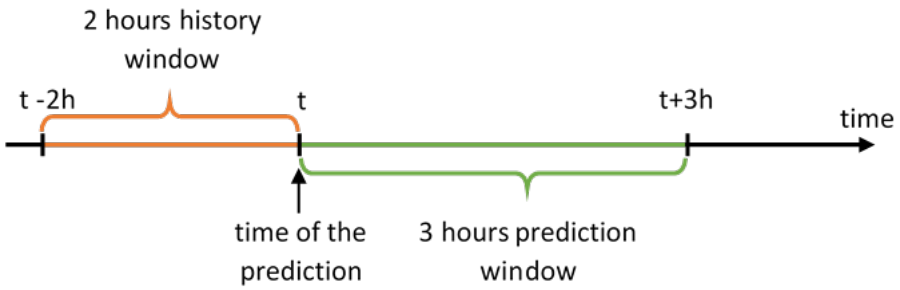


Figure 4-1: Predictive task

4.4 Description of the models

4.4.1 First CNN model description

Input features

A set of features were deemed crucial from the opinion of clinicians. It consists of the 12 following parameters:

- **Vital signs**: heart rate (HR), respiratory rate (RR), rectal temperature (TRECT), oxygen saturation (SpO2), mean arterial pressure (MAP), systolic arterial pressure (SAPS), diastolic arterial pressure (SAPD) and fraction of inspired oxygen (FiO2).
- **Arterial blood gas**: partial pressure of oxygen (PaO2), partial pressure of carbon dioxide (PaCO2)
- **ICU scoring systems**: cardiac LODS sub-score
- **Others**: time of the day

Architecture

This model is a sequential stack of 3 convolutional layers and 2 dense layer, using 12 filters of size 4 in the first convolutional layer, 12 filters of size 3 in the second convolutional layer, and 24 filters of size 3 in the third convolutional layer. The Rectified Linear Unit (ReLU) is used as the activation function following convolution,

producing zero when x $<$ 0 and then linear with slope 1 when x $>$ 0. A zero padding is applied beyond the borders of the input such that the output of the convolution has the same length as the original input. Then, each convolutional layer is followed by a max pooling layer with a pooling size of 2. During the training, dropout layers are applied after each max pooling layers, dropping randomly 20% from the first and second max pooling layers, and 50% from the third. An additional dense layer of 100 nodes is placed as the last hidden layer. Finally, a softmax function renders probabilities. The architecture of the first model is illustrated in the figure 4-2

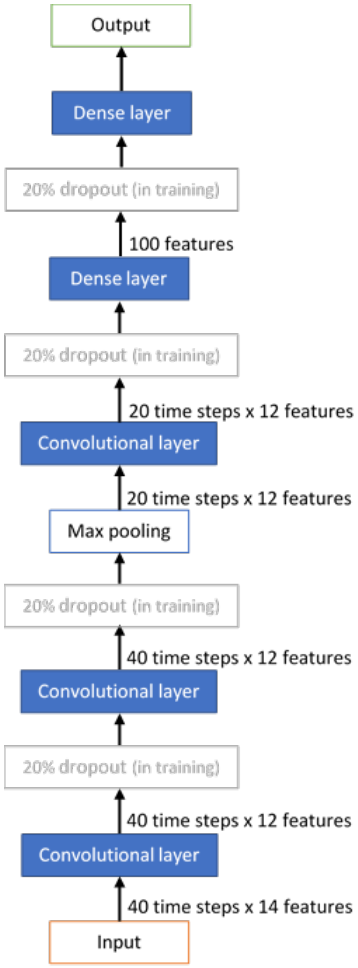


Figure 4-2: 1st model architecture

4.4.2 Second CNN model description

Input features

Additional features were later proposed by experts in an attempt to improve prediction. They were incorporated in this CNN model. Those 32 parameters include:

- **blood sample data**: pH, lactate, hemoglobin (HB), glucose, bilirubin, urea, creatinine, potassium (K), sodium (Na), white blood cell count (WBC), C-reactive protein (CRP), Base excess (BE), and central venous pressure (CVP)
- **Ventilator settings**: Positive end-expiratory pressure (PEEP)
- **Drugs**: Drugs: Norepinephrine (NOR), epinephrine (EPI), dopamine (DOP), and dobutamine (DOB), Nonsteroidal anti-inflammatory drug and steroids
- **ICU scoring systems**: Cardiac LODS subscore, cardiac and pulmunory SOFA subscores, Glascow coma scale, APACHE score, APACHE II score, and chronic health status scores
- **Others**: AIDS, acute and chronic liver disease, patient isolation, height and duration of stay.

Architecture

As a result of including more input features, the number of filters in the convolutional layers is increased. The first convolutional layer uses 20 filters of size 3, the second uses 40 filters of size 3, and the third uses 60 filters of size 3. Another difference from the first CNN model is that a max pooling layer is applied after the third convolutional layer, with the aim of reducing the number of weights after introducing more filters into the model. The figure 4-3 illustrates the architecture of the second model.

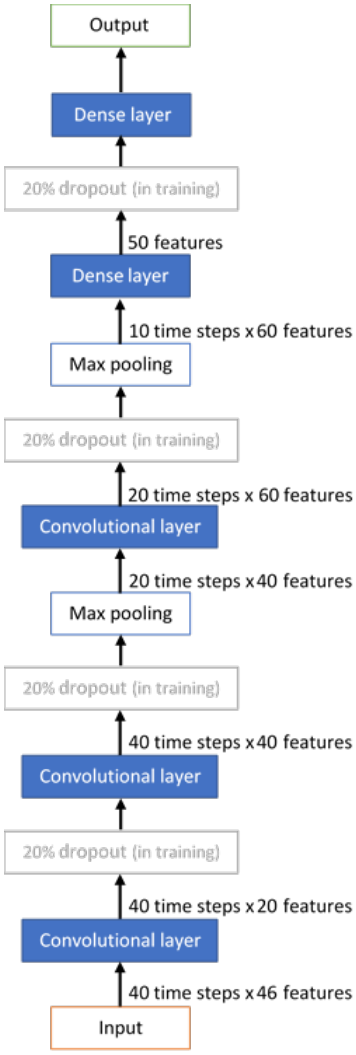


Figure 4-3: 2nd model architecture

4.4.3 Third model description

The third model is different from the previously introduced ones in that it involves Long short-term memory recurrent networks and multitask learning. This model relies on the same 14 parameters as the first introduced CNN model.

Architecture

The architecture of the third model is a combination of convolutional neural networks and long short term memory networks. It involves using convolutional layers for feature extraction on input data followed by LSTM layers to process the extracted

sequences of features with the help of their internal memory. This option is motivated by a similar approach in speech recognition [35], which also deal with temporal signals. It is based on the premise that convolutional layers are capable of learning sequences of hidden patterns in the signal (e.g. spectral features) while LSTM layers can analyze temporal dependency in the generated sequences.

Multitask learning

Since our target is a combination of 2 sub-scores, one could first learn to predict the cardiac and respiratory sub-scores then simply deduce the overall score. Unless it results in worse prediction performance, predicting cardiac and respiratory separately brings more insight into the future patient state from the clinician perspective. To that end, the main prediction problem breaks down into 2 sub-tasks regarding each of cardiac and respiratory deterioration. A typical approach is to learn each sub-task at a time, generating 2 separate predictive models. However, this approach does not leverage commonalities across the sub-tasks as their respective models are trained separately. In contrast, multitask learning is an approach in which tasks are learned jointly. As explained by Rich Caruana [36], using domain information that is underlying related tasks as an inductive bias leads to a better generalization. In practice, a multitask learning approach creates a shared representation for all tasks, and learning them is done in parallel.

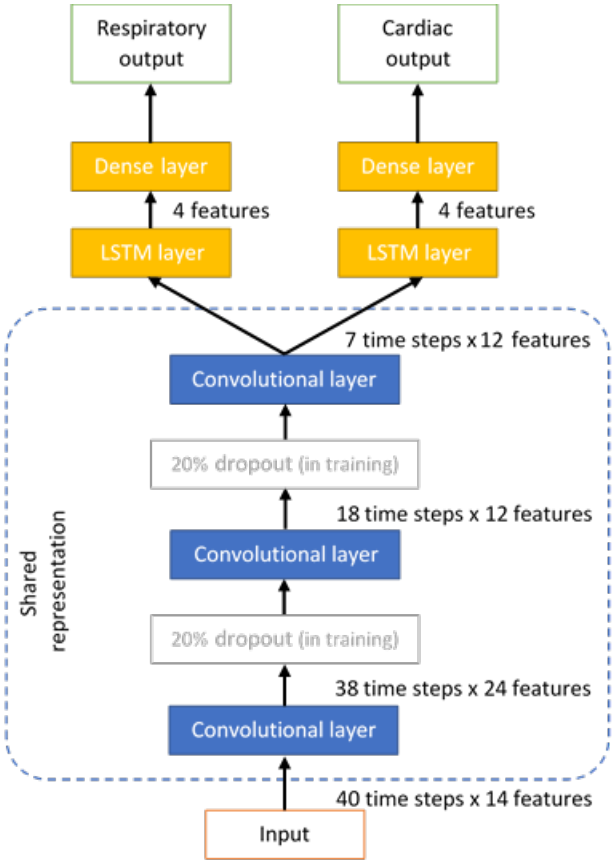


Figure 4-4: 3rd model architecture

4.5 Preprocessing

The raw data-set necessitates some preprocessing steps in order to tackle the challenges of irregular sample rates, outliers, and missing data. The time interval between two successive measurements of the same feature varies throughout the patient records, and different features are given at different sample rates. For instance, a time analysis of the heart rate (HR) and the systolic arterial pressure (SAPS) in FINNAKI, which are some of the frequently sampled variables, gives the following statistics of time intervals between successive measurements (in minutes):

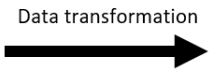
Feature	Mean	Standard deviation	Median	Minimum	Maximum	25^{th} percentile	75^{th} percentile
HR	2.21	7.89	2	0	10151	1	2
SAPS	2.21	6.02	2.07	0	8625	1.55	2.82

Table 4.2: Sampling rate statistics in minutes

Data resampling

In order for the predictive models to process the data, data variables were resampled and discretized so that one and only one value for each feature represents a sampling interval. The choice of the appropriate sampling rate is constrained by a trade-off between preserving all the data information and reducing the dimensionality of the input for predictive models. That is, using a low sampling rate (e.g. one sample per 10 minutes), unlike a high sampling rate (e.g. one sample per minute), will require aggregating values belonging to the same sampling interval. However, higher sampling rates induce smaller sampling intervals that require imputation when observations are missing. As a result, the input of the models will include more variables taking redundant values. Finally, after a time analysis of the frequent vital signs, a 3 minutes sampling rate is adopted. The patient ICU stay was divided into 3 min intervals, measurement timestamps were rounded up to the later side of the interval, and the average of measurements within the same interval was taken. An example of the transformation is shown in the figure 4-5.

Timestamp	Observation
21:02:18	o1
21:04:35	o2
21:05:21	o3
21:10:40	o4



Timestamp	Observation
21:03:00	o1
21:06:00	avg(o2, o3)
21:09:00	avg(o2, o3)
21:12:00	o4

Figure 4-5: Example of data transformation

Missing data

Missing values are dealt with using forward filling or imputation. That is, for intervals without observations, the feature takes the value in the most recent available observation. In case no observation of drugs is available during the whole ICU stay, the patient is assumed not being administrated drugs and their input values are zeros. In the case of missing temperature, the missing data are filled with the value of 37. For the FiO2 and PEEP parameters, the patient is assumed not being under ventilation by default, and their variables take respectively the values of 21% and 0. For other features, the patient record is simply discarded from the data-set.

Normalization

Finally, the variables are normalized so that the values of each feature in the data have zero-mean and unit-variance. This method of feature scaling is the most common in practice for artificial neural networks [15]. The resulting values, called z-scores, are obtained from the raw values x by calculating the distribution mean $\bar{x}$ and standard deviation σ for each feature, then applying the formula:

$$z = \frac{x - \bar{x}}{\sigma} \tag{4.1}$$

4.6 Learning

During the training, the cost function to minimize is a binary cross-entropy error function. Given a batch of N samples, each with a label y_i and a predicted value $\hat{y}_i$, a binary cross-entropy is mathematically expressed as follows:

$$H = -\frac{1}{N} * \sum_{i}^{N} (y_i * log(\hat{y}_i) + (1 - y_i) * log(1 - \hat{y}_i)) \tag{4.2}$$

The loss function is optimized using the Adam optimizer on mini-batches of 128 samples, it is configured as in the original paper [37]. The criterion for stopping the learning process is the decrease in the area under the ROC curve (AUC) from testing

the model on validation set. All models were implemented in Keras with TensorFlow as the backend engine.

4.7 Evaluation

The evaluation is based on the confusion matrix, receiver operating characteristic (ROC) curve, histograms of probabilities by organ system and sensitivity across time before the onset of deterioration

4.7.1 Confusion matrix and receiver operating characteristic (ROC) curve

We use a confusion matrix to report the number of false positives (fp), false negatives (fn), true positives (tp), and true negatives (tn). We refer to the correctly identified samples by *true*, and by *false* we refer to the incorrectly identified samples. Furthermore, a prediction is said to be positive when it claims deterioration within the prediction window (predicted class = 1). Otherwise, it is said to be negative (predicted class = 0). In this study, the comparison between the different models is based on multiple statistical measures, particularly the sensitivity ($TPR = \frac{tp}{tp+fn}$), the negative predictive value ($NPV = \frac{tn}{tn+fn}$), and the area under the ROC curve. We choose these performance metrics for evaluation in order particularly to evaluate missing deterioration and false alarms, and for comparison between the different models.

From the predicted probabilities of deterioration , the receiver operating characteristic (ROC) curve is drawn. It plots sensitivity (TPR) against false positive rate ($FPR = \frac{fp}{fp+tn}$) while varying the threshold of predicted probability for the classification of samples. It is worth noting that the confusion matrix, the sensitivity and the negative predictive value can change with respect to that threshold for output probabilities. By default, they are reported for the 50% threshold probability above which a sample is classified as indicating deterioraton (predicted class = 1). In contrast, the area

under the ROC curve offers a broader basis of comparison as it does not depend on that threshold. Moreover, it can be noticed from the literature review that the area under the ROC curve is worth reporting in the results.

4.7.2 Histograms of predicted probabilities by type of deterioration

In order to investigate how well the model performs with regard to each of the relevant organ systems (i.e cardiac or respiratory organ systems), data samples are divided into 4 different categories:

- **No deterioration**: samples in this category are of actual class 0 which means that no deterioration occurs during the prediction window.
- **Cardiac deterioration only**: samples in this category are of actual class 1 and represent cases in which a cardiac deterioration occurs within the prediction window.
- **Respiratory deterioration only**: samples in this category are of actual class 1 and represent cases in which a respiratory deterioration occurs within the prediction window.
- **Both cardiac and respiratory deterioration**: samples in this category are of actual class 1 and represent cases in which both cardiac and respiratory deterioration occur within the prediction window.

Then a histogram of the predicted probabilities is made for each category to estimate its predicted probability distribution. The aim of interpreting those histograms is to analyze the performance of the prediction regarding each type of deterioration.

4.7.3 Sensitivity across time before the onset of deterioration

It is important to make sure that a model does not only identify deterioration shortly after the prediction time, but also predict deterioration that starts later in the pre-

diction window. To that end, sensitivity is analyzed relative to the time before the onset of deterioration, using a plot of sensitivity against the time period between the prediction and the deterioration onset.

4.7.4 Comparison with baseline methods

The implemented neural networks are compared with the k-nearest neighbors (k-NN) algorithm and the random forest classifier (RF) in terms of performance. The k-nearest neighbors algorithm relies on the k nearest neighbors to classify a sample by their most common class. The standard Euclidean distance is selected as the metric for computing distance to neighbors. The random forest classifier is a collection of decision trees labeling with the most common class among the outputs of its individual trees. The Gini impurity is selected as the attribute selection criterion for constructing the decision trees. Each of these methods is tested with different settings so as to find the best AUC, mainly the number of nearest neighbors for the k-NN and the number of decision trees for the RF. The results chapter shows the performance of 5, 10 and 20-nearest neighbors algorithms and random forests with 10, 100, 300 and 500 estimators (trees).

4.7.5 Feature selection by backward feature selection

In order to assess the importance of each feature, a backward feature selection is performed using the architecture and parameters of the best predictive model. Starting with the set of all the features, the algorithm removes one by one a feature from the set based on the AUC score. The predictive model is trained after each feature removal from the FINNAKI training samples, then an AUC score is calculated from the FINNAKI test samples. This approach relies on the premise that the magnitude of change in performance after removing a feature informs of the extent to which it affects the classification. The result of this analysis is a ranking of the features according to their importance to the prediction.

Chapter 5

Results

As the aim is to construct predictive models that perform well in both FINNAKI data and the MIMIC III clinical database, those data-sets are compared in order to raise differences that may hinder that purpose. A selection of subsets of the data-sets is made so as to reduce the differences in the structure of the data-sets. Then the models are trained on a subset of the FINNAKI data-set and are tested on subsets of FINNAKI data and MIMIC III clinical database. Finally, this thesis presents an analysis of feature importance through a backward feature selection.

5.1 Comparison between FINNAKI and MIMIC III critical care database and selection of subsets

FINNAKI and MIMIC III critical care databases differ in many aspects. Relating international comparisons of intensive care [38], Meghan et al. outline variations in the ICU population and resources across different countries. The U.S have relatively a high proportion of ICU beds per capita, allowing for more patients being transferred directly from the emergency room instead of going to the general ward.

5.1.1 Sampling rate

One major difference between MIMIC III clinical database and FINNAKI is in the sampling rate. This is demonstrated by comparing statistics of the time spans (in minutes) between consecutive measurements of two prominent features: heart rate (HR) and systemic arterial pressure - systolic (SAPS).

	Mean (min)	Median (min)	25^{th} percentile (min)	75^{th} percentile (min)
FINNAKI	2.21	2	1	2
MIMIC III	48.64	60	30	60

Table 5.1: HR sampling rate statistics

	Mean (min)	Median (min)	25^{th} percentile (min)	75^{th} percentile (min)
FINNAKI	2.21	2.07	1.55	2.82
MIMIC III	50.02	60	30	70

Table 5.2: SAPS sampling rate statistics

As shown in the tables 5.1 and 5.2, a difference in the frequency of HR and SAPS between Finnaki and MIMIC III data sets is noticeable. Specifically, HR and SAPS are recorded in FINNAKI at a higher sampling rate.

5.1.2 Cardiac and respiratory SOFA

Another difference arises from the distribution of SOFA in each data-set. From random subsets of patients selected from both data-sets, the tables 5.3 and 5.4 show a difference in the distribution of the cardiac and respiratory SOFA sub-score. Particularly, the predominant cardiac sub-score is 3 in FINNAKI whereas it is 0 in MIMIC III.

Cardiac SOFA sub-score	**FINNAKI** (%)	**MIMIC III** (%)
0	27.17	**67.32**
1	4.05	17.68
2	1.09	2.15
3	**52.39**	7.91
4	15.30	4.93

Table 5.3: Cardiac SOFA distribution

Respiratory SOFA sub-score	**FINNAKI** (%)	**MIMIC III** (%)
0	**43.71**	**60.78**
1	33.31	15.36
2	15.57	10.98
3	6.33	8.80
4	1.06	4.06

Table 5.4: Respiratory SOFA distribution

The prediction is only made when the patient is in a stable condition. The table 5.5 shows the percentage of samples in which the patient is in an stable state for each of the data-sets.

	FINNAKI (%)	**MIMIC III** (%)
Stable state	19.78	40.82
Unstable state	80.22	59.18

Table 5.5: Distribution of states

5.1.3 Selection of comparable subsets

The difference between the MIMIC III clinical database and FINNAKI in terms of the sampling rate imposes selecting comparable subsets in order to apply models on both data-sets. For instance, one could select a subset of patients from the MIMIC III clinical database on the basis of the recording frequency of one or multiple of the parameters. In this study, a subset of the MIMIC III clinical database is defined based on the median of the time spans between successive measurements of heart rate. The subset is a clinical data-set of 531 patients of which at least 50% of heart rate data are originally recorded at least once per 15 minutes. In other words, the subset is restricted to patients with the median of the variable X = "time spans between 2 successive measurements of heart rate" less or equal 15 minutes. The purpose of this subsetting is to reduce the effect of the sampling rate and hence obtain a data-set from the MIMIC III clinical database that is valid for testing models trained on FINNAKI data.

FINNAKI

For the sake of comparison, we use the same training, validation and test sets for all the models for which we present the results. The data-set includes in total 983 patients of which 500 are for training, 100 for validation, and 383 for testing. Each sample of the data-set consists of input data and a class. In the training set, 135195 out ot the total 285125 samples are of class 1, and in the test set, 89875 out of the total 284074 samples are of class 1 (i.e. deterioration occurs within the prediction window).

	label(class)	
	No deterioration (0)	**deterioration (1)**
Number of samples	149930	135195
Percentage of samples	52.58%	47.42%

Table 5.6: Class distribution in the FINNAKI training set

	label(class)	
	No deterioration (0)	**deterioration (1)**
Number of samples	194199	89875
Percentage of samples	68.36%	31.64%

Table 5.7: Class distribution in the FINNAKI test set

MIMIC III

The MIMIC test set includes in total 531 patients. Specifically in the test set, 55422 (40.91%) out of the total 135462 samples are of class 1 (i.e. deterioration occurs within the prediction window).

	label(class)	
	No deterioration (0)	**deterioration (1)**
Number of samples	80040	55422
Percentage of samples	59.09%	40.91%

Table 5.8: Class distribution in the MIMIC III test set

5.2 First CNN model

5.2.1 Confusion matrices

The table 5.9 of confusion shows that applying the 1st predictive model to the FINNAKI test set results in a sensitivity of 76.20% and a specificity of 63.54%. Applied to the MIMIC III test set, this model performs better by 12% in sensitivity but worse by 32% in specificity, dragging down the accuracy by 12% as shown in the table of confusion 5.10.

		Predicted class		
		0	1	
Actual	0	**123397**	70802	Specificity = 63.54%
class	1	21389	**68486**	Sensitivity = 76.20%
		NPV = 85.23%	PPV = 49.17%	Accuracy = 67.55%

Table 5.9: Confusion matrix of the 1st predictive model on FINNAKI

		Predicted class		
		0	1	
Actual	0	**25570**	54470	Specificity = 31.19%
class	1	6186	**49236**	Sensitivity = 88.84%
		NPV = 80.52%	PPV = 47.48%	Accuracy = 55.22%

Table 5.10: Confusion matrix of the 1st predictive model on MIMIC

The performance metrics derived from the confusion matrices are summarized in the table 5.11

	FINNAKI	**MIMIC III**
Sensitivity (%)	76.20	88.84
Specificity (%)	63.54	31.19
PPV (%)	49.17	47.48
NPV (%)	85.23	80.52
Accuracy (%)	67.55	55.22

Table 5.11: Performance metrics of the 1st predictive model

5.2.2 ROC curves

The ROC curves 5-1 and 5-2 show that the 1st predictive model performs better in predicting samples from FINNAKI. The area under the ROC curve is 0.7665 from the

FINNAKI test set but decreases to 0.6722 when the model is applied to the MIMIC III test set.

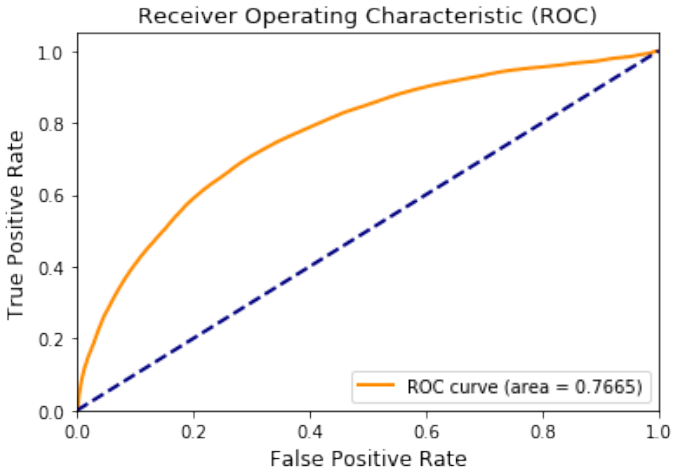


Figure 5-1: ROC curve of the 1st predictive model on FINNAKI

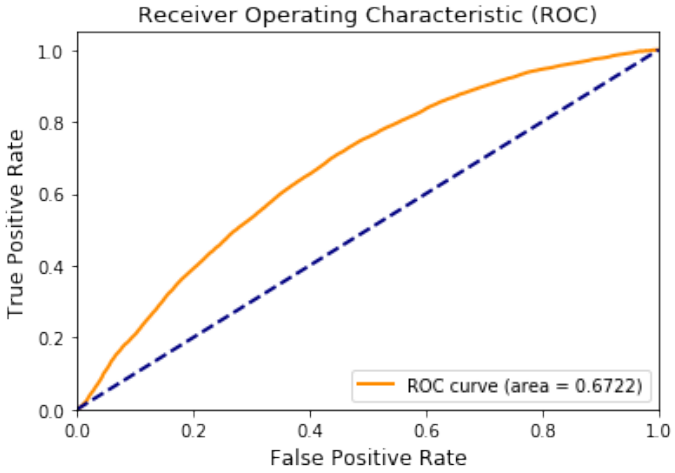


Figure 5-2: ROC curve of the 1st predictive model on MIMIC III

5.2.3 Histograms of predicted probabilities by organ system

The histograms of predicted probabilities generated from the FINNAKI test set, shown in the figure 5-3, depict a stretched histogram related to respiratory deterioration with 38.30% of the total predicted probabilities beyond 0.5. In contrast, the histograms regarding cardiac deterioration and both deterioration types show less

dispersion with respectively 21.08% and 12.42% of the total predicted probabilities beyond 0.5. On the other hand, the histograms of predicted probabilities generated from the MIMIC III test set (figure 5-4 depict a large false positive rate (68.05%) but greater positive predictive values for both types of deterioration.

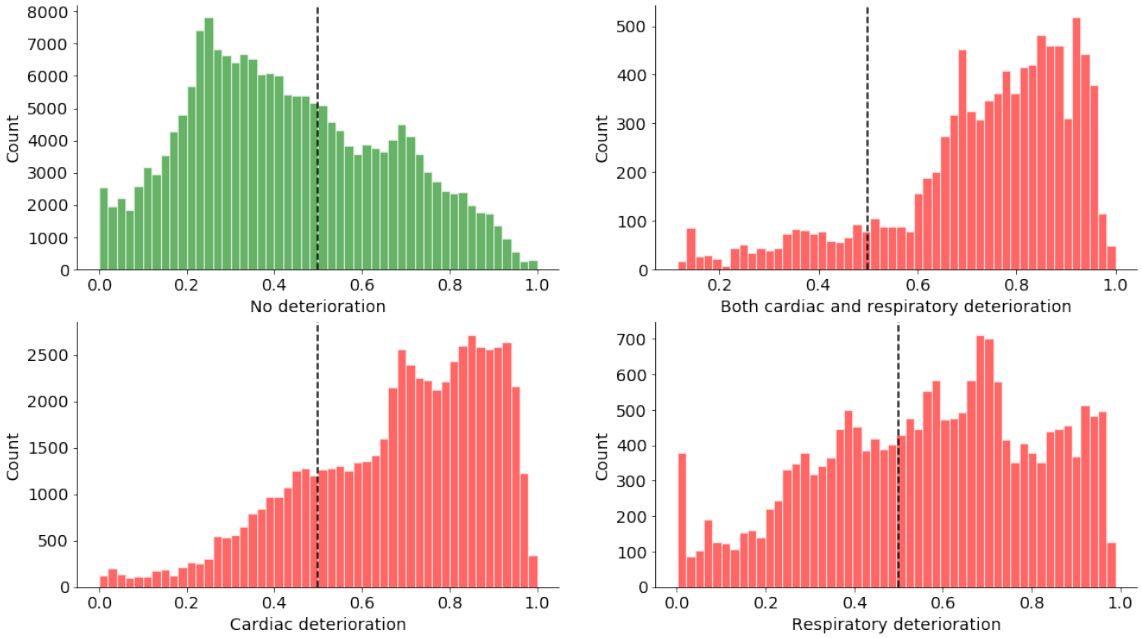


Figure 5-3: Histograms of predicted probabilities by type of deterioration (FINNAKI)

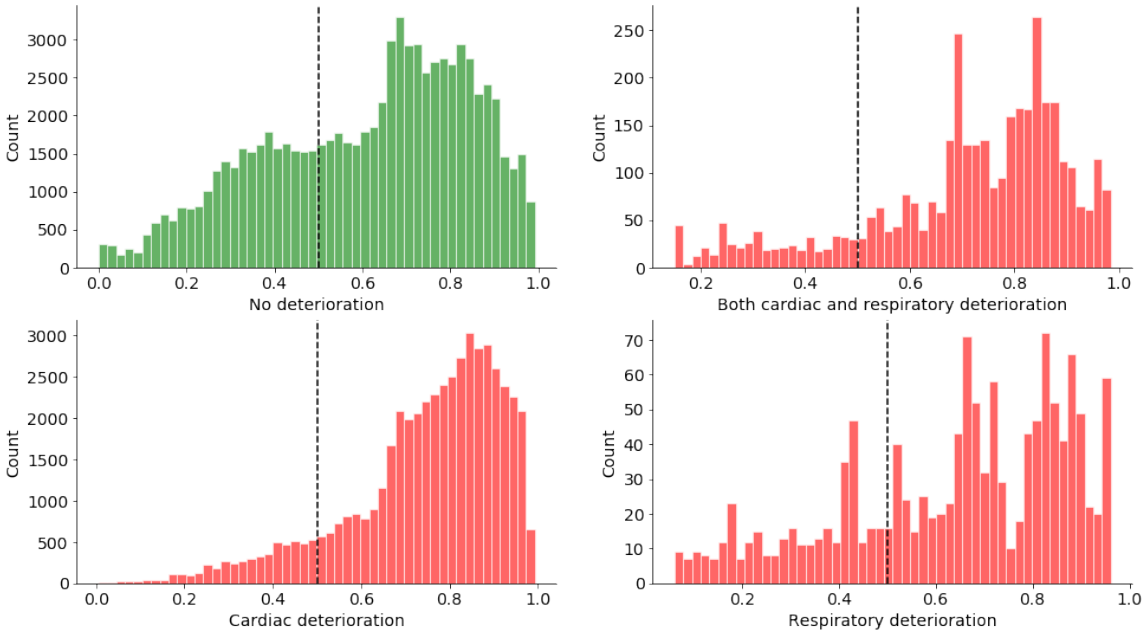


Figure 5-4: Histograms of predicted probabilities by type of deterioration (MIMIC III)

5.2.4 Sensitivity across time before deterioration

The figures 5-5 and 5-6 show the increase in the sensitivity as the time of prediction approaches deterioration. Although the graphs show a higher sensitivity from the prediction on the MIMIC III test set, it is worth noting that the corresponding specificity (31.19%) is lower than that of the prediction on the FINNAKI test set (63.54%).

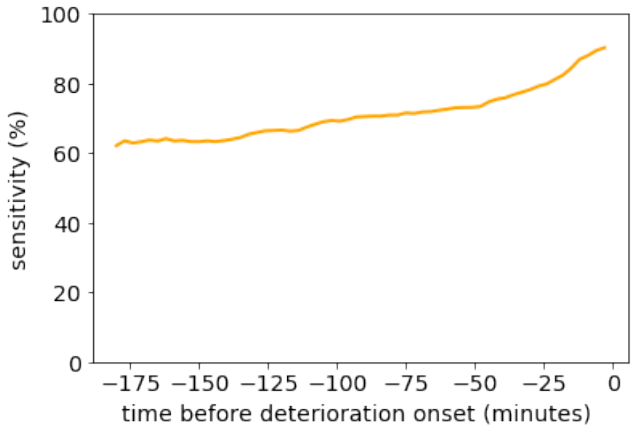


Figure 5-5: Sensitivity across time before deterioration onset (FINNAKI)

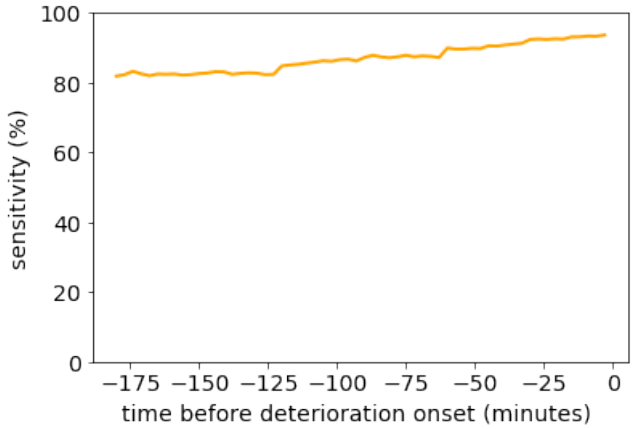


Figure 5-6: Sensitivity across time before deterioration onset (MIMIC III)

5.3 Second CNN predictive model

As the additional input features for the 2nd model were only retrieved from FINNAKI data only, the following results concern the FINNAKI test set. The confusion matrix shown in the table 5.12 is obtained from the 2nd model tested on FINNAKI data with a sensitivity of 85.23% and a specificity of 63.54%. The area under the corresponding ROC curve traced in figure 5-7 is 0.7741.

		Predicted class		
		0	1	
Actual	0	**94799**	99400	Specificity = 48.82%
class	1	13274	**76601**	Sensitivity = 85.23%
		NPV = 87.72%	PPV = 43.52%	Accuracy = 60.34%

Table 5.12: Confusion matrix of the 2nd predictive model on FINNAKI

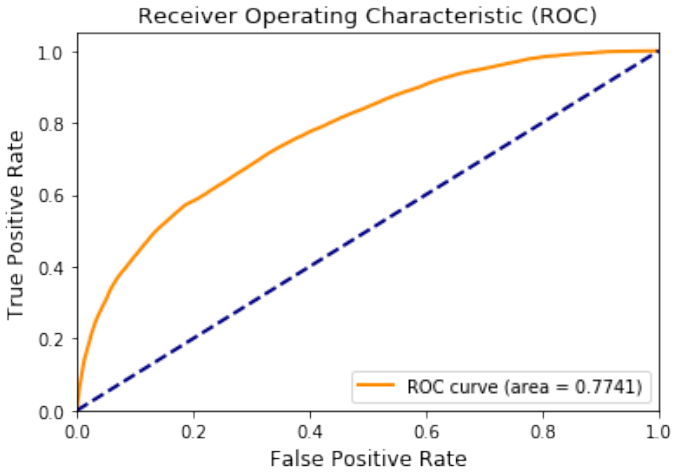


Figure 5-7: ROC curve of the 2nd predictive model

The figure 5-8 suggests that the model predicts all types of deterioration with high sensitivity (84.91% of cardiac, 81.25% of respiratory and 95.36% of both deterioration samples are detected). However, It fails to predict 51.18% of non deteriorating samples as shown in the upper-left histogram.

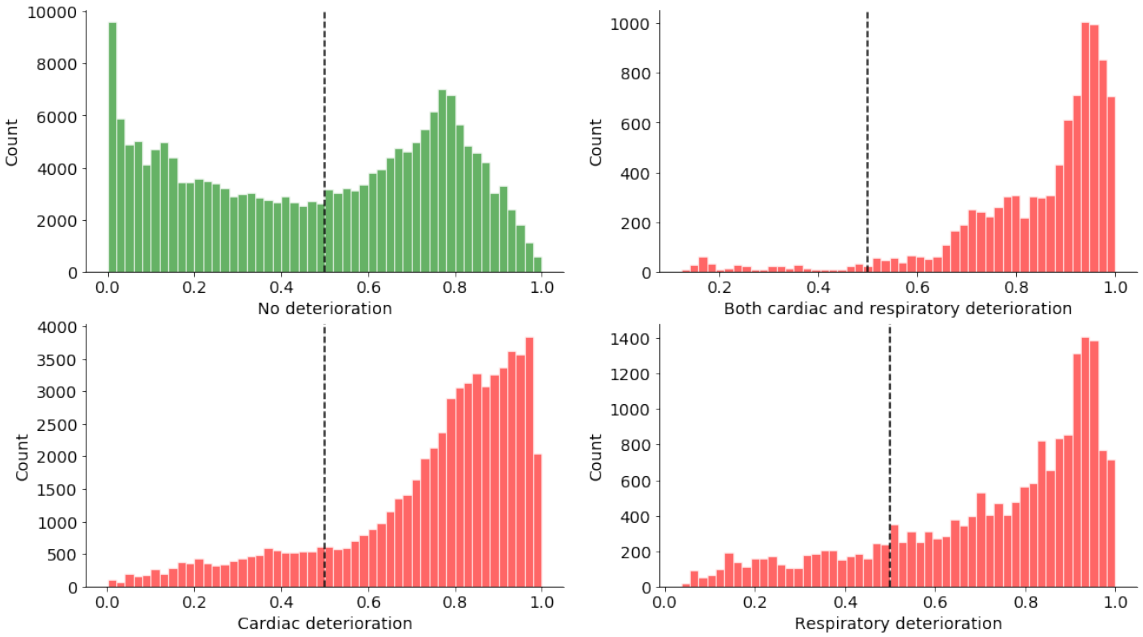


Figure 5-8: Histogram of predicted probabilities by type of deterioration

The graph of sensitivity across time as displayed in the figure 5-9 shows that the model is also capable of detecting the onset of deterioration even later in the prediction window.

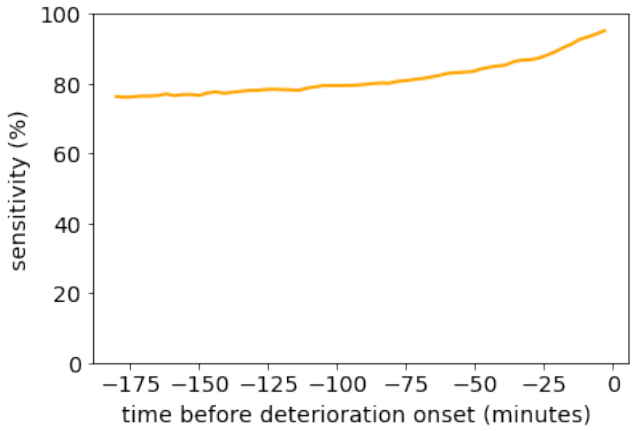


Figure 5-9: Sensitivity across time before deterioration onset

5.4 Third predictive model

5.4.1 Confusion matrices

The table of confusion 5.13 shows the outcome of testing the 3rd predictive model on FINNAKI data. The model achieves a sensitivity of 74.36% and a specificity of 68.22%. Applied to the MIMIC III test set, this predictive model scores higher by 14% in sensitivity but lower by 35% in specificity, dragging down the accuracy by 24% as shown in the table of confusion 5.14.

		Predicted class		
		0	1	
Actual	0	**132476**	61723	Specificity = 68.22%
class	1	23041	**66834**	Sensitivity = 74.36%
		NPV = 85.18%	PPV = 51.99%	Accuracy = 70.16%

Table 5.13: Confusion matrix of the 3rd predictive model on FINNAKI

		Predicted class		
		0	1	
Actual	0	**26453**	53587	Specificity = 33.05%
class	1	5970	**49452**	Sensitivity = 89.23%
		NPV = 81.59%	PPV = 47.99%	Accuracy = 56.03%

Table 5.14: Confusion matrix of the 3rd predictive model on MIMIC

The performance metrics derived from the confusion matrices are summarized in the table 5.15

	FINNAKI	MIMIC III
Sensitivity (%)	74.36	89.23
Specificity (%)	68.22	33.05
PPV (%)	51.99	47.99
NPV (%)	85.18	81.59
Accuracy (%)	70.16	56.03

Table 5.15: Performance metrics of the 3rd predictive model

5.4.2 ROC curves

The ROC curves 5-10 and 5-11 show that the 1st predictive model performs better at predicting samples from FINNAKI. The area under the ROC curve is 0.7812 from the FINNAKI test set but decreases to 0.6816 when the model is applied to the MIMIC III test set.

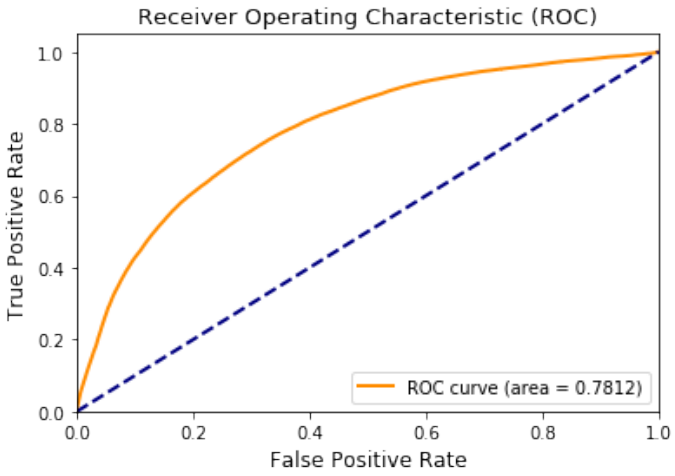


Figure 5-10: ROC curve of the 3rd predictive model on FINNAKI

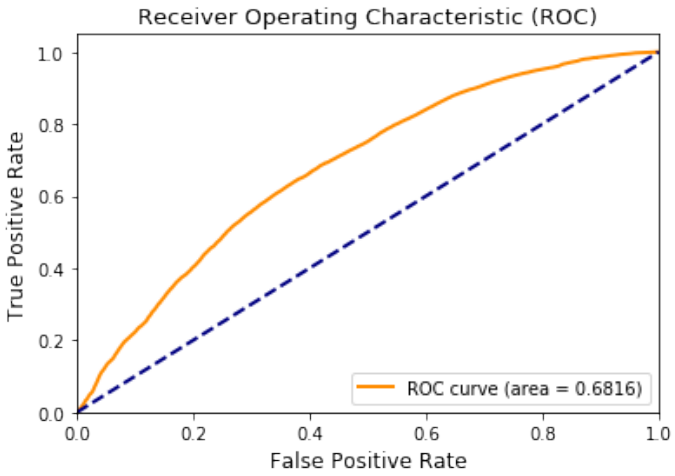


Figure 5-11: ROC curve of the 3rd predictive model on MIMIC III

5.4.3 Histograms of predicted probabilities by type of deterioration

The figure 5-12 depicts skewed distribution of the predicted probabilities peaking close to 0.10 for non deteriorating samples and close to 0.90 for samples with cardiac deterioration alone or combined with respiratory deterioration. 76.74% of samples with cardiac deterioration as well as 84.95% of samples with both deterioration types are correctly identified. This shows that the model performs well at detecting those types of deterioration. The bimodal still not symmetric distribution in the bottom-right histogram shows that the model is also capable of predicting respiratory deterioration alone but it omits 38.66% of samples representing this deterioration type.

On the other hand, the histograms of predicted probabilities generated from the MIMIC III test set (figure 5-13) depict a large false positive rate (61.34%) and a multimodal distribution of the predicted probabilities of respiratory deterioration alone. Similarly to the test on FINNAKI data, the model demonstrates a good sensitivity in predicting cardiac deterioration alone (90.77%) or combined with respiratory deterioration (79.13%).

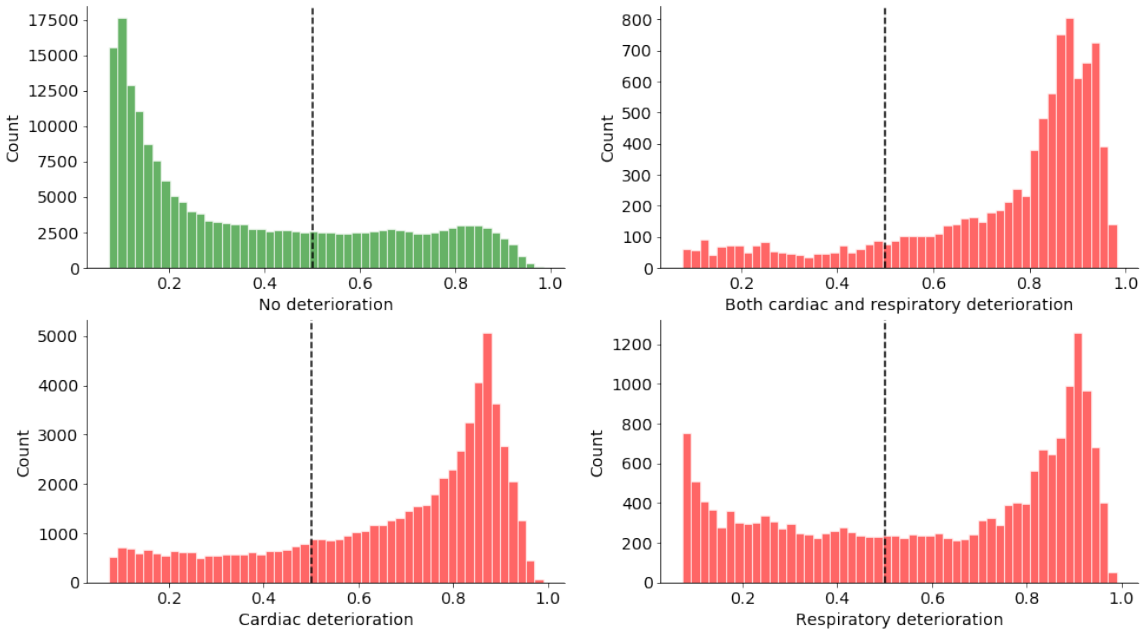


Figure 5-12: Histograms of predicted probabilities by type of deterioration (FINNAKI)

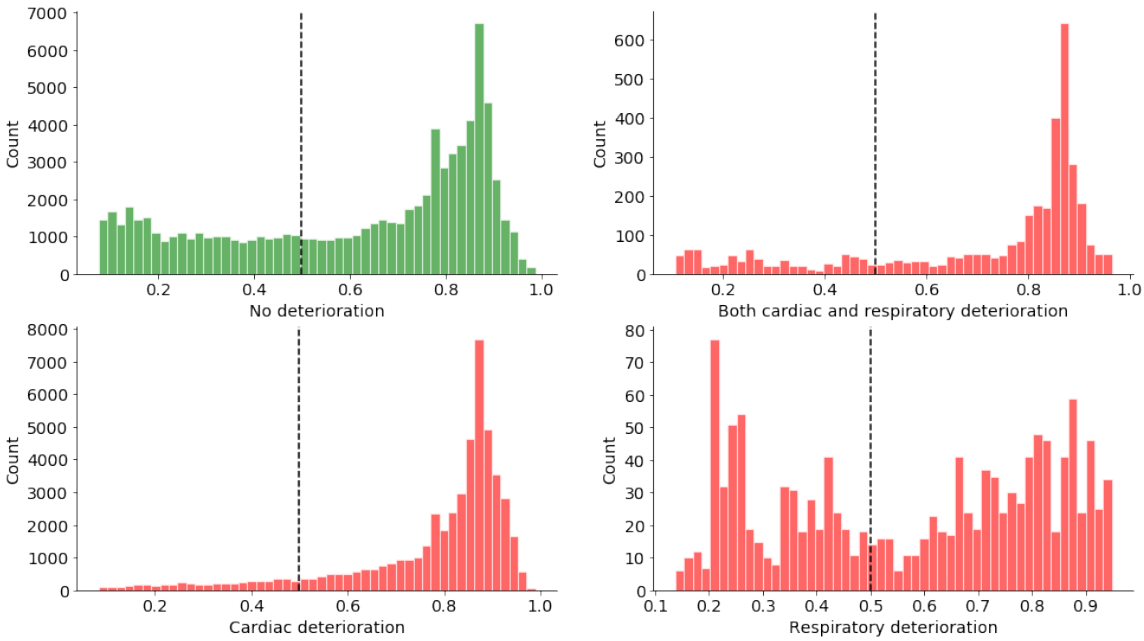


Figure 5-13: Histograms of predicted probabilities by type of deterioration (MIMIC III)

5.4.4 Sensitivity across time before deterioration

The figures 5-14 and 5-15 show the increase in sensitivity as the time of prediction approaches deterioration. Although the graphs show a higher sensitivity from the prediction on the MIMIC III test set, it is worth noting that the corresponding specificity (3.05%) is lower than that of the prediction on the FINNAKI test set (68.22%).

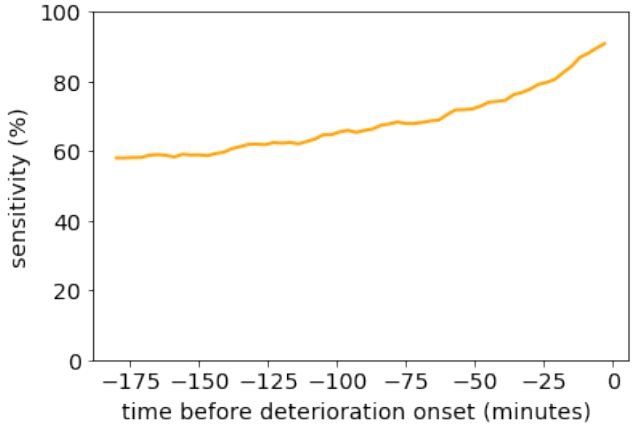


Figure 5-14: Sensitivity across time before deterioration onset (FINNAKI)

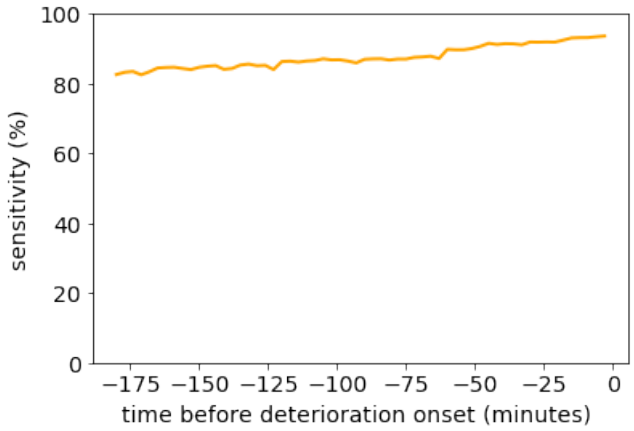


Figure 5-15: Sensitivity across time before deterioration onset (MIMIC III)

5.5 Comparison of performance between the developed models and other baseline methods (random forest and nearest-neighbors algorithm)

The table 5.16 shows the performance metrics of the presented neural networks calculated from the test sets. In boldface are the highest performance values on each data set. In comparison, the table 5.17 reports the performance of the 9, 29, 49 and 89 nearest-neighbors algorithms, and the table 5.18 reports the performance of random forests (RF) with 10, 100, 300 and 500 trees. Several nearest-neighbors algorithms were tested, ranging from 9 to 99. Their AUCs are plotted in the figure 5-16, depicting an increase of the AUC as more neighbors are considered. Random forests with other numbers of trees were tested as well, ranging from 10 to 500. AUCs are plotted in the figure 5-17 for 10, 20, 30, 50, 100, 300 and 500 trees, which shows the AUC increasing until an upper bound as the number of trees grows.

The performance metrics are displayed side by side in the graphs 5-18 and 5-19 resulting from the application of the models on Finnaki and MIMIC III respectively.

	First model		**Second model**		**Third model**	
	FINNAKI	**MIMIC III**	**FINNAKI**	**MIMIC III**	**FINNAKI**	**MIMIC III**
Sensitivity (%)	76.20	88.84	**85.23**	-	74.36	**89.23**
Specificity (%)	63.54	31.19	48.82	-	**68.22**	**33.05**
PPV (%)	49.17	47.48	43.52	-	**51.99**	**47.99**
NPV (%)	85.23	80.52	**87.72**	-	85.18	**81.59**
Accuracy (%)	67.55	55.22	60.34	-	**70.16**	**56.03**
AUC	0.7665	0.6722	0.7741	-	**0.7812**	**0.6816**

Table 5.16: Performance metrics of the implemented neural networks

	9-nearest neighbors		29-nearest neighbors		49-nearest neighbors		89-nearest neighbors	
	FINNAKI	MIMIC III	FINNAKI	MIMIC III	FINNAKI	MIMIC III	FINNAKI	MIMIC III
Sensitivity (%)	49.03	54.69	50.96	59.44	52.39	62.32	**53.27**	**65.22**
Specificity (%)	69.32	**54.80**	68.81	51.19	69.46	48.72	**70.61**	47.00
PPV (%)	42.50	45.74	43.04	45.90	44.23	45.85	**45.60**	**46.16**
NPV (%)	74.63	63.45	75.21	64.42	75.93	64.98	**76.57**	**65.98**
Accuracy (%)	62.91	**54.75**	63.17	54.57	64.06	54.30	**65.13**	54.48
AUC	0.5917	0.5474	0.5989	0.5531	0.6092	0.5552	**0.6193**	**0.5611**

Table 5.17: Performance metrics of 5, 10 and 20 nearest neighbors algorithms

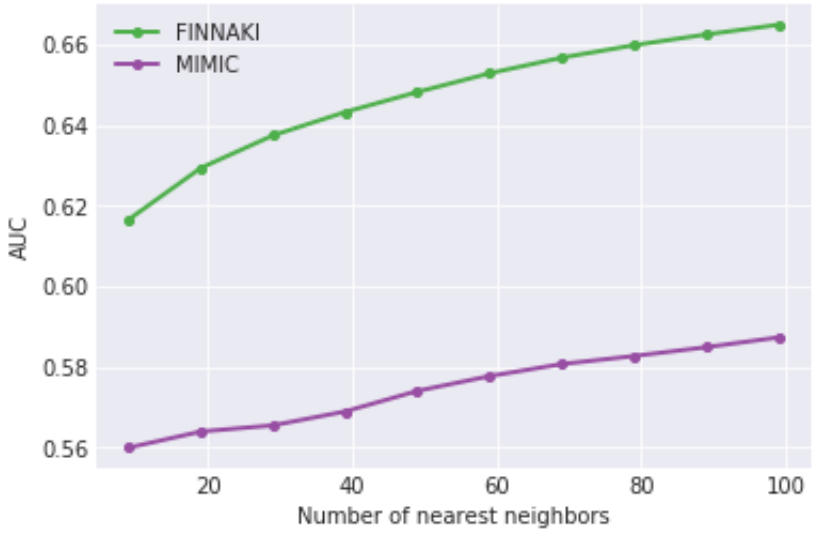


Figure 5-16: AUCs of k-NN with k ranging from 9 to 99

	RF of 10 estimators		RF of 100 estimators		RF of 300 estimators		RF of 500 estimators	
	FINNAKI	MIMIC III	FINNAKI	MIMIC III	FINNAKI	MIMIC III	FINNAKI	MIMIC III
Sensitivity (%)	51.72	52.92	59.93	63.85	60.77	66.85	**61.23**	**68.03**
Specificity (%)	**77.72**	**62.13**	74.49	60.90	73.78	60.17	73.55	58.77
PPV (%)	51.80	49.18	**52.09**	53.07	51.75	**53.75**	51.72	53.33
NPV (%)	77.67	65.58	80.07	70.87	80.25	72.39	**80.39**	**72.64**
Accuracy (%)	69.49	58.36	**69.88**	62.11	69.66	**62.90**	69.65	62.56
AUC	0.7049	0.6056	0.7281	0.6622	**0.7317**	**0.6696**	0.7310	0.6676

Table 5.18: Performance metrics of random forests with 10, 100, 300, and 500 estimators

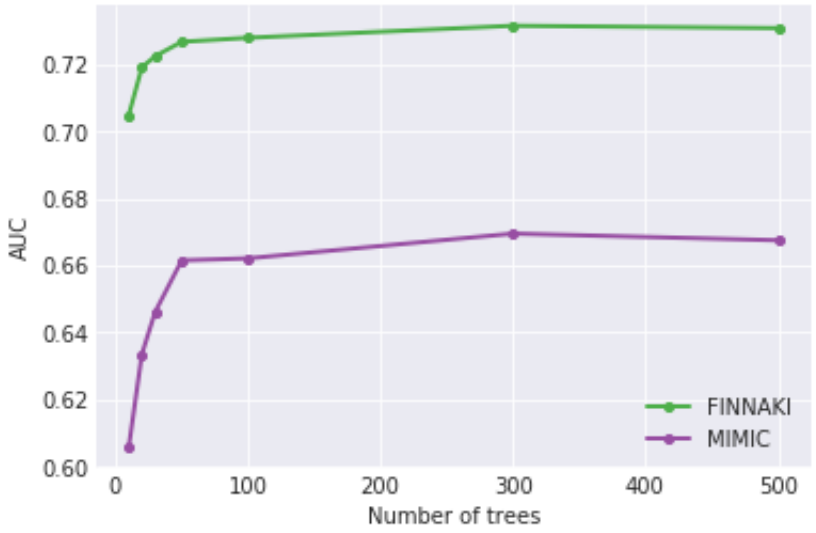


Figure 5-17: AUCs of random forests with number of trees ranging from 10 to 500

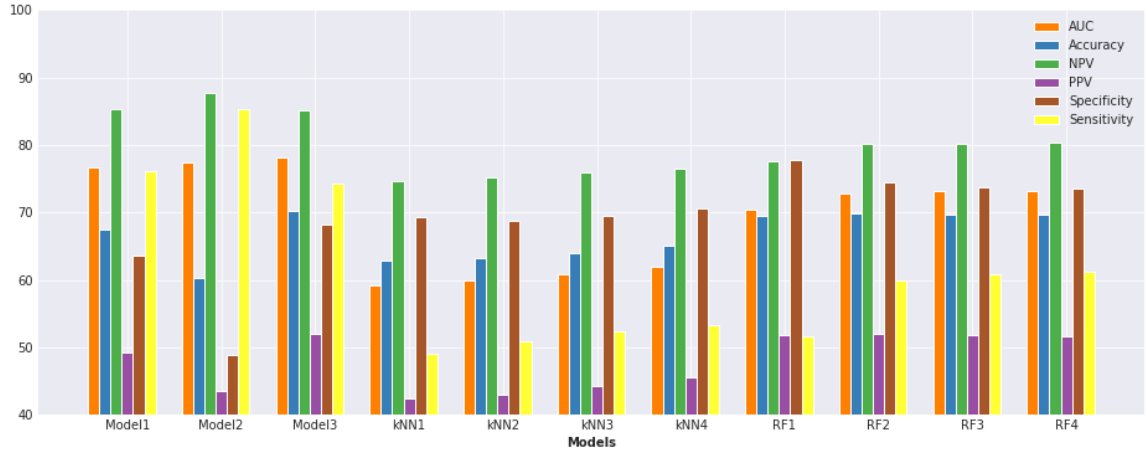


Figure 5-18: Performance metrics from the FINNAKI test set

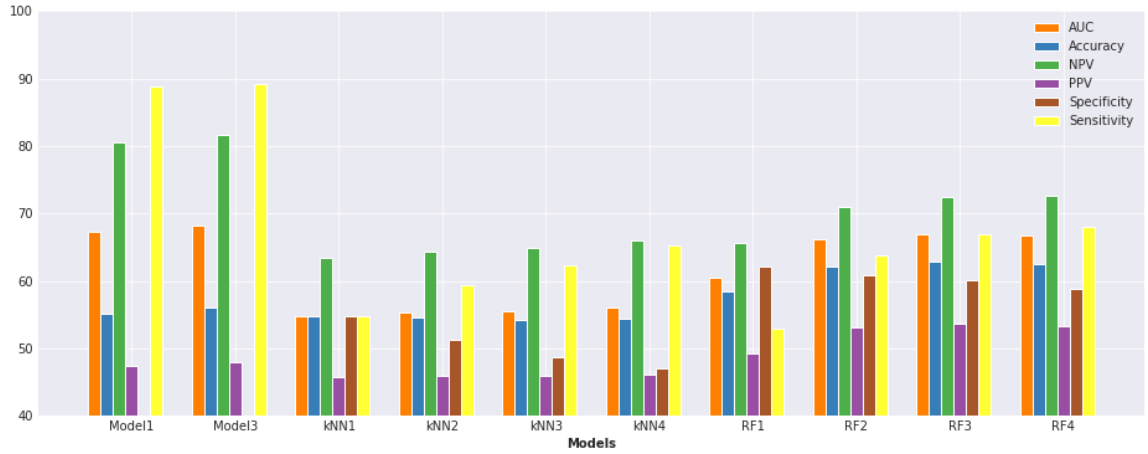


Figure 5-19: Performance metrics from the MIMIC III test set

Based on the AUC, the results suggest that the neural networks outperform the baseline methods. The k-nearest neighbors algorithms perform clearly the worst. Their AUC increases with the number of nearest neighbors but their implementation is comparatively slow in time. The random forest classifiers score up to 0.7317 in FINNAKI and 0.6696 in MIMIC III which are less than the AUC values of the neural networks.Finally, the third model surpasses the other neural networks. Yet the gap between their AUC values is smaller than the difference between the neural networks and the baseline methods.

5.6 Feature importance from the third model

As explained in the methods chapter, a backward feature selection is conducted using the same architecture and parameters as the third model. It is based on the AUC score that is calculated from the FINNAKI test samples. The table 5.19 shows the gradual steps of the backward selection process starting from the top row down to the bottom row. Beside each step number is the name of the next discarded feature and the resulting AUC score from the classification of the test samples. The age is first omitted from the set of features, then other features are gradually omitted in the following steps until only one feature is left. It is worth mentioning that the AUC of the classification of the training set is 0.7812 when all the features are included. Along the steps, the AUC increases up to 0.8046 then decreases after the removal of RR. The classification based on the whole set of features is outperformed by a classification based only on a subset of features, which indicates that the model may have been overfitted with irrelevant features. For instance, the omission of age, gender, SAPD, temperature and SAPS yields a higher AUC score (0.8046) than considering the whole set of 14 features (0.7812).

Step	**Discarded feature**	**Resulting AUC on Finnaki test set**
1	Age	0.7849
2	Gender	0.7885
3	SAPD	0.7920
4	Temperature	0.7970
5	SAPS	0.8046
6	RR	0.8025
7	HR	0.7979
8	Cardiac LODS	0.7810
9	Time of day	0.7803
10	SpO2	0.7754
11	PaO2	0.7709
12	PaCO2	0.7600
13	FiO2	0.7077
14	MAP	0.500

Table 5.19: Backward selection steps

Chapter 6

Discussion

The results of the predictive models reveal that short term cardiac and respiratory deterioration can be predicted from a short history of medical records. A high prediction performance can be achieved using neural networks with temporal convolutional layers and LSTMs. The 3rd predictive model shows the best performance in terms of the AUC. Unlike the other tested models, its architecture leverages the related patterns of future cardiac and respiratory organ dysfunctions through a shared representation, and enables to determine the organ system that is subject to deterioration through, which offers more details from the clinician's point of view. This suggests that such architectures involving multitask learning and LSTMs are worth implementing in predicting deterioration in the ICU.

As shown in the results of the 2nd model, the additional parameters taken in the previous 2 hours seem to improve only by little than expected the performance of prediction within the next 3 hours. This leads to the idea that many of them are not important for the prediction task. Nevertheless, it may also imply that these features affect more significantly predictions in longer terms, or that it requires even more complex machine learning techniques to demonstrate a bigger impact on the performance metrics and avoid overfitting. Yet, the development of more complex models is limited by the size of the data due to the increase in the number of weights that have to be adjusted during the learning process. Adding a convolutional layer to the 2nd model was also tested but did not show success in improving the performance

metrics.

As explained in the literature review, differences in the formulation of deterioration and time scope do not help in making a valid comparison between related studies. The presented models are capable of predicting whether the patient is incurring a drop in the mean arterial pressure (MAP) to a value less than 70 mmHg or needing mechanical ventilation within the next 3 hours, whereas the most related research majorly predicts vasopressor intervention and ventilation intervention [26] [25]. It is also worth noting that these studies rely on hourly averaged data while this study also considers cases in which a decline in the condition is shorter in time.

The histograms of output probabilities by organ system suggest that respiratory deterioration is harder to predict than cardiac deterioration when the prediction task combines both types of deterioration. This finding matches to some extent the study of Harini et al. [26] in which vasopressor ventilation is predicted with a higher AUC (0.77) compared to invasive ventilation (0.75).

Overall, the models perform better in the prediction of the FINNAKI test set with respect to the AUC as compared to the MIMIC III test set. It should be recalled that the FINNAKI data and MIMIC III clinical data differ in terms of the sampling rate and distribution of the SOFA sub-scores. Points of dissimilarity between the datasets and variations in the ICU population and resources may explain the difference in the predictive performance. It should also be remembered that the models have been trained exclusively on FINNAKI data. It remains possible for the model to learn from multiple sources so as to fit regardless of the geographical location of the ICU.

Regarding the available data, it also allows to predict recovery from the samples in which the patient is in an unstable state at the time of prediction. However from these samples, those in which the patient is recovering in the future represent a small portion (e.g. less than 0.1% in FINNAKI data). One reason suggested by professor Ville Pettilä is that the patient is often discharged in less than 3 hours after the end of vasopressor or ventilation administration.

The feature importance analysis shows that some parameters (e.g. MAP and FiO2) are crucial whereas other parameters (e.g. age) hardly influence the classification. The

classification also reveals that the omission of a subset of features (e.g. age, cardiac LODS, temperature) can yield a similar or even better performance than considering the whole set of 14 features. This analysis provides an insight about the significance of features in this prediction task. However, the use of neural networks poses a challenge in investigating the patterns, from the input along time, that are responsible for the output predictions. Compared to other machine learning approaches, complex neural networks offer higher accuracy to the detriment of easy interpretability [39]. In a clinical perspective, it is important to know the pattern triggering the alarm so as to provide the right treatment. Yet the aim remains to provide an assistant tool to help the caregivers detect intensive care patients that are likely deteriorating in the near future. These models then remain useful in the clinical context. One could also use interpretable models (e.g. decision trees) alongside neural networks, or other solutions suggested by literature for interpreting neural networks [39].

The predictive models offers the possibility to choose an acceptable rate of false alarms by varying the threshold for classifying output probabilities. However, reducing the false positive rate comes at the expense of sensitivity. The decision on the threshold value depends on the extent of tolerance towards false alarms and unidentified cases of deterioration, and a clear improvement of the performance is made when an increase of sensitivity is achieved while preventing the decrease of specificity.

Chapter 7

Conclusion

In this thesis, deep neural networks are implemented to predict if a patient is prone to cardiac or respiratory deterioration within the next 3 hours given health records from the past 2 hours. Trained and tested on FINNAKI data, deep neural networks could score up to 0.7812 in the area under the ROC curve. On the other hand, testing those trained models on a sub-set of the MIMIC III clinical database results in a drop of the area under the ROC curve to 0.6816, which can be explained by the dissimilarity between the data-sets. Temporal convolutional neural networks and long short term memory networks demonstrated their ability of leveraging the temporal features networks, and multitask learning was useful for combining in one model the tasks related to cardiac and respiratory deterioration with a shared representation. The predictors that are developed in this work can prove valuable as part of clinical decision support in that they can help the caregivers narrow their focus on the intensive care patients of potentially critical state so that earlier diagnosis or preventive treatment can be made. Interpreting the patterns influencing the classification and analyzing an implementation of the predictors in the intensive care unit could be subjects of further research.

Bibliography

[1] Christopher Olah. Understanding lstm networks, 2015, accessed: April 3, 2018. `http://colah.github.io/posts/2015-08-Understanding-LSTMs`.

[2] Pirkko Nykanen and Niilo Saranummi. Clinical decision systems. In J.D. Bronzino, editor, *Biomedical Engineering Handbook*. Taylor & Francis, 1999.

[3] Amy Grace Rapsang and Devajit C. Shyam. Scoring systems in the intensive care unit: A compendium. *Indian Journal of Critical Care Medicine : Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 18(4):220–228, -4 2014.

[4] Robert C. Hyzy. Icu scoring and clinical decision making. *CHEST*, 107(6):1482–1483, -06-01 1995.

[5] Ville Pettilä, Markus Pettilä, Seppo Sarna, Petri Voutilainen, and Olli Takkunen. Comparison of multiple organ dysfunction scores in the prediction of hospital mortality in the critically ill. *Critical Care Medicine*, 30(8):1705, August 2002.

[6] Daleen Aragon Penoyer. Nurse staffing and patient outcomes in critical care: A concise review. *Critical Care Medicine*, 38(7):1521–1528, /07/01 2010.

[7] J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and LG Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure, 1996.

[8] Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the sofa score to predict outcome in critically ill patients. *JAMA, the journal of the American Medical Association*, 286(14):1754–1758, 2001.

[9] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[10] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[11] Frank H Guenther. Neural networks: Biological models and applications. *Internation Encyclopedia of the Social and Behavioural Sciences*, 2001.

[12] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. page 2643–2651, USA, 2013. Curran Associates Inc.

[13] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. pages 160–167. ACM, 07/05/2008.

[14] Nima Tajbakhsh, Jae Y. Shin, Suryakanth R. Gurudu, R. Todd Hurst, Christopher B. Kendall, Michael B. Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[15] J. Heaton. *Artificial Intelligence for Humans: Deep learning and neural networks*. Artificial Intelligence for Humans Series. Createspace Independent Publishing Platform, 2015.

[16] Manli Sun, Zhanjie Song, Xiaoheng Jiang, Jing Pan, and Yanwei Pang. Learning pooling for convolutional neural network. *Neurocomputing*, 224:96–104, Feb 8, 2017.

[17] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[18] Santiago Fernández, Alex Graves, and Jürgen Schmidhuber. An application of recurrent neural networks to discriminative keyword spotting. In *International Conference on Artificial Neural Networks*, pages 220–229. Springer, 2007.

[19] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] Romain Pirracchio, Maya L. Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J van der Laan. Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet. Respiratory Medicine*, 3(1):42–52, Jan 2015.

[22] Mervyn Singer, Craig C. M. Coopersmith, Richard R. S. Hotchkiss, Mitchell Levy, John C. Marshall, Greg G. Martin, Steven Opal, Gordon Rubenfeld, Tomvan Der T D Poll, Jean Louis Vincent, Derek Angus, Clifford S. Deutschman, Christopherwarren C. Seymour, Manu M. Shankar-Hari, Djillali Annane, Michael M. Bauer, Rinaldo Bellomo, Gordon Bernard, and Jean Daniel J D Chiche. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, Feb 2016.

[23] Jacob S. Calvert, Daniel A. Price, Uli K. Chettipally, Christopher W. Barton, Mitchell D. Feldman, Jana L. Hoffman, Melissa Jay, and Ritankar Das. A computational approach to early sepsis detection. *Computers in Biology and Medicine*, 74:69–73, 07 01, 2016.

[24] Hye Jin Kam and Ha Young Kim. Learning representations for the early detection of sepsis with deep neural networks. *Computers in Biology and Medicine*, 89:248–255, Oct 01, 2017.

[25] Mike Wu, Marzyeh Ghassemi, Mengling Feng, Leo A. Celi, Peter Szolovits, and Finale Doshi-Velez. Understanding vasopressor intervention and weaning: risk prediction in a public heterogeneous clinical time series database. *Journal of the American Medical Informatics Association: JAMIA*, 24(3):488–495, May 01, 2017.

[26] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding using deep networks. May 23, 2017.

[27] Andre S Fialho, Leo Anthony Celi, Federico Cismondi, SM Vieira, SR Reti, JMC Sousa, and SN Finkelstein. Disease-based modeling to predict fluid response in intensive care units. *Methods of information in medicine*, 52(6):494, 2013.

[28] Cátia M Salgado, Susana M Vieira, Luís F Mendonça, Stan Finkelstein, and João MC Sousa. Ensemble fuzzy models in personalized medicine: Application to vasopressors administration. *Engineering Applications of Artificial Intelligence*, 49:141–148, 2016.

[29] Cindy Crump, Sunil Saxena, Bruce Wilson, Patrick Farrell, Azhar Rafiq, and Christine Tsien Silvers. Using bayesian networks and rule-based trending to predict patient status in the intensive care unit. *AMIA Symposium*, 2009:124–128, Nov 14, 2009.

[30] Xiaowu Bai, Wenkui Yu, Wu Ji, Zhiliang Lin, Shanjun Tan, Kaipeng Duan, Yi Dong, Lin Xu, and Ning Li. Early versus delayed administration of norepinephrine in patients with septic shock. *Critical care*, 18(5):532, 2014.

[31] Jean-Michel Boles, Julian Bion, Alfred Connors, Margaret Herridge, Brian Marsh, Christian Melot, Ronald Pearl, Henry Silverman, Michael Stanchina,

Antoine Vieillard-Baron, et al. Weaning from mechanical ventilation. *European Respiratory Journal*, 29(5):1033–1056, 2007.

[32] Hatem Bouabana. Predicting deterrioration for patients in the intensive care unit. Master's thesis, University of Tampere, 2018.

[33] Sara Nisula, Kirsi-Maija Kaukonen, Suvi T Vaara, Anna-Maija Korhonen, Meri Poukkanen, Sari Karlsson, Mikko Haapio, Outi Inkinen, Ilkka Parviainen, Raili Suojaranta-Ylinen, et al. Incidence, risk factors and 90-day mortality of patients with acute kidney injury in Finnish intensive care units: the FINNAKI study. *Intensive care medicine*, 39(3):420–428, 2013.

[34] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[35] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 4580–4584. IEEE, 2015.

[36] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[37] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[38] Meghan Prin and Hannah Wunsch. International comparisons of intensive care: informing outcomes and improving standards. *Current opinion in critical care*, 18(6):700, 2012.

[39] Xuan Liu, Xiaoguang Wang, and Stan Matwin. Interpretable deep convolutional neural networks via meta-learning. *arXiv preprint arXiv:1802.00560*, 2018.