IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 March 2023, accepted 2 April 2023, date of publication 3 May 2023, date of current version 30 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3272745

RESEARCH ARTICLE

Modeling of Electronic Health Records for Time-Variant Event Learning Beyond Bio-Markers–A Case Study in Prostate Cancer

J. HERP[1,2], JAN-MATTHIAS BRAUN[1,2], M. L. CANTUARIA[1,2], ASHKAN TASHK[1], (Senior Member, IEEE), T. B. PEDERSEN[3], M. H. A. POULSEN[3,4,5], M. KROGH[6], E. S. NADIMI[1,2], (Senior Member, IEEE), AND S. P. SHEIKH[6,7]

[1]Unit of Applied Artificial Intelligence and Data Science, The Maersk Mc-Kinney Møller Institute, Faculty of Engineering, University of Southern Denmark, 5230 Odense, Denmark
[2]Center for Clinical Artificial Intelligence (CAI-X), 5000 Odense, Denmark
[3]Department of Urology, Odense University Hospital, 5000 Odense, Denmark
[4]Department of Urology, Hospital South West Jutland, 6700 Esbjerg, Denmark
[5]Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, 5000 Odense, Denmark
[6]Open, Odense University Hospital (OUH), 5000 Odense, Denmark
[7]Department of Clinical Biochemistry and Pharmacology, Institute of Clinical Research, Faculty of Health Sciences, University of Southern Denmark, 5000 Odense, Denmark

Corresponding author: Jan-Matthias Braun (j-mb@mmmi.sdu.dk)

This work was supported by the Danish Agency for Digitisation (Digitaliseringsstyrelsen) through the project "Better Diagnostics of Prostate Cancer (Bedre Diagnosticering af Prostatacancer)" under Grant DIGST 2019-5773.

ABSTRACT Electronic health records (EHR) of large populations constitute a vast untapped resource for data-driven diagnosis and disease progression. We develop a model capable of predicting future steps in a patient's journey for prostate cancer (PC) and its metastases without relying on direct biomarker-measurements on a set of 18 529 EHR. To this end, we 1) harmonise EHR without presumptions–events are sorted and grouped by fundamental a priori principles; 2) develop a new Long-Short-Term Memory (LSTM) recurrent neural network node for learning temporal relations, on which we build an autoencoder based model; 3) derive a graph representation based on unsupervised k-means clustering of events related to PC in the autoencoder's latent layer. We report 88 % predicting accuracy for the targeted metastasis-related events, and lower accuracies for more general events. The model gains interpretability with a graph representation illustrating the patient journey. Most importantly, we predict that 20 % of all PC diagnosed patients will progress into metastatic disease one visit ahead of time. For the remaining patients we can predict the next step in their journey. We conclude that the model based on the new LSTM node provides a valuable tool for earlier diagnosis of life threatening metastases and quality assurance of the procedure.

INDEX TERMS Autoencoder, electronic health records, event prediction, metastasis, prostate cancer, recurrent neural networks.

I. INTRODUCTION

Prostate cancer (PC) is the second most common cancer in men with 19 % prevalence, over 4 500 diagnoses in Denmark in 2019, and an absolute number of 1.3 M diagnoses worldwide. In 2019, PC was also the second most frequent cause of cancer related death with a lifetime risk of 15 % and accounting for 4.7 % of all registered death causes for men in Denmark [1], [2]. Although these numbers underline the risk related to PC, PC patient screening is under prioritised. Guidelines in Denmark [3], [4] do not recommend neither systematic, nor opportunistic early screening, given that PC

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

manifests itself seldom before the age of 50, and half of men of the age of 60 will be diagnosed with a clinically insignificant PC. However, autopsy studies have shown that PC can be detected significantly earlier than that [5]. Looking back over the last decades, only a minor reduction in PC mortality has been observed, even though advances in detection, treatment, life-prolonging and palliation have been made. One of the main challenges with PC is the lack of ability to predict which of the patients will develop metastatic and thereby lethal PC – and which of the patients will continue to have an indolent PC. Biomarker panels and pathology nomograms have yet to show they can predict the course of the disease for a PC patient, which has lead to massive over-treatment while having marginal effect on PC mortality. In consequence, foreseeing the course of a PC patient, or any cancer patient, is of tremendous clinical relevance to select interventions with best patient-outcome while minimising side effects.

Since the introduction of prostate specific antigen (PSA), diagnoses and disease progression of PC have been guided by this bio-marker. Under normal conditions, only low levels of PSA can be detected in the blood, and the increase of serum PSA found in PC can represent abnormalities in prostate gland architecture. Historically, PSA has been utilised for monitoring the progression of patients already diagnosed with PC, or for recurrent curative therapy. In 1987, a large study demonstrated PSA as the most sensitive bio-marker for monitoring PC progression [6]. Here, it was shown that the PSA level increases with advancing clinical stage and that it is useful for detecting recurrence after therapy. Subsequent studies have explored the PSA's ability for early disease prediction. In 1991, it was demonstrated that a combination of PSA measurements of more than 4.0 ng/mL with other clinical findings improved the early detection of PC among 1 653 healthy men without predisposition of cancer [7], [8].

However, like other diagnostic tools, so does PSA have its limitations. For instance, the diagnostic test performance of PSA is volatile. Particularly, the specificity ranges from 20% to 40% [9]. The relatively low specificity can be explained, as other non-cancerous circumstances, such as inflammation, infection and benign prostatic hyperplasia, can elevate the PSA level. Furthermore, up to 15% of men with low levels of PSA have PC [10], [11]. It is therefore not possible to reliably predict the risk of severe cases of PC with PSA alone. Adding to this, PSA has lead to an increase in detection of insignificant PC findings [5].

Despite its shortcomings, PSA remains an inexpensive and sensitive bio-marker for PC detection and disease progression monitoring. These features have made PSA usage common in screening procedures such that additional bio-markers for clinical evaluation of PC often are obtained after the initial diagnoses. As a consequence, PSA retains its place as a primary clinical tool for PC diagnostics alongside imaging and biopsy based approaches – unless new methodology can be put forward to include other historic medical data previously thought to be unrelated to PC.

While PSA measurements retain a role in diagnosing a patient with PC, the Danish National Patient Registry (DNPR) offers an untapped resource in terms of historic clinical recordings for other than PC related diagnostic procedures. These records are digitised and can be accessed as electronic health records (EHR). Schmidt et al. [12] provide a comprehensive review of the content, quality, and research potential of the DNPR. In the time during 1977 to 2012, the DNPR registered more than 8 million persons with detailed administrative and clinical data. In addition, the DNPR provides data sources for disease identification, examination records, in-hospital medical treatments, and surgical procedures. Schmidt et al. [12] value the DNPR as a source for long-term temporal trend analysis.

Under the hypothesis that EHR contain information about disease progression, these data can be used for the development of predictive models able to predict patient trajectories under different medical contexts. These models are expected to help improve the quality of medical assessment in general, but also to propel the development of personalised medicine based on individual, i.e. per patient, predictions of disease progression [13]. Based on such predictions, individual risks can be estimated or potential treatment options evaluated. A good model should be able to be applied to unfiltered data and ideally be able to identify relevant elements and time scales in EHRs, to allow for the generalised application to different diseases without manual tuning.

In this paper, we present a generalised predictive model derived from EHR data using PC as a case study. We thereby aim to predict the series of events (diagnostic, treatment, and procedure codes) up to and from a diagnosis of PC. Subsequently, we propose a way this model can abstract higher level patient journeys without domain knowledge. This model is free of pre-selected diagnostic biomarkers, such as PSA or testosterone. The contribution of this work is twofold. i) After an introduction to the patient data included in this study (Section II), we present a tool to model and predict non-uniform sampled EHR events in Section III. We are offering an adaptation of long-short-term-memory (LSTM) recurrent neural networks (RNNs) that is able to learn relevant temporal relations between EHR events while discarding redundant and unrelated events. ii) In Section IV, we facilitate the results, validation, and discussion of the model outcomes with respect to PC, highlighting the clinical relevance for data driven solutions, such as the proposed model. In this part of the manuscript, we are focusing on the model's ability to generalise the prediction of patient journeys.

We conclude that the presented model is able learn from unfiltered datasets based on EHR records, thereby achieving superior performance in domain relevant predictions, highlighting the ability to extract relevant time scales and events in an unsupervised manner.

II. CLINICAL DATA DESCRIPTION AND PREPARATION

This study is based on a dataset composed of 18 529 patients diagnosed with PC at least once between January 1, 2004, and

December 31, 2019, in a state-funded clinic in the Region of Southern Denmark.[1] The average age of a patient in this study on January 1, 2004 is 62 years.

A diagnosis for PC is defined as the presence of a DC619 code in a patient's journal using the Danish Care Classification System (SKS, from Danish, Sundhedsvasenets Klassifikations System). The SKS codes associated with PC and a possible metastatic disease are listed in Table 1. The average age for the first PC diagnosis is 71±9 years, with the youngest patient being 30 and the oldest 98 years old at the time of diagnosis. Based on metastasis related codes in Table 1, 12.6 % of the patients are identified to have a metastasis related PC diagnoses - which is below the estimated 20−30 % worldwide [14]. For these patients, the average age for their first metastasis related diagnosis is 75 ± 8 years. The average time between the first diagnosis of PC and a confirmation of a metastatic disease is 1 132 days (approx. 3 years), with the lower and upper quartile measuring 0 and 1 988 days (approx. 5.5 years) respectively.

A. DATA DESCRIPTION

The data provided by the Region of Southern Denmark contains a wide range of SKS codes. These codes are used for sharing and delivering structured information for different information systems. This study does not exclude any SKS codes that might be present in the underlying dataset. For an exhaustive list of all SKS codes we refer the interested reader to the medinfo.dk database [15].

The SKS database consists of 17 classes, ranging from diagnostic codes, over medical procedure, to administrative codes. These are further divided into chapters and sections. Table 2 shows this hierarchy by the example of *malignant plasma cells neoplasms* (DC90). DC90 is categorised under classes *classification of disease and health related conditions* (D), in the chapter of Neoplasms (Chapter 2), in section *Cancer in lymphatic and hematopoietic tissue* (Section 15), and has four sub classifications *DC900*, *DC901*, *DC902*, and *DC903*.

In a Danish patient journal, these codes are stored as what we refer to an event. An event is a code recorded at a specific time, see table 3a for an example.

B. DATA PREPARATION

For each possible event in the dataset, we assigned a unique variable E_n, with $1 \leq n \leq N$, N the number of unique event codes. Table 3 shows how the aforementioned unstructured SKS codes are converted to a running variable E_n. Given the sparse and volatile nature of patient data, it is likely that any E_n can be less frequent than others, or simple appear by chance. In order to avoid modelling of sparse events which are not represented well in the dataset, we agglomerated (clustered) these events, exploiting the hierarchical structure of the event codes, i.e., we included them into higher relevant medical categories, if possible. Table 2 illustrates this process for *myelomatosis leikemia* (DC900) and *solitary bone plasmacytoma* (DC903). Both have a low frequency in the dataset, but since they both belong to the same subsection of "Malignant plasma cells neoplasms" (DC90) that contains frequent events, we can agglomerate DC900 and DC903 into an eventset DC90, and thereby not lose potential information these events may carry. For a full description of the compilation of the event list we refer to Appendix A.

After having compiled the complete list of events, we focused on recording the time between any set of events. In order to model temporal dependencies, we assumed that a random variable δ_t can model the time between, since, or to events.

Carrying on with the example of Table 3, we created one eventset $\mathcal{E}_t$ of zeros and ones, where ones stand for events coded for that visit, and determined a numerical value for the time δ_t since the previous visit. Concatenating the eventset and the time since the previous visit (cmp. Eq. (21), Table 3b), we arrived at a data vector $\mathbf{x}_t^{(p)} = \mathcal{E}_t \mid \delta_t$ coding the t^{th} visit of patient (p), with P the number of patients:

$$\mathbf{x}_1^{(1)} = [1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^\top \tag{1a}$$
$$\mathbf{x}_2^{(1)} = [0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 45]^\top \tag{1b}$$
$$\mathbf{x}_3^{(1)} = [0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 3]^\top \tag{1c}$$
$$\mathbf{x}_4^{(1)} = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 41]^\top \tag{1d}$$
$$\mathbf{x}_1^{(2)} = [0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0]^\top \tag{1e}$$

We denoted the dimension of these vectors as $N + 1$, which is the sum of the number of unique events in all patient journals plus the time since the last visit. Note that the first visit always codes with $\delta_t = 0$. Collecting all data vectors which code for patient p's visits, we can define a matrix where columns i contain the data vectors $\mathbf{x}_i$:

$$\mathbf{X}_p = [\mathbf{x}_1, \ldots, \mathbf{x}_{T_p}] \in \mathbb{R}^{m \times T_p}. \tag{2}$$

As the number T_p of visits in patient p's journal depends on the patient under consideration, the matrices $\mathbf{X}_p$ have varying numbers of columns. For further processing, we zero padded them to have the same number $T = \max_{1 \leq p \leq P} T_p$ of columns and then turned them into a tensor, e.g.:

$$\boldsymbol{\chi} = [\mathbf{X}_1, \ldots, \mathbf{X}_P]^\top \in \mathbb{R}^{P \times m \times T} \tag{3}$$
$$= \begin{bmatrix} [\mathbf{0}, \ldots, \mathbf{0}, \mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{T_1}^{(1)}] \in \mathbb{R}^{m \times T} \\ \vdots \\ [\mathbf{x}_1^{(P)}, \ldots, \mathbf{x}_t^{(P)}, \ldots, \mathbf{x}_T^{(P)}] \in \mathbb{R}^{m \times T} \end{bmatrix} \tag{4}$$

III. METHODOLOGY

Our model is based on the hypothesis that there exists a causal relation, or at least a correlation, between consecutive events in EHR, especially, we make no assumption on why the data for these events were acquired.

[1]The Region of Southern Denmark is an administrative entity which operates the healthcare service in its 22 municipalities, covering health, social services and special education, psychiatry, and regional development.

TABLE 1. SKS codes associated with PC and metastatic PC translated from Danish.

Code name	Code Description (translated)	Code Description (danish)
DC619	Prostate cancer	Prostatakræft
DC619M	Prostate cancer with metastasis	Prostatakræft med metastaser
AZCD40	No metastasis	Ingen fjernmetastaser
AZCD49	Information on metastasis is missing	Oplysning om fjernmetastaser mangler
AZCD41	Distant metastasis	Fjernmetastaser
AZCD41A	Non-nearby lymph nodes metastasis	Ikke-regionale lymfeknudemetastaser
AZCD41B	Bone tumor	Knoglemetastaser
AZCD41C	Metastasis in other organs	Andre organ-metastaser

TABLE 2. SKS hierarchy example for Malignant plasma cells neoplasms (DC90), including the absolute frequency # of the codes in the underlying dataset. Note, that the colour codes for the frequency of the event being high (green), or low (red, beyond a threshold of 50 events). Based on the frequency, we agglomerated the low frequency events into events on a higher hierarchy level, until the sum of event frequencies crosses the threshold. Events DC900 and DC903, for example, wouldn't be taken into account individually, but can be agglomerated into a higher level event DC90. Appendix A contains a detailed description of the procedure.

Code	Description (translated)	Class	Chapter	Section	Subsection	#
D	Classification of disease and health related conditions					21203
Chapter 2	Chapter II: Neoplasms[DC00-DD48]	D				10807
Section 15	Cancer in lymphatic and hematopoietic tissue [DC81-DC96]	D	Chapter 2			436
DC90	Malignant plasma cells neoplasms	D	Chapter 2	Section 15		139
DC900	**Myelomatosis leukemia**	**D**	**Chapter 2**	**Section 15**	**DC90**	**2**
DC901	Plasma cell leukemia	D	Chapter 2	Section 15	DC90	73
DC902	Solitary non-bone plasmacytoma	D	Chapter 2	Section 15	DC90	50
DC903	**Solitary bone plasmacytoma**	**D**	**Chapter 2**	**Section 15**	**DC90**	**14**

Modelling EHR is no new endeavour, rule-based and regression methods have long been playing a role in diagnostic decision support and provide accurate results in selected disease detection and prediction studies [16], [17], [18], [19]. With the increase in data volume and computational power of the last decades, neural networks are gradually replacing regression and statistical models as work horse in data mining and data driven modelling. However, big data analytics comes at the price that the interpretability of the models often decreases. Models such as presented in Lipton et al. and Choi et al. [20], [21] allow for modelling patient data with high accuracy. Nonetheless, these models do not inform about the patient journey, i.e. they do not give an account for when information are relevant in time. Pham et al. [22] include the time between events as a frequency component in their modelling. Pham's DeepCare model weights the time between events as a decreasing function $h(\delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}}) = 1/\delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}}$. Baytas et al. [23] proposed a time-aware LSTM structure, by adding a *discounted* forget gate in the LSTM node (see Section III-A). While this study follows the same approach on the LSTM node level, Baytas has its focus on modelling decreasing times similar to Pham. Further, they propose a parametric approach to weight the times for informed time scales measured in days [23]. Still, this approach implies and requires knowledge on the dominant time-scales for the diseases modelled. In contrast to the assumptions Baytas et al. and Pham implemented when including time into their models, the proposed model maintains an unsupervised approach to learning relevant time scales, which reflects the lack of knowledge concerning the driving time-scales within PC modelling. Therefore, we will use and compare the proposed model's performance to Pham's DeepCare model as a state of the art reference model with time-awareness.

$$h \propto \left(\frac{\delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}}}{60}, \left(\frac{\delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}}}{180} \right)^2, \left(\frac{\delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}}}{360} \right)^3 \right), \quad (5)$$

In the following, we will lay out the foundation for a specific kind of neural networks, a so called auto encoder. While the fundamental operations are similar to that of existing models, and temporal considerations have been introduced as hard rules by Pham and Baytas et al. [22], [23], we will arrive at a model interpretation that generalises the patient journey for PC, without assuming any underlying decaying structure or parameterisation on the temporal side – hence we call the proposed model *sample frequency independent*. It is thus the Iν-LSTM's main contribution to learn relevant time scales in an unsupervised manner from the presented data.

A. INDEPENDENT FREQUENCY LSTM (Iν-LSTM)

There are many ways in which event patterns can be abstracted, ranging from a priori associated rule mining [24] to RNNs [25]. Given the recurrent and successive nature of the data, particular the volume at hand, RNNs are a suitable tool to handle the task. LSTM RNNs, proposed first in 1997 [26] and in their current version by Graves et al. [27], are now widely used in a variety of applications.

TABLE 3. Example of the conversion process from SKS codes (Table 3a) to a running variable (Table 3b) and from there to an eventset $\mathcal{E}_t$. We concatenated the time δ_t (also Table 3b) to the eventset to create a data vector which the presented method can work on.

Patient	Time of Visit	SKS codes					
1	06/04/2010	DH911		BDDD8			
1	07/19/2010	DC619	DR391	ZZ1291	550104U		
1	07/22/2010	DC619		BWDB80	550104U	AZCD41	AZCD15C
1	09/01/2010	DH911					
2	06/23/2009	DZ090		ZZ9990	BDDD8		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) Example of a subset of anonymised SKS data for two patients. Each row lists the SKS codes which were registered for one visit.

Patient	Time of Visit	Events $\mathcal{E}$						Eventset $\mathcal{E}_t$											δ_t
								E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	
1	06/04/2010	E_1		E_2				1	1										./.
1	07/19/2010	E_3	E_4	E_5	E_6					1	1	1	1						45
1	07/22/2010	E_3		E_7	E_6	E_8	E_9			1			1	1	1	1			3
1	09/01/2010	E_1								1									41
2	06/23/2009	E_{10}		E_{11}	E_2				1								1	1	./.
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

data vector $\mathbf{x}_t$

(b) SKS codes from Table IIIa uniquely mapped to a running variable E_i. Each SKS code is replaced in order of occurrence, i.e., in this example "DH911" is mapped to E_1. This mapping allows to later construct eventsets $\mathcal{E}_t$ as vectors of boolean values encoding the existence of a specific event E_i by a 1 at position vector element i for the visit with time t, (1). The time span δ_t measures the time since the previous visit. Concatenating the event set $\mathcal{E}_t$ with the time span δ_t, we create data vectors $\mathbf{x}_t$, which capture all information considered per patient visit.

LSTMs are explicitly designed to retain long-term temporal dependencies. Like all RNNs, they are composed of repeated computational nodes and can produce a sequence to sequence output. Fig. 1a shows a high-level view of how a (recurrent) LSTM node conserves historic data for reasoning. Here $\mathbf{x}_t$ and $\mathbf{h}_t$ denote the input and output sequences, respectively, up to sequence length S. In contrast to naïve neural network nodes, i.e. nodes that are composed of only one activation function, LSTMs are composed of a gated structure whose components can be characterised as input-, output-, forget-gates, a current and candidate memory cell, as well as a hidden state variable. Given weight matrices $\mathbf{W}$, $\mathbf{U}$ and bias vector $\mathbf{b}$ in a computational node structure, as shown in Fig. 1b, a LSTM neural network node is defined through four operations:

LSTM Forget Gate

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \cdot f\left(\mathcal{E}_t\right) + \mathbf{U}_f \cdot \mathbf{h}_{t-1} + \mathbf{b}_f\right) \tag{6}$$

LSTM Memory

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i \cdot \mathbf{h}_{t-1} + \mathbf{U}_i \cdot f\left(\mathcal{E}_t\right) + \mathbf{b}_i\right) \tag{7a}$$

$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}_c \cdot \mathbf{h}_{t-1} + \mathbf{U}_c \cdot f\left(\mathcal{E}_t\right) + \mathbf{b}_c\right) \tag{7b}$$

LSTM Node Update

$$\mathbf{c}_t = \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \tag{8}$$

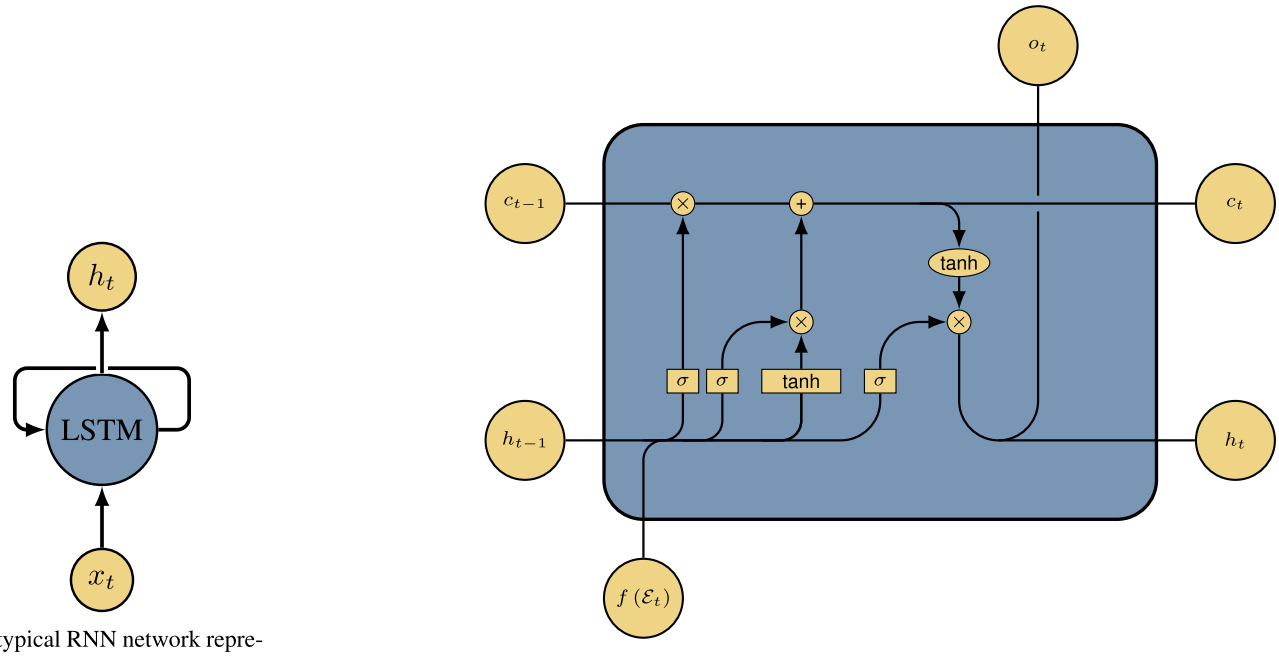
LSTM Node Activation

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o \cdot \mathbf{h}_{t-1} + \mathbf{U}_o \cdot f\left(\mathcal{E}_t\right) + \mathbf{b}_0\right) \tag{9a}$$

$$\mathbf{h}_t = \mathbf{o}_t \circ \tanh\left(\mathbf{c}_t\right) \tag{9b}$$

A RNN composed of LSTM nodes described by Eqs. (6) to (9b) assumes temporal regularity between events. This implicit assumption in the RNN architecture, i.e. that the elapsed time between patient visits is uniformly sampled, makes it unsuited for the dataset at hand. In general, patient visits are not scheduled to follow any distribution and underlie random influences from their environments. Non-uniform time between events is in itself not of concern if the variability of one patient's behaviour can be generalised to many patients. However, patients come with their very individual schedule, thus, temporal variability for the times between the same visits of different patients is to be expected. In the following, we propose an adaptation to the LSTM structure in order to include the aforementioned time between events, $\delta_{t'} \equiv \delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}}$.

We treat the irregularities in the timing of patient visits as an inter- and intra-patient independent sample frequency ν. The independent sample frequency model, Iν-LSTM, adds a separate temporal input to the LSTM structure that amplifies or suppresses an eventset depending on the time that passed since the previous visit. In order to account for long gaps between events, an additional forget gate is introduced. The



(a) A typical RNN network representation.

(b) Illustration of a traditional LSTM [27].

FIGURE 1. **Generalisation of a LSTM RNN.**

proposed Iν-LSTM is laid out in Fig. 2 - following the LSTM node structure as proposed by Baytas et al. [23].

Given the same weight matrices as for a LSTM node, a Iν-LSTM is defined through five operations and a temporal map of type $h(\delta_t, \mathbf{b}_t) = \cosh^{-2}(b_1\delta_t - b_2)$:

Iν-LSTM Memory Gate same as in Eqs. (7a), (7b) and

$$\hat{\mathbf{c}}_{t-1} = \tanh\left(\mathbf{U}_a \cdot \mathbf{c}_{t-1} + \mathbf{b}_a\right) \tag{10a}$$

$$\hat{\hat{\mathbf{c}}}_{t-1} = \hat{\mathbf{c}}_{t-1} \circ h(\delta_t, \mathbf{b}_t) \tag{10b}$$

$$\hat{\hat{\hat{\mathbf{c}}}}_{t-1} = \mathbf{c}_{t-1} - \hat{\mathbf{c}}_{t-1} \tag{10c}$$

$$\tilde{\tilde{\mathbf{c}}}_{t-1} = \hat{\hat{\hat{\mathbf{c}}}}_{t-1} + \hat{\hat{\mathbf{c}}}_{t-1} \tag{10d}$$

Iν-LSTM Node Update

$$\mathbf{c}_t = \mathbf{f}_t \circ \tilde{\tilde{\mathbf{c}}}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \tag{11}$$

Iν-LSTM Node Activation same as in Eqs. (9a) and (9b).

B. EVENT PREDICTION

Considering a chain of successive eventsets $\mathcal{E}_t^{(P)}$, we assume an ordered ($t < t'$) pair of eventsets, $\mathcal{E}_t^{(P)}$ and $\mathcal{E}_{t'}^{(P)}$, which is causally related or correlated. If we can predict the latter eventset $\mathcal{E}_{t'}^{(P)}$, when using $\mathcal{E}_t^{(P)}$ as input, we say that $\mathcal{E}_{t'}^{(P)}$ depends on $\mathcal{E}_t^{(P)}$:

$$\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)} \quad t < t' \tag{12a}$$

$$\mathcal{E}_t^{(P)}, \mathcal{E}_{t'}^{(P)} \subseteq \mathcal{E} \tag{12b}$$

Many neural networks are concerned with multiple-input-single-output or multiple-input-multiple-output maps. In this section we want to address the prediction of the next events in a sequence. In this study, the input space is of size $N+1$ and we wish to retain the option to predict any event in the dataset, thus the desired output space is of the dimension N. As we can argue that the diagnostic procedure is taking place in fewer dimensions and that we aim for an abstraction of the underlying diagnostic process, a desired model should be able to encode information in a lower dimensional latent space $\mathcal{L}$ of dimension l.

Sequence-to-sequence autoencoders [28] have been used to learn a representation of data by mapping from an input $\mathcal{X}$ to a desired output $\mathcal{Y}$. In order to fulfil Eqs. (12), we are thus seeking a map between an encoder ψ and a decoder ϕ, with latent space $\mathcal{L}$ (Fig. 3). For an input in $\mathbb{R}^{N+1}$ the encoder

$$\psi(\text{I}\nu\text{-LSTM}(m_0,S),\text{LSTM}_1(m_1),\ldots,\text{LSTM}_k(m_k)) : \mathcal{X} \to \mathcal{L}, \tag{13}$$

consists of a layer of m Iν-LSTM nodes with sequence length S and subsequent k LSTM layers with decreasing amount of nodes, $m_0 > m_1 > \ldots > m_k$, that map into $\mathcal{L}$ of size l. The latent space is defined as an element-wise tanh activation function, σ, with weight matrix $\mathbf{W}$ and bias vector $\mathbf{b}$:

$$\mathbf{l} = \sigma\left(\mathbf{W}\psi(\mathbf{x}_t) + \mathbf{b}\right), \quad \mathbf{l} \in \mathcal{L}. \tag{14}$$

The decoder then employs a sequence of LSTM layers with increasing number of nodes to map to the desired output in $\mathbb{R}^N$:

$$\phi(\text{LSTM}_{k'}(m_k), \ldots, \text{LSTM}_{0'}(m_0)) : \mathcal{L} \to \mathcal{Y} \tag{15}$$

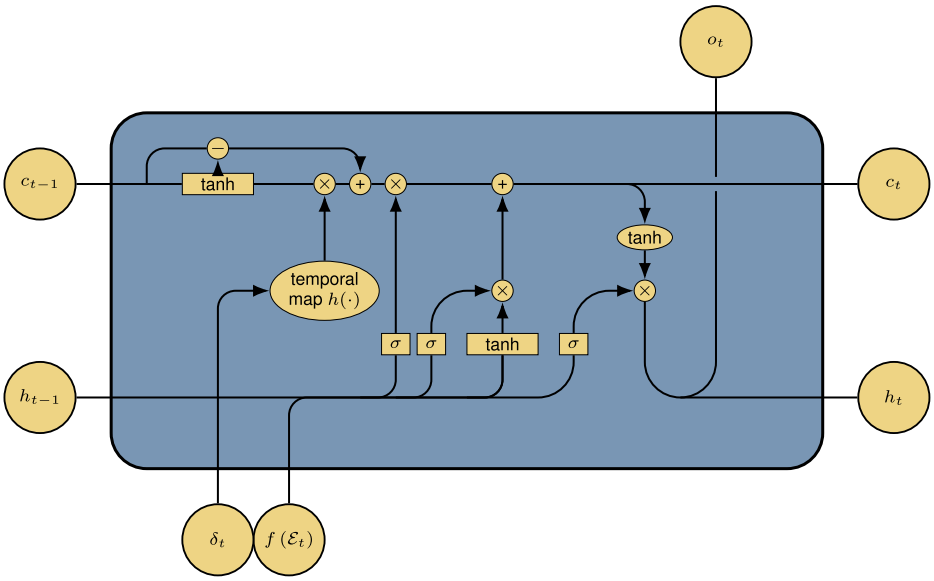


FIGURE 2. Illustration of the proposed independent frequency adjusted LSTM (Iν-LSTM) node. The modifications to the original LSTM (Fig. 1b) are added on the left side, to incorporate the temporal information δ_t into the memory gate update. During training, the temporal map $h(\cdot)$ learns relevant timespans from the presented data.

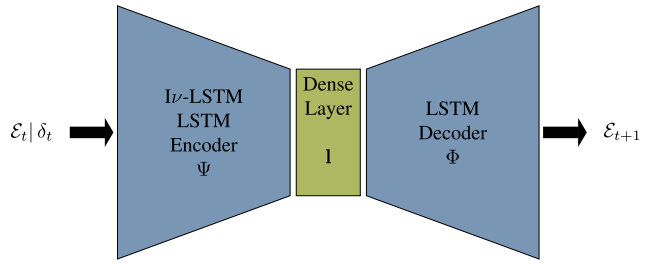


FIGURE 3. Schematics of an autoencoder, the high dimensional input on the left side is transferred into an intermediate representation, the Dense Layer l, before projecting again into a higher dimensional representation, a prediction of the next eventset. Thus, the dense representation l has to retain relevant information, to be able to project back to a high dimensional eventset.

TABLE 4. Model selection, including the depth k of the encoder Ψ and k' of the decoder Φ. $k = k'$, the LSTM layer sizes m_i, the sequence length S, and the latent space dimension l.

Model	$k = k'$	$[m_0 > m_1 > \ldots > m_k]$	S	l
Iν-LSTM$_1$	3	[75, 50, 40, 20]	5	5
Iν-LSTM$_2$	2	[50, 40, 20]	5	5
Iν-LSTM$_3$	3	[75, 50, 40, 20]	5	3
Iν-LSTM$_4$	2	[50, 40, 20]	2	5
Iν-LSTM$_5$	3	[75, 50, 40, 20]	2	3
Iν-LSTM$_6$	2	[50, 40, 20]	2	3
LSTM$_1$	3	[75, 50, 40, 20]	5	5
LSTM$_2$	2	[50, 40, 20]	5	5
LSTM$_3$	3	[75, 50, 40, 20]	2	3
LSTM$_4$	2	[50, 40, 20]	2	3

From here on we use

$$\text{I}\nu\text{-LSTM}(k, k', [m_0, \ldots, m_k], S, l) \tag{16}$$

for autoencoders of different topology, respectively LSTM $(k, k', [m_0, \ldots, m_k], S, l)$ when only LSTM layers are used. Training of the autoencoders is achieved by minimising the root-mean-squared error. For an exhaustive list of different loss function and optimisation techniques we refer the interested reader to Bianchi et al., Goodfellow et al., and Bishop [25], [28], [29], [30].

C. MODEL TRAINING AND VALIDATION

We are proposing topologies for Iν-LSTMs and LSTM auto encoders. The models are constrained to be symmetric w.r.t. the number of layers ($k = k'$) and nodes in each layer, and allow the width of each layer to be decreasing or increasing with either:

$$[m_0 > m_1 > \ldots > m_k] = \begin{cases} [75, 50, 40, 20] & \text{for } k = 3 \\ [50, 40, 20] & \text{for } k = 2 \end{cases} \tag{17}$$

Further, the latent dimension l is limited to be either 5 or 3 nodes wide, and the sequence lengths S are limited to 5 or 2 time instances (Table 4).

Alongside, we are training DeepCare [22], Associated Rule Mining [24], and Random Forest classification for comparison.

Model performance is evaluated in terms of a Jaccard coefficient based accuracy (i) for predicting the next PC related eventset including metastases (**Acc. PC w. Metastases**), i.e. the intersection over union of PC metastases related events only, (ii) for predicting the next PC related eventset excluding

TABLE 5. Model comparison for test results in terms of accuracy, sensitivity, and specificity for predicting events related to prostate cancer with metastases (first three columns), accuracy for events related to prostate cancer with and without metastases (fourth column), and prediction accuracy for all events in the dataset (in the last column). The best performance in each category is highlighted in bold.

Model	Acc. PC w. Metastases	Sen. PC w. Metastases	Spe. PC w. Metastases	Acc. PC	Acc. $\mathcal{E}_t$
Iν-LSTM$_1$	0.87	**0.76**	0.89	**0.78**	**0.64**
Iν-LSTM$_2$	**0.88**	0.71	**0.92**	0.73	0.62
Iν-LSTM$_3$	0.81	0.70	0.87	0.72	0.58
Iν-LSTM$_4$	0.82	0.71	0.89	0.73	0.60
Iν-LSTM$_5$	0.81	0.69	0.87	0.72	0.56
Iν-LSTM$_6$	0.69	0.70	0.62	0.72	0.49
LSTM$_1$	0.68	0.68	0.69	0.58	0.52
LSTM$_2$	0.68	0.67	0.69	0.58	0.49
LSTM$_3$	0.61	0.62	0.61	0.52	0.47
LSTM$_4$	0.59	0.58	0.59	0.51	0.42
DeepCare	0.55	0.55	0.56	0.59	0.49
Associated Rule Mining	0.53	0.53	0.54	0.46	0.40
Random Forest	0.46	0.39	0.50	0.44	0.31

metastases (**Acc. PC**), i.e. the intersection over union of PC related events only, and (iii) for predicting the next following eventset (**Acc.** $\mathcal{E}_t$), i.e. the intersection over union between the predicted eventset and the actual eventset. For the first case, the sensitivity (**Sen. PC w. Metastases**) and specificity (**Spe. PC w. Metastases**) are provided as well.

IV. RESULTS, VALIDATION, AND DISCUSSION

A. MODEL PERFORMANCE AND COMPARISON

After data preparation, the models are trained on 10 000 out of 18 529 patients, in a 70 % to 30 % training to validation split. The remainder of 8 529 patients is reserved for testing. The results for model validation of the proposed and literature models are summarised in Table 5.

Concerning the accuracy for the prediction of events with respect to metastatic PC, the proposed models scored between 0.59 to 0.88. Apart from of Iν-LSTM$_6$, the independent frequency models performed significantly better compared to the simple LSTM or literature models. In terms of sensitivity, the range of independent frequency models comprises 0.69 to 0.76, while specificity ranges from 0.87 to 0.92, not including the measure for Iν-LSTM$_6$ with 0.62. Iν-LSTM$_1$ and Iν-LSTM$_2$ perform overall better than the other tested models, where Iν-LSTM$_2$ presents marginally better accuracy and specificity than Iν-LSTM$_1$. Conditioned on the combinations of the encoder topology, the latent space dimension, and the input sequence length, as given by Table 4, the model accuracy is most sensitive to changes in the latent dimension and sequence length – the latter to a smaller degree. We conclude that three latent dimensions are not as suited to facilitate and discriminate an embedded representation of the sequential data as higher dimensional latent spaces are. The lower influence of the sequence length S could be understood in terms of the LSTMs internal memory capacity already providing a trace of the previous sequence. The same behaviour can be observed for **Acc. PC** and **Acc.** $\mathcal{E}_t$. Due to the high accuracy in predicting metastatic PC and the lower computational complexity than Iν-LSTM$_1$, we select Iν-LSTM$_2$ as the best candidate model for further investigation for the remainder of this work. For a 10-fold cross-validation of the two best performing models we refer the reader to Appendix C.

When comparing the accuracies between predicting events for the sets **PC w. Metastases**, **PC**, and $\mathcal{E}_t$, we can see that the accuracy is higher the more specialised the subset is, i.e., eventsets for **PC w. Metastases** show the best prediction accuracy, whereas the generic eventsets $\mathcal{E}_t$ are hardest to predict. This observation holds for almost all models, with the exceptions of DeepCare, Random Forests, and Iν-LSTM$_6$. This effect is most pronounced for the here presented Iν-LSTM family of models. We argue that this effect represents the Iν-LSTM models' ability to learn relevant time frames from the data. As all patients were selected by the property of having a PC diagnosis, we can expect that all models should be able to pick up common sequences of PC events, whereas all models should have trouble to properly predict eventsets which are not part of the common set of diagnoses but randomly contributed by individual patients and therefore do not correlate with common events. To be able to provide better predictions, a model needs to be able to focus on the relevant events and neglect irrelevant data. The ability to extract the time scales relevant for a disease's progression and treatment can help to make this distinction and is a unique feature of the here presented Iν-LSTMs.

For DeepCare, we observe that the accuracy for predicting events related to **PC w. Metastases** is in fact lower than the accuracy for predicting events related to **PC**. Otherwise, the performance is close to that of standard LSTMs. In comparison to Iν-LSTMs, DeepCare models have fixed time scales which are included with a $1/\delta_t$-characteristic. Thus, apart from the topology, the proposed model is the more flexible

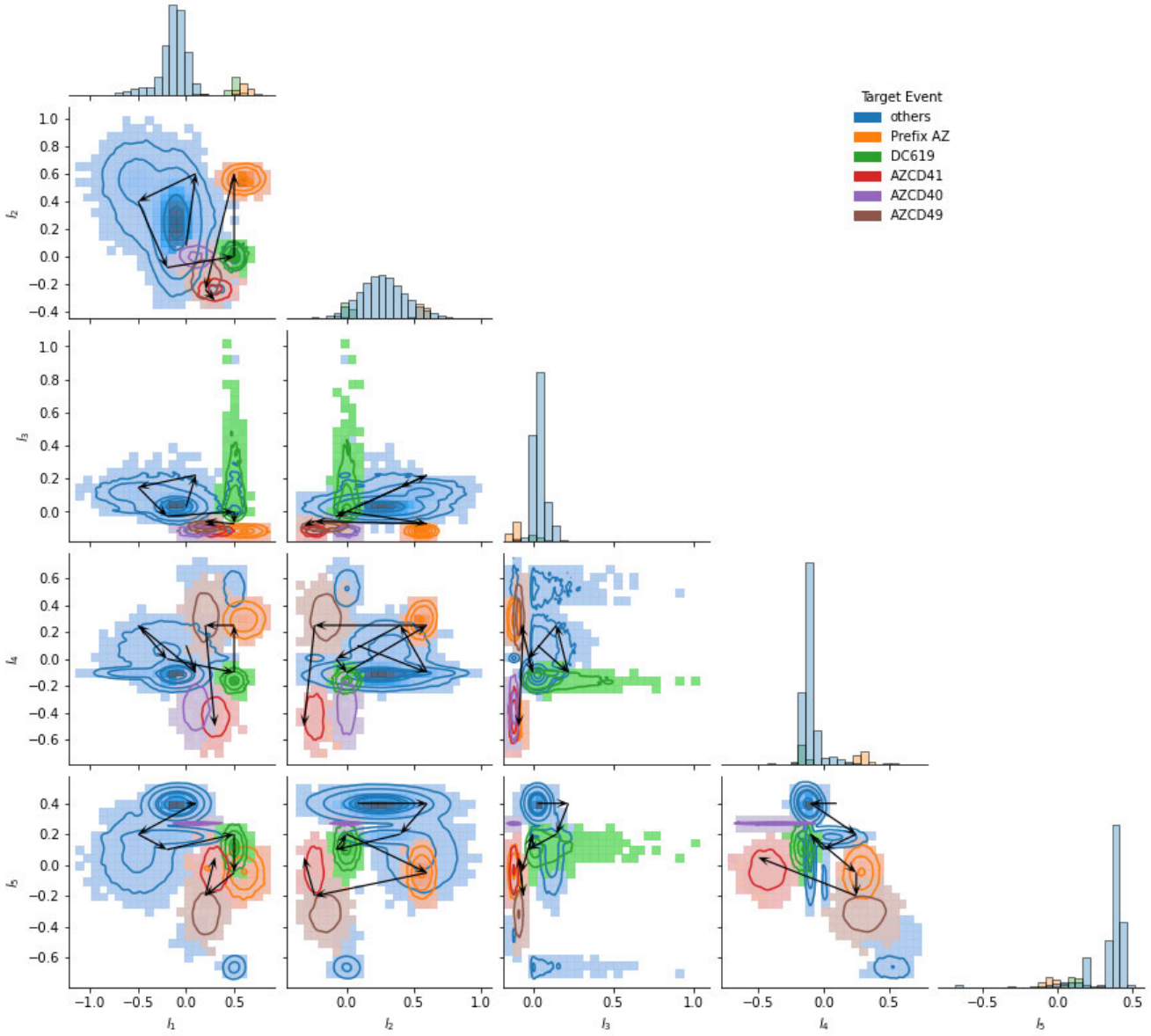


FIGURE 4. Scatter matrix of Iν – $LSTM_2$ latent layer. Arrows: superimposed single patient journey.

approach; weighting time contributions is the only difference which can be taken into account when explaining the difference in the ability to learn the specialised underlying structure in the EHR. It is noteworthy that the introduced temporal map is learned unsupervised and therefore is able to adapt to time scales inherent to the presented data. The question whether this advantage of the Iν-LSTMs can be used to improve the prediction of, e.g., risk of metastases or PC progression in general, is the subject of ongoing investigations.

The other literature candidates, Associated Rule Mining and Random Forests, scale poorly with large data, thus they are unsuited in this case.

B. LATENT LAYER INTERPRETATION

Given the nature of autoencoders, we can investigate the lower dimensional embedded structure when passing information through the network. For each eventset $\mathcal{E}_t$, we store the latent vector $\mathbf{l}$. Fig. 4 shows the latent space $\mathcal{L}$ as a scatter matrix for all eventsets in the training data. Here the data is represented as a density plot over all captured latent vectors $\mathbf{l}$ over all pairwise permutations of latent dimensions and histograms when projected into single dimensions. For a more intuitive interpretation of the latent space we have labelled each cluster for visual purposes only (for the full clustering procedure see Appendix B) to a target eventset as:

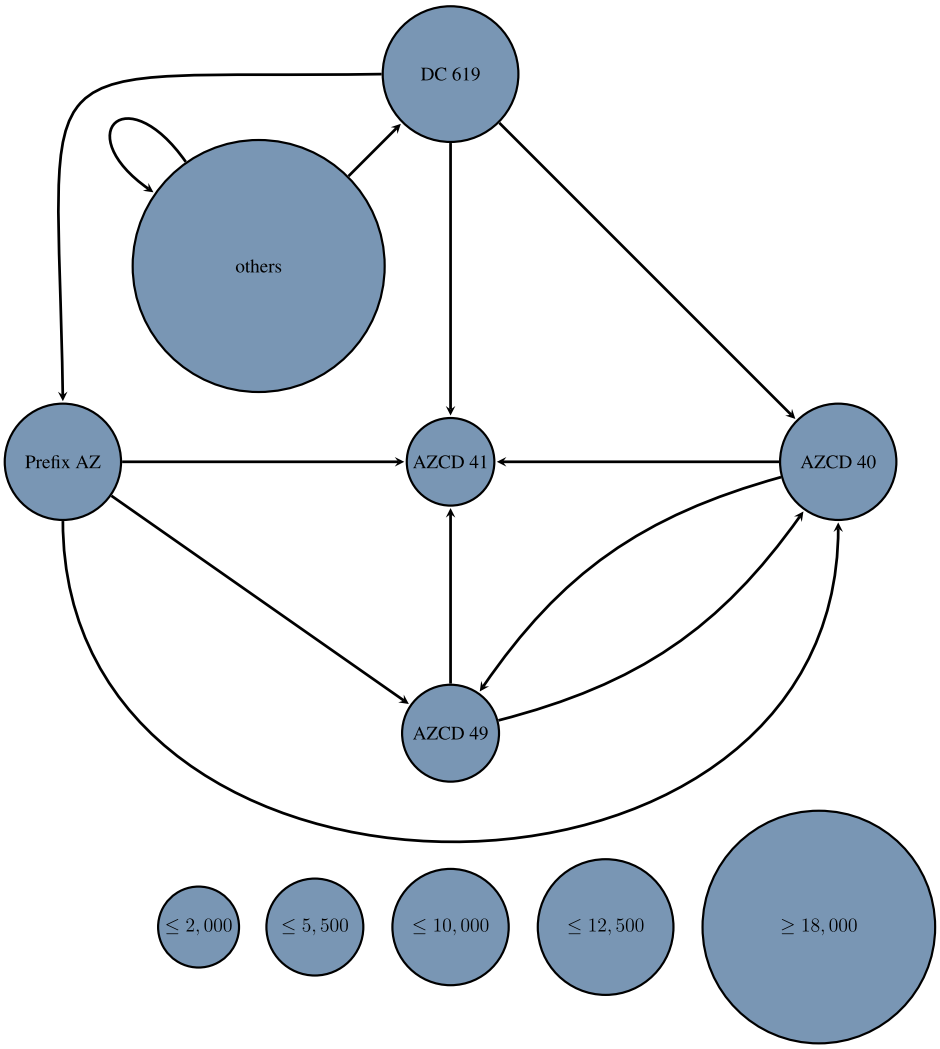


FIGURE 5. Collective disease progression of patients in the test set towards metastatic and non-metastatic PC. Based on the agglomerative event clustering, the dynamics of patient journeys in the latent space, labelled after *k*-means clustering, allows to reconstruct the relation between diagnostic codes. To focus on the relevant transitions, all transitions with a frequency below 15 % have been neglected in the construction of this graph. The graph clearly shows that PC with metastasis (AZCD41) is the definite end of the diagnostic chain. The non-metastatic code (AZCD40) on the other hand can be the final diagnostic code, but can also transition to AZCD41 directly or via the AZCD49 code, which indicates a state of waiting for a result which can either result in a AZCD40 or AZCD41 diagnostic code.

DC619 Latent space leading to the first eventset containing a DC619 diagnostic code.

AZCD40 Latent space leading to the eventset containing the combination DC619 and AZCD40.

AZCD41 Latent space leading to the eventset containing the combination DC619 and AZCD41.

AZCD49 Latent space leading to the eventset containing the combination DC619 and AZCD49.

Prefix AZ Latent space leading to diagnostic codes DC619 with the prefix appendix code AZ that are not AZCD40, AZCD41, or AZCD49.

others Latent space leading to eventsets that are not any of the above.

The accuracy of the model predictions already gives a good indication that information relevant to disease progression is present in the abstract, latent representation. Furthermore, clear cluster separation and structure, for example along the dimensions l_1 and l_2 for prefix AZ vs others, or along the dimensions l_1 and l_3 for DC619 vs others, are evidence that the network has embedded correlations and associations between different eventsets. It is also evident that some dimensions, for example l_4 and l_5, carry no visual interpretable information. Generally, the label ''others'' comes from several clusters, and their projections do not necessarily form continuous shapes. However, the decrease in accuracy for a lower dimensionality l of the latent space is a clear

indication that the network has learned along the additional dimensions, although presumably carrying noise.

C. EVENT PREDICTION

While this way of interpreting the embedded space explores how well related eventsets can be discriminated, it lacks an interpretation how the network reasons over time. In the following, we attempt to answer this question by interpreting the latent space in terms of a spatial-temporal representation. We consider now the sequence of events for one patient only, i.e. $\mathcal{E}_t^{(p)}, t = \{1, \ldots, T_p\}$, and track the change of $\mathbf{l}$ over time. Fig. 4 shows the mentioned latent space with the distribution of all eventsets for all patients, overlaid by one patient's journey as sequence of eventsets. For this patient, each successive eventset is connected by an arrow forming a graph representation of which regions activate. The selected patient's journey can easiest be followed at the top left panel of Fig. 4 with the latent dimensions l_1 and l_2 on the axes. We conclude that the clusters used in Fig. 4 correspond to the relevant milestones in a PC patient journal. The journey starts with several diagnoses unrelated to PC (large, blue distribution), before the first "DC619" diagnosis (green distribution) appears. Then, the transition goes over "AZCD49" to "AZCD41", the final metastasis diagnosis of this journey.

Repeating this exercise for all patients in the test set, and introducing a node for each cluster, we can generate a spatial-temporal graph displaying the collective disease progression of all patients (Fig. 5). When evaluating all patients from the test dataset, we can derive that 20.4 % of all patient journeys end in a metastatic stage. This is significant higher than the 12.6 % of the patients' data descriptive statistic. As noted earlier 12.6 % is low compared to the estimated worldwide 20 to 30 % [14]. Calculating the median over PC diseases with metastasis from 2017 to 2019, based on data from The Danish Prostate Cancer Database (Dansk Prostata Cancer Database, [31]) and the Danish Health Data Authority (Sundhedsdatastyrelsen, [1]), metastatic PC stages can be found in 26.1 % out of all PC diagnoses in Denmark, which is still higher than the patients' average. Danish urologists and nurses believe the underlying reasons to be varying practice in recording SKS codes and changing systems over time. The issue of this discrepancy in coding and reporting is thus well known and is iteratively handled via updated reporting guidelines [32]. We are thus postulating the hypothesis that the proposed network can generalise past recorded data, even if they were erroneously recorded, therefore painting a picture that resembles closer the individual patient journey as it should have been recorded. Extending on this observation, the ability to use the presented method to check the self-consistency of the dataset in an unsupervised manner makes it a candidate for quality assurance in addition to modelling patient journeys, providing input to policy makers to improve health care on an administrative level.

Discussing these graphs, we have to keep in mind that patients in this study were selected by having a PC related diagnosis, and that the labels were chosen accordingly (Section IV-B) to reflect diagnoses around PC. Based on this dataset, the presented method can clearly reproduce a meaningful representation of disease progression from the initial diagnosis to the possible metastasis diagnoses at the end of the patient journey. In this light, the higher accuracy for predicting PC and especially PC with metastasis is an indication that the model could learn PC specific disease progression, while it performed significantly worse to predict seemingly unrelated events, which are presumable randomly represented in the history of selected patients. Therefore, in comparison to DeepCare, Associated Rule Mining, and Random Forests, this model seems to have the ability to prioritise events based on common elements in the patient history, which allows higher accuracy eventset predictions in regard to the common diagnostic complex, i.e. PC. The question if this ability to prioritise events based on common elements in the patient history can be used to find additional indicators for PC diagnosis has to be answered in a future study.

V. CONCLUSION

In this study, we present a modified LSTM node, which is able to extract time scales from non-uniformly sampled inputs. This unsupervised approach to relevant time scales in a data set sets this model apart from other state of the art methods, which use functional or parametric approaches to weight relevant time scales. Applied to a dataset which features patient journals from the Danish National Patient Registry, which were selected based on exhibiting a prostate cancer related diagnostic code, the model based on these modified LSTM nodes is able to achieve high accuracy, sensitivity, and specificity for metastasis related event prediction, which surpasses state of the art methods significantly in absolute numbers. Yet the main characteristic is that the presented model is not predicting all events in the dataset with the same accuracy. While all tested state of the art models perform slightly better for prostate cancer related events, the presented method shows a significant improvement of $\approx$10 % points when predicting PC related events over the whole event set, and another $\approx$10 % points improvement can be seen when restricting the evaluation to PC with metastasis. In this case, the tested state of the art methods do not show a significant change.

The model includes an autoencoder whose dimension has a clear influence on the resulting performance of the model. Furthermore, we demonstrated that the latent variables can be used to create an abstract representation of the patient journey which can be used to reconstruct typical patient journeys and found the reconstruction to be in good agreement with the guidelines. Evidence that the model has the ability to predict the course of PC is a step closer to fill the gap for predictive models in a clinical setting.

In consequence, we conclude that the presented model can focus learning on a subset of events in a problem specific dataset that matches the problem at hand. This sets the presented method apart from the state of the art and allows for many applications.

In future studies, we want to focus on the question if this domain specific knowledge, which is implicitly accumulated in the model, can be used to improve prediction of patient risk factors and to identify relevant diagnostic codes which are not directly related to PC. The identification of such markers would allow to improve the diagnostics of PC by harnessing knowledge about other, seemingly unrelated medical incidents and to suggest additional procedures or treatment options. Based on the observation that graphs are a powerful basis for the understanding of cause-effect relations [33], we will specifically investigate in how far the relations identified with this method can be used for causal modelling and the identification of e.g. relevant preconditions, interventions, and life-style options, by including relevant data into the generated graphs. Extending beyond prostate cancer, future activities will clarify if the method can be applied to model other disease complexes and how it handles data which is selected by more than one disease complex as criteria. We base this on the hypothesis that for any data set drawn from a larger set, by selecting for a specific disease, the model should similarly extract a meaningful latent space representation if the assumption of correlation between the data and the selected disease exists. Generalising further, research into the applicability as quality assurance tool bears the chance of vastly improving health care on an administrative level.

APPENDIX A
AGGLOMERATIVE EVENT CLUSTERING

For each event in the dataset we encode a unique variable E_n such that

$$\mathcal{E} = \{E_1, \ldots, E_n, \ldots, E_N\} \tag{18}$$

is a set of all N possible recorded events (e.g. diagnoses, procedures, treatments, etc.), with $t = \{1, \ldots, T_p\}$ being the times at which events can be observed for a patient p. For each t there then naturally exists a subset of $\mathcal{E}$ (including the empty set), denoted $\mathcal{E}_t^{(p)}$.

Defining the support of an event as countable

$$\text{supp}(E_n) \equiv |\{t | E_n \in \mathcal{E}_t^{(p)} \subseteq \mathcal{E}, t = \{1, \ldots, T_p\}|, \quad \forall p \tag{19}$$

we demand a minimum support, i.e. frequency of the corresponding code, before a given SKS code can be stored in $\mathcal{E}$. For each E_n those support is smaller than the minimum support, we calculate the support of an agglomerative event $E_n^{\uparrow}$ by accumulating the support of all lower hierarchy SKS codes $E_n^{\downarrow}$, that do not fulfil the minimum support, until it exceeds the minimum support required.

In the example of Table 2, for $\text{supp}_{\min} = 15$, the codes DC901 and DC902 would become separate events, whereas DC900 and DC903 would not cross the threshold. Codes which do not cross the threshold will be collected on higher hierarchy levels. In this case, going up one level (DC90) and combing DC900 and DC903 into one agglomerated event would cross the frequency threshold of 15.

For the present analysis, we chose $\text{supp}_{\min} = 50$.

A. TIME BETWEEN EVENTS

We refer to δ_t as the time between two events. It is always positive and unbound, $[0, \infty)$ and without loss of generality can be continuous or discrete:

$$\delta_{t'} = \delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}} = \begin{cases} t' - t & \exists \{t, t' \mid t' > t\} \in \mathbb{R}^{+*} \\ 0 & \text{else.} \end{cases} \tag{20}$$

Carrying on with the example of Table 3 we arrive at the extended subsets $\mathcal{E}_t^{(p)}|\delta_{\mathcal{E}_t^{(p)} \Rightarrow \mathcal{E}_{t'}^{(p)}}$:

$$\mathcal{E}_1^{(1)}|\delta_{\mathcal{E}_0^{(1)} \Rightarrow \mathcal{E}_1^{(1)}} = \{E_1, E_2\}|0 \tag{21a}$$

$$\mathcal{E}_2^{(1)}|\delta_{\mathcal{E}_1^{(1)} \Rightarrow \mathcal{E}_2^{(1)}} = \{E_3, E_4, E_5, E_6\}|45 \tag{21b}$$

$$\mathcal{E}_3^{(1)}|\delta_{\mathcal{E}_2^{(1)} \Rightarrow \mathcal{E}_3^{(1)}} = \{E_3, E_7, E_6, E_8, E_9\}|3 \tag{21c}$$

$$\mathcal{E}_4^{(1)}|\delta_{\mathcal{E}_3^{(1)} \Rightarrow \mathcal{E}_4^{(1)}} = \{E_1\}|41 \tag{21d}$$

$$\mathcal{E}_1^{(2)}|\delta_{\mathcal{E}_0^{(2)} \Rightarrow \mathcal{E}_1^{(2)}} = \{E_{10}, E_{11}, E_2, \}|0 \tag{21e}$$

For simplicity we consider only models that have $N + 1$ inputs, comprising N Boolean conditions for each E_n, i.e. $f : \mathcal{E}_t^{(P)} \rightarrow \{0, 1\}^N$, and the random variable $\delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}}$. We define the input to such a model as the vector

$$\mathbf{x}_t^{(P)} = \left[f\left(\mathcal{E}_t^{(P)}\right), \delta_{\mathcal{E}_t^{(P)} \Rightarrow \mathcal{E}_{t'}^{(P)}} \right]^{\top} \in \mathbb{R}^m. \tag{22}$$

Resulting in the vectors given in equation (1).

APPENDIX B
k-MEANS LATENT SPACE CLUSTERING

We label the learned eventsets by using k-means clustering. We will first summarise k-means clustering, before we explain the label selection below.

To find the clusters, we partition all latent vectors $\mathbf{l}_i$ into k sets $\mathcal{S} \in \{S_1, \ldots, S_k\}$ so as to minimise the total cluster Euclidean distance $\Delta(\mathbf{l}|k)$ between the latent samples $\mathbf{l}_i$ and the cluster centres $\mathbf{C}_j$.

$$\Delta(\mathbf{l}|k) \equiv \frac{1}{T} \sum_{j=1}^{k} \sum_{\mathbf{l}_i \in S_j} ||\mathbf{l}_i - \mathbf{C}_j||^2, \tag{23}$$

and maximising the cost for how well samples of $\mathbf{l}_i$ lie within a cluster:

$$\Psi(\mathbf{l}_i|k) = \frac{1}{k} \sum_{i=1}^{T} \frac{\mu_i^{(a)} - \mu_i^{(b)}}{\max\left(\mu_i^{(a)}, \mu_i^{(b)}\right)}, \tag{24}$$

where

$$\mu_i^{(a)} = \frac{1}{n_j} \sum_{\substack{\mathbf{l}_i, \mathbf{l}_m \in \mathcal{S}_j \\ m \neq i}} ||\mathbf{l}_i - \mathbf{l}_m||^2 \tag{25}$$

TABLE 6. 10-fold cross-validation for selected models, measurements presented as $\mu \pm \sigma^2$.

Model	Acc. PC w. Metastases	Acc. PC	Acc. $\mathcal{E}_t$
Iν-LSTM$_1$	0.86 ± 0.08	0.79 ± 0.07	0.63 ± 0.15
Iν-LSTM$_2$	0.88 ± 0.04	0.80 ± 0.10	0.69 ± 0.13

is the average dissimilarity of $\mathbf{l}_i$ to latent samples within the same cluster j with n_j samples, while

$$\mu_i^{(b)} = \min_{\mathcal{S}_h \neq \mathcal{S}_j} \left(\frac{1}{n_h} \sum_{\substack{\mathbf{l}_s \in \mathcal{S}_j \\ \mathbf{l}_r \in \mathcal{S}_h}} ||\mathbf{l}_s - \mathbf{l}_r||^2 \right) \tag{26}$$

is the lowest dissimilarity of $\mathbf{l}_i$ with respect to any other cluster h with n_h samples.

The decision for an optimal number of clusters is thus a trade-off between the geometric distance to each cluster and the cluster separation, that is

$$\textbf{find } k \textbf{ s.t.} \tag{27a}$$

$$\arg\min_k \Delta(\mathbf{l}|k) \tag{27b}$$

$$\arg\max_k \Psi(\mathbf{l}|k) \tag{27c}$$

We find $k = 15$ clusters. Based on the criteria described in Section IV-B, we select 6 ad hoc labels. The labels are **DC619**, **AZCD40**, **AZCD41**, **AZCD49**, **Prefix AZ**, and **others**, based on the majority of the samples in these clusters following the above criteria. For each new predicted eventset we assign a cluster label by the shortest euclidean distance to the nearest centroid.

APPENDIX C
10-FOLD CROSS-VALIDATION

Table 6 summarises the 10-fold cross-validation for selected models with mean score and standard deviation.

ACKNOWLEDGMENT

(E. S. Nadimi and S. P. Sheikh are co-first authors.)

REFERENCES

[1] Sundhedsdatastyrelsen, "Dødsårsagsregisteret 2019," København, Denmark. [Online]. Available: https://sundhedsdatastyrelsen.dk/da/om-os and https://sundhedsdatastyrelsen.dk/-/media/sds/filer/find-tal-og-analyser/doedsaarsager/doedsaarsagsregisteret_2019.pdf

[2] K. Bekæmpelse, "Kræftens Bekæmpelse Forkning 2021—Forskning sårsrapport," København, Denmark. [Online]. Available: https://www.cancer.dk/ and https://www.cancer.dk/dyn/resources/File/file/7/9617/1647419889/kbf_forskningsrapport_2021.pdf

[3] M. Borre, H.-E. Wittendorff, L. Bentzen, K. Brasso, A. J. Christensen, J. Elversang, O. Graumann, S. Hansen, H. Hvarness, H. Jakobsen, H. Lindberg, M. Moe, H. Møller, M. T. Pank, A. Petersen, P. M. Petersen, M. H. A. Poulsen, and H. D. Zacho, "Diagnostik af prostatacancer (diagnostic of prostate cancer) v. 2.0," in *Clinical Guideline, Sekretariatet for Kliniske Retningslinjer på Kræftområdet (Secretary for Clinical Guidelines in Cancer Fields)*. Aarhus, Denmark: Danske Multidisciplinære Cancer Grupper DMCG.dk, 2020. [Online]. Available: https://www.dmcg.dk/siteassets/kliniske-retningslinjer—skabeloner-og-vejledninger/kliniske-retningslinjer-opdelt-pa-dmcg/daproca/daproca_5.2-diagnostik_v2_291019adm.pdf

[4] O. Graumann, H. D. Zacho, K. Brasso, M. Borre, H.-E. Wittendorff, L. Bentzen, A. J. Christensen, J. Elversang, H. Hvarness, H. Jakobsen, H. Lindberg, M. Moe, H. Møller, M. T. Pank, A. Petersen, P. M. Petersen, and M. H. A. Poulsen, "Billeddiagnostik ved prostatacancer (image diagnostic in prostate cancer) c. 1.0," in *Clinical Guideline, Sekretariatet for Kliniske Retningslinjer på Kræftområdet (Secretary for Clinical Guidelines in Cancer Fields)*. Aarhus, Denmark: Danske Multidisciplinære Cancer Grupper DMCG.dk, 2020. [Online]. Available: https://www.dmcg.dk/siteassets/kliniske-retningslinjer—skabeloner-og-vejledninger/kliniske-retningslinjer-opdelt-pa-dmcg/daproca/daproca_8.0-billeddiagnostik_v1.pdf

[5] T. Kimura, S. Sato, H. Takahashi, and S. Egawa, "Global trends of latent prostate cancer in autopsy studies," *Cancers*, vol. 13, no. 2, p. 359, Jan. 2021.

[6] T. A. Stamey, N. Yang, A. R. Hay, J. E. McNeal, F. S. Freiha, and E. Redwine, "Prostate-specific antigen as a serum marker for adenocarcinoma of the prostate," *New England J. Med.*, vol. 317, no. 15, pp. 909–916, Oct. 1987.

[7] M. Catalona, J. William, P. Smith, S. Deborah, P. Ratliff, L. Timothy, R. Dodds, M. Kathy, M. Coplen, E. Douglas, M. Yuan, M. Petros, A. John, M. Andriole, and L. Gerald, "Measurement of prostate-specific antigen in serum as a screening test for prostate cancer," *New England J. Med.*, vol. 324, pp. 1156–1161, Apr. 1991.

[8] C. Parkes, N. J. Wald, P. Murphy, L. George, H. C. Watt, R. Kirby, P. Knekt, K. J. Helzlsouer, and J. Tuomilehto, "Prospective observational study to assess value of prostate specific antigen as screening test for prostate cancer," *BMJ*, vol. 311, no. 7016, pp. 1340–1343, Nov. 1995.

[9] M. K. Brawer, "Prostate-specific antigen: Current status," *CA, A Cancer J. Clinicians*, vol. 49, no. 5, pp. 264–281, Sep. 1999.

[10] M. S. Lucia, A. K. Darke, P. J. Goodman, F. G. La Rosa, H. L. Parnes, L. G. Ford, C. A. Coltman, and I. M. Thompson, "Pathologic characteristics of cancers detected in the prostate cancer prevention trial: Implications for prostate cancer detection and chemoprevention," *Cancer Prevention Res.*, vol. 1, no. 3, pp. 167–173, Aug. 2008.

[11] M. Thompson, "Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 Ng per milliliter," *New England J. Med.*, vol. 350, pp. 46–2239, May 2004.

[12] M. Schmidt, S. A. J. Schmidt, J. L. Sandegaard, V. Ehrenstein, L. Pedersen, and H. T. Sørensen, "The Danish national patient registry: A review of content, data quality, and research potential," *Clin. Epidemiology*, vol. 7, pp. 449–490, Nov. 2015.

[13] A. Rajkomar, "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, p. 18, 2018.

[14] A. Felici, M. S. Pino, and P. Carlini, "A changing landscape in castration-resistant prostate cancer treatment," *Frontiers Endocrinology*, vol. 3, p. 85, Jul. 2012.

[15] The Danish Health Data Authority. (2015). *Sundhedsdatastyrelsen*. [Online]. Available: https://sundhedsdatastyrelsen.dk/da/english

[16] F. Ahmed, M. S. Hossain, R. U. Islam, and K. Andersson, "An evolutionary belief rule-based clinical decision support system to predict COVID-19 severity under uncertainty," *Appl. Sci.*, vol. 11, no. 13, p. 5810, Jun. 2021.

[17] A. Awaysheh, J. Wilcke, F. Elvinger, L. Rees, W. Fan, and K. L. Zimmerman, "Review of medical decision support and machine-learning methods," *Veterinary Pathol.*, vol. 56, no. 4, pp. 512–525, 2019.

[18] S. Berrouiguet, R. Billot, M. E. Larsen, J. Lopez-Castroman, I. Jaussent, M. Walter, P. Lenca, E. Baca-García, and P. Courtet, "An approach for data mining of electronic health record data for suicide risk management: Database analysis for clinical decision support," *JMIR Mental Health*, vol. 6, no. 5, p. e9766, May 2019.

[19] J. M. Hardin and D. C. Chhieng, *Data Mining and Clinical Decision Support Systems*. New York, NY, USA: Springer, 2007, pp. 44–63.

[20] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, no. 4. [Online]. Available: https://iclr.cc/archive/www/doku.php%3Fid=iclr2016:main.html

[21] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. 1st Mach. Learn. Healthcare Conf.*, vol. 56, F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, Eds. Boston, MA, USA: Northeastern University, Aug. 2016, pp. 301–318.

[22] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "DeepCare: A deep dynamic memory model for predictive medicine," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*, 2017, pp. 30–41.

[23] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware LSTM networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*. New York, NY, USA: Association for Computing Machinery, Aug. 2017, pp. 65–74.
[24] M. Hahsler, B. Grün, and K. Hornik, "Arules—A computational environment for mining association rules and frequent item sets," *J. Stat. Softw.*, vol. 14, no. 15, pp. 1–25, 2005. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v014i15. [Online]. Available: https://www.researchgate.net/publication/246525355_Introduction_to_arules_-_Mining_Association_Rules_and_Frequent_Item_Sets and https://www.jstatsoft.org/article/view/v014i15, doi: 10.18637/jss.v014.i15.
[25] F. M. Bianchi, E. Maiorino, M. C. Kampffmeyer, A. Rizzi, and R. Jenssen, *Recurrent Neural Networks for Short-Term Load Forecasting*. Cham, Switzerland: Springer, 2017.
[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[27] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5, pp. 602–610, 2005.
[28] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org
[29] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Germany: Springer-Verlag, 2006.
[30] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford Univ. Press, 1995.
[31] Regionernes Kliniske Kvalitetsudviklingsprogram (RKKP), "Dansk prostata cancer database," Aarhus, Denmark. [Online]. Available: https://www.rkkp.dk/om-rkkp/ and https://ducg.dk/fileadmin/ingen_mappe_valgt/daproca_aarsrapport_2020_offentlig.pdf
[32] Sundhedsdatastyrelsen, "Indberetning til Landspatientregisteret (LPR3)," København, Denmark. [Online]. Available: https://sundhedsdatastyrelsen.dk/da/om-os and https://sundhedsdatastyrelsen.dk/da/rammer-og-retningslinjer/om-patientregistrering/indberetning-lpr3
[33] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

J. HERP received the B.S. and M.S. degrees in physics from the Institute of Physics, Aarhus University, Denmark, in 2011 and 2013, respectively, and the Ph.D. degree in engineering science in the topic of fault detection and prediction in wind turbines from the Applied Statistical Processing Group, Maersk Mc-Kinney Møller Institute, University of Southern Denmark, in 2017. From 2017 to 2021, he was an Assistant Professor with the Group of Machine Learning and AI (former Applied Statistical Signal Processing Group), Embodied Systems for Robotics and Learning (ESRL). Since 2021, he has been an Associate Professor with the Maersk Mc-Kinney Møller Institute, Applied Artificial Intelligence and Data Science Section (AID), University of Southern Denmark. He has published and coauthored more than 25 journal and conference contributions and supervised more than 20 students in various projects. Besides research in machine learning (ML) applications on medical data, his research interests include fault tracking and prediction in wind turbines (facilitated in ML) and statistical framework. He is also involved in ML applications in medical devices, mainly in wireless capsule endoscopy image analysis, focusing on virtual reconstruction of human organs, and detection of diseases based on video feeds.

JAN-MATTHIAS BRAUN received the Diploma (M.Sc. equivalent) degree in physics on classical and quantum chaos from the Technical University of Dresden, in 2007, and the Ph.D. degree in physics on adaptive orthosis control from the University of Göttingen, in 2015.

He is currently an Assistant Professor with the Applied AI and Data Science Group, University of Southern Denmark. His research interests include artificial intelligence and machine learning, from adaptive and learning systems to applications in clinical contexts and physics simulations.

Dr. Braun is a member of the Bernstein Network Computational Neuroscience e.V. and the German Physical Society.

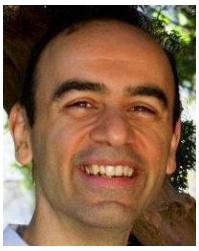
M. L. CANTUARIA received the B.S. and M.Eng. degrees in chemical/environmental engineering from the State University of Campinas, Brazil, and the Ph.D. degree in epidemiology from the University of Southern Denmark. After the Ph.D. degree, she worked as a Research Fellow and a Postdoctoral Researcher with the University of São Paulo, Brazil, and the University of Southern Denmark, respectively. She is currently an Assistant Professor with the University of Southern Denmark and a Guest Researcher with the Danish Cancer Society Research Center. Her research interest includes epidemiological modeling and machine learning techniques applied to different exposures and health-related outcomes.

ASHKAN TASHK (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Shiraz University, Shiraz, Iran, in 2006, and the M.Sc. degree in electrical engineering and the Ph.D. degree in telecommunication engineering from the Shiraz University of Technology, Shiraz, in 2010 and 2015, respectively. Since 2019, he has been a Postdoctoral Researcher with the Unit of Applied AI and Data Science, Maersk Mc-Kinney Møller Institute (MMMI), University of Southern Denmark, Odense, Denmark. He has several publications, including three books in Persian, ten journal articles, and 45 conference proceedings in various fields of signal and image processing, bioinformatics, and biomedical topics. His research interests include biomedical signal and image proceeding, the development of new artificial intelligence (AI) and machine learning (ML) applications, and data science research.

T. B. PEDERSEN received the M.S. degree in medicine from Aarhus University, Aarhus, Denmark, in 2010. He is currently pursuing the Ph.D. degree in medicine. He is a Urology Consultant with Odense University Hospital, Odense, Denmark. His research interest includes prostate cancer, with a special interest in diagnostics and test validation.

M. H. A. POULSEN received the medical degree from the University of Southern Denmark, in 2004, and the Ph.D. degree in health science from the Department of Clinical Research, University of Southern Denmark, in 2014. Since 2019, he is an Associate Professor with the Institute of Clinical Research, University of Southern Denmark, and since 2020, he is a Consultant with the Department of Urology, Odense University Hospital. The cornerstone of his clinical and research work is prostate cancer, with a focus on early detection, staging, and treatment. He is part of a number of local, national, and international studies, exploring both preclinical and clinical topics within prostate cancer.

M. KROGH received the degree in registered nursing from Odense University Hospital and the master's degree in project and innovation management (MPI) from the University of Southern Denmark. She was involved in healthcare innovation and IT for 16 years. She has in-depth insight into clinical perspectives. She works using participatory design methods.



E. S. NADIMI (Senior Member, IEEE) received the M.S. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2004, and the Ph.D. degree in electrical engineering and control system theory from Aalborg University, Aalborg, Denmark, in 2008. In 2011, he joined the School of Engineering and Applied Sciences (SEAS), Harvard University, for a year. He is currently a Professor of artificial intelligence and clinical machine learning with the Maersk Mc-Kinney Møller Institute, Faculty of Engineering, University of Southern Denmark, where he is the Head of the Team of Applied AI and Data Science (AID) and the Technical Research Leader of the Center for Clinical Artificial Intelligence (CAI-X), University of Southern Denmark. His primary research interests include artificial intelligence, clinical machine learning and causal inference, and non-invasive medical robots. His secondary research interest includes green and renewable energy.

S. P. SHEIKH received the medical degree and the Ph.D. degree in medicine from the University of Copenhagen, in 1985 and 1992, respectively, and the master's degree in business communication from Copenhagen Business School, in 2006. Since 2007, he has been a Professor of molecular cardiology with the University of Southern Denmark. Since 2006, he has been the Head of the Department of Biochemistry and Pharmacology, Odense University Hospital, where he has also been the Director of the Danish Center for Regenerative Medicine, since 2014. Throughout his career, he has supervised 16 Ph.D. students and several postdoctoral candidates, has published 151 journal articles and reviews, as well as holds one patent. His research activities have mainly focused on stem cell fate and differentiation (adipose-derived stem cells and iPS cells and differentiation into cardiomyocytes and endothelial cells), stem cell effects in animal models, and implementation of stem cell treatment in clinical medicine. In addition, research activities in molecular mechanisms of signal transduction of seven TM receptors (especially in the cardiovascular systems), transcriptome, and proteome profiling, and the role of DLK1 in stem cell biology and cardiac diseases.