



Time Series Prediction Using Deep Learning Methods in Healthcare

MOHAMMAD AMIN MORID, Santa Clara University

OLIVIA R. LIU SHENG and JOSEPH DUNBAR, University of Utah

2

Traditional machine learning methods face unique challenges when applied to healthcare predictive analytics. The high-dimensional nature of healthcare data necessitates labor-intensive and time-consuming processes when selecting an appropriate set of features for each new task. Furthermore, machine learning methods depend heavily on feature engineering to capture the sequential nature of patient data, oftentimes failing to adequately leverage the temporal patterns of medical events and their dependencies. In contrast, recent deep learning (DL) methods have shown promising performance for various healthcare prediction tasks by specifically addressing the high-dimensional and temporal challenges of medical data. DL techniques excel at learning useful representations of medical concepts and patient clinical data as well as their nonlinear interactions from high-dimensional raw or minimally processed healthcare data.

In this article, we systematically reviewed research works that focused on advancing deep neural networks to leverage patient structured time series data for healthcare prediction tasks. To identify relevant studies, we searched MEDLINE, IEEE, Scopus, and ACM Digital Library for relevant publications through November 4, 2021. Overall, we found that researchers have contributed to deep time series prediction literature in 10 identifiable research streams: DL models, missing value handling, addressing temporal irregularity, patient representation, static data inclusion, attention mechanisms, interpretation, incorporation of medical ontologies, learning strategies, and scalability. This study summarizes research insights from these literature streams, identifies several critical research gaps, and suggests future research opportunities for DL applications using patient time series data.

CCS Concepts: • **Applied computing** → **Life and medical sciences; Health informatics**; • **Computing methodologies** → **Machine learning**;

Additional Key Words and Phrases: Systematic review, patient time series, deep learning methods, healthcare predictive analytics

ACM Reference format:

Mohammad Amin Morid, Olivia R. Liu Sheng, and Joseph Dunbar. 2023. Time Series Prediction Using Deep Learning Methods in Healthcare. *ACM Trans. Manage. Inf. Syst.* 14, 1, Article 2 (January 2023), 29 pages.

<https://doi.org/10.1145/3531326>

Authors' addresses: M. A. Morid, 500 El Camino Real, Leavey School of Business, Santa Clara University, Santa Clara, CA 95053; email: mmorid@scu.edu; O. R. L. Sheng and J. Dunbar, 1655 E Campus Center Dr. David Eccles School of Business, The University of Utah, Salt Lake City, UT 84112; emails: {olivia.sheng, joseph.dunbar}@eccles.utah.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

2158-656X/2023/01-ART2 \$15.00

<https://doi.org/10.1145/3531326>

1 INTRODUCTION

As the digital healthcare ecosystem expands, healthcare data is increasingly being recorded within **electronic health records (EHRs)** and **Administrative Claims (AC)** systems [1, 2]. The widespread adoption of these information systems has become popular with government agencies, hospitals, and insurance companies [3, 4], capturing data from millions of individuals over many years [5, 6]. As a result, physicians and other medical practitioners are increasingly overwhelmed by the massive amounts of recorded patient data, especially given these professionals' relatively limited access to time, tools, and experience wielding this data on a daily basis [7, 8]. This problem has caused **machine learning (ML)** methods to gain attention within the medical domain, since ML methods effectively use an abundance of available data to extract actionable knowledge, thereby both predicting medical outcomes and enhancing medical decision making [3, 9]. Specifically, ML has been utilized in the assessment of early triage, the prediction of physiologic decompensation, the identification of high cost patients, and the characterization of complex, multi-system diseases [10, 11], to name a few. Some of these problems, such as early triage assessment, are not new and date back to at least World War I, but the success of ML methods and the concomitant, growing deployment of EHR and AC information systems have sparked broad research interest [4, 12].

Despite the swift success of traditional ML in the medical domain, developing effective predictive models remains difficult. Due to the high-dimensional nature of healthcare data, typically only a limited set of appropriate features from among thousands of candidates are selected for each new prediction task, necessitating a labor-intensive and time-consuming process. This often requires the involvement of medical experts to extract, preprocess, and clean data from different sources [13, 14]. For example, a recent systematic literature review found that risk prediction models built from EHR data use a median of 27 features from among many thousands of potential variables [15]. Moreover, to handle the irregularity and incompleteness prevalent in patient data, traditional ML models are trained using coarse-grain aggregation measures, such as mean and standard deviation, for input features. These depend heavily on manually crafted features, and they cannot adequately leverage the temporal sequential nature of medical events and their dependencies [16, 17]. Another crucial observation is that patient data evolves over time. The sequential nature of medical events, their associated long-term dependencies, and confounding interactions (e.g., disease progression and intervention) offer useful but highly complex information for predicting future medical events [18, 19]. Aside from limiting the scalability of traditional predictive models, these complicating factors unavoidably result in imprecise predictions, which can often overwhelm practitioners with false alarms [20, 21]. Effective modeling of high-dimensional, temporal medical data can help to improve predictive accuracy and thus increase the adoption of state-of-the-art models in clinical settings [22, 23].

Compared with the traditional ML counterpart, **deep learning (DL)** methods have shown superior performance for various healthcare prediction tasks by addressing the aforementioned high dimensionality and temporality of medical data [12, 16]. These enhanced neural network techniques can learn useful representations of key factors, such as esoteric medical concepts and their interactions, from high-dimensional raw or minimally processed healthcare data [5, 20]. DL models achieve this through repeated sequences of training layers, each employing a large number of simple linear and nonlinear transformations that map inputs to meaningful representations of distinguishable temporal patterns [5, 24]. Released from the reliance on experts to specify which manually crafted features to use, these end-to-end neural net learners have the capability to model data with rich temporal patterns and can encode high-level representations of features as nonlinear combinations of network parameters [25, 26].

Not surprisingly, the recent popularity of DL methods has correspondingly increased the number of their associated publications in the healthcare domain [27]. Several studies have reviewed

such works from different perspectives. Pandey and Janghel [28] and Xiao et al. [29] describe a wide variety of DL models and highlight the challenges of applying them to a healthcare context. Yazhini and Loganathan [30], Srivastava et al. [31] and Shamshirband et al. [32] summarize various applications in which DL models have been successful. Unlike the aforementioned studies, which broadly review DL in various health applications, ranging from genomic analysis to medical imaging, Shickel et al. [27] exclusively focus on research involving EHR data. They categorize deep EHR learning applications into five categories: information extraction, representation learning, outcome prediction, computational phenotyping, and clinical data de-identification, while describing a theme for each category. Finally, Si et al. [33] focus on EHR representation learning and investigate their surveyed studies in terms of publication characteristics, which include input data and preprocessing, patient representation, learning approach, and evaluative outcome attributes.

In this article, we review studies focusing on DL prediction models that leverage patient structured time series data for healthcare prediction tasks from a technical perspective. We do not focus on unstructured patient data, such as images or clinical notes, since DL methods that include natural language processing and unsupervised learning tend to ask research questions that are quite different due to the unstructured nature of the data types. Rather, we summarize the findings of DL researchers for leveraging structured healthcare time series data, of numeric and categorical types, for a target prediction task in terms of the network architecture and learning strategy. Furthermore, we methodically organize how previous researchers have handled the challenging characteristics of healthcare time series data. These characteristics notably include incompleteness, multimodality, irregularity, visit representation, the incorporation of attention mechanisms or medical domain knowledge, outcome interpretation, and scalability. To the best of our knowledge, this is the first review study to investigate these technical characteristics of deep time series prediction in healthcare literature.

2 METHOD

2.1 Overview

The primary goal of this systematic literature review is to extract and organize the findings from research on structured time series prediction in healthcare using DL approaches, and to subsequently identify related, future research opportunities. Because of their fundamental importance and potential impact, we aimed to address the following review questions:

- (1) How are various healthcare data types represented as input for DL methods?
- (2) How do DL methods handle the challenging characteristics of healthcare time series data, including incompleteness, multimodality, and irregularity?
- (3) What DL models are most effective? In what scenarios does one model have advantages over another?
- (4) How can established medical resources help DL methods?
- (5) How can the internal processes of DL outcomes be interpreted to extract credible medical facts?
- (6) To what extent do DL methods developed in limited healthcare settings become scalable to larger healthcare data sources?

To answer these questions, we identify 10 core characteristics including medical task, database, input features, preprocessing, patient representation, DL architecture, output temporality, performance, benchmark, and interpretation for extraction from each study. Section 2.4 elaborates on these 10 core characteristics. In addition, we find that asserted research contributions of the deep time series prediction literature can be classified into the following 10 categories: patient

representation, missing value handling, DL models, addressing temporal irregularity, attention mechanisms, incorporation of medical ontologies, static data inclusion, learning strategies, interpretation, and scalability. Section 3 introduces selected papers exhibiting research contributions in each of these 10 categories and further describes their research approaches, hypotheses, and evaluation results. In Section 4, we discuss strengths and weaknesses of the main approaches and identify research gaps based on the same identified 10 categories.

2.2 Literature Search

We searched for eligible articles in MEDLINE, IEEE, Scopus, and ACM Digital Library published before February 7, 2021. To show a complete picture of the studies published in 2020, we performed the search and selection process again on November 4, 2021, and added all studies published in 2020. Our specific search queries for each of these databases can be found in Table S1 of the online supplement.

2.3 Inclusion and Exclusion Criteria

We followed PRISMA guidelines [34] to include English-language, original research studies published in peer-reviewed journals and conference proceedings. Posters and preprints were not included. We specifically selected papers that employed DL methods to leverage structured patient time series data for healthcare prediction tasks. Reviewed works can be broadly classified under the outcome prediction category of Shickel et al. [27]. We excluded studies based on unstructured data, as well as those lacking key information on the core study characteristics listed in Table 1.

2.4 Data Extraction

We focused the review of each study to center on the 10 identifiable features relating to its problem description, input, methodology, and output. Table 1 provides a brief description and explains the main usage of each of these 10 characteristics.

2.5 Data Analysis

Each selected study in the systematic review has either proposed a technical contribution to advancing the methods for a deep time series prediction pipeline or adopted an extant method for a new healthcare application to make a domain contribution. The focus of this systematic review is summarizing the findings of the former research stream with technical contributions and identifying the associated research gaps. Nevertheless, we also briefly summarize and discuss articles with domain contributions.

Based on the technical contributions noted in the included studies, we classify identifiable contributions into one of 10 categories: patient representation, missing value handling, DL models, addressing temporal irregularity, attention mechanisms, incorporation of medical ontologies, static data inclusion, learning strategies, interpretation, and scalability. For each given category, Section 3 summarizes deep patient time series learning approaches identified in the reviewed studies. Section 4 compares the strengths and weaknesses of these DL techniques and the associated future research opportunities.

3 RESULTS

Our literature search initially resulted in 1,524 studies, with 511 of them being duplicates (i.e., indexed in multiple databases). The remaining 1,014 works underwent a title and abstract screening. Following our exclusion criteria, 621 studies were excluded. Out of these 621 omitted studies, 74 did not use EHR or AC data, 81 did not use multivariate temporal data, 171 did not use DL methods for their prediction tasks, and 295 studies were based on unstructured data, such as

Table 1. Core Study Characteristics

Feature	Description		Usage
Medical task	Describes the medical time series prediction goal		Helps to understand if a certain network quality fits a specific task or if it is generalizable to more than one task
Database	Determines the healthcare data source and scope used for the experiments		Helps to understand whether the experimental dataset is public or not; in addition, since patient data in different countries is recorded with different coding systems, this aids in identifying the adopted coding system
Input	Demographic	Determines if patient demographic data is used as input	Helps to understand the variety of structured patient data used as input in the study
	Vital sign	Determines if patient vital signs are used as input	
	Lab test	Determines if patient lab tests are used as input	
	Procedure codes	Determines if patient procedure codes are used as input	
	Diagnosis codes	Determines if patient diagnosis codes are used as input	
	Medication codes	Determines if patient medication codes are used as input	
	Others	Describes other EHR or AC input features	
Preprocessing	Describes the windowing and missing value imputation methods		Helps to understand how data preprocessing affects the outcome
Patient representation	Shows the final format of the time series data fed into the DL model		Helps to identify whether sequence representation or matrix representation has been used to represent patient time series data
DL architecture	Shows the DL model architecture used for the time series prediction		Helps to compare and contrast the learning architectures, and also to identify architectural contributions
Output temporality	Determines whether the target is static or dynamic		Specifies whether the output is the same for a sequence of events or if it changes over time for each event
Performance	Shows the highest achieved performance based on the primary outcome		Provides researchers in each learning task with state-of-the-art prediction performance
Benchmark	Lists the models used as a baseline for comparison		Identifies traditional ML or DL models that are outperformed by the proposed model
Interpretation	Shows the methods used for DL model interpretation		Aids in understanding how a DL “black-box” model has been interpreted

images, clinical notes, or sensor data. The remaining 393 papers were then selected for a full-text review, and we subsequently removed 316 additional papers because they lacked one or more of the core study characteristics listed in Table 1. Specifically, 64 of the removed papers did not provide distinctive input features (e.g., medical code types), 99 did not have patient representation (e.g., embedding vector creation), 129 did not sufficiently describe their DL network architectures (e.g., RNN network type), and 24 did not specify their output temporality (i.e., static or dynamic) designs. Figure 1 summarizes the article extraction procedure, and Figure 2 shows the distribution of the 77 included studies based on their publication year. A majority of the studies (77%) were published after 2018, signaling a recent surge in interest among researchers for DL models applied to healthcare prediction tasks.

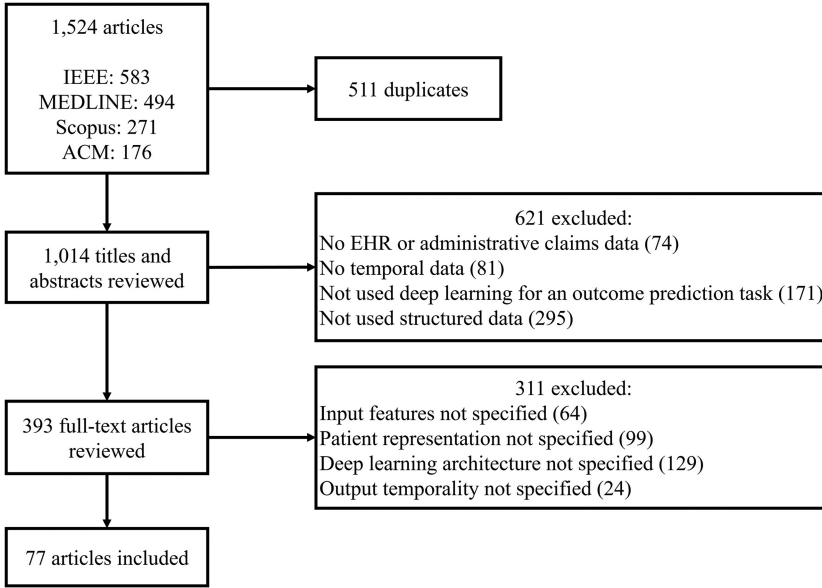


Fig. 1. Inclusion flow of the systematic review.

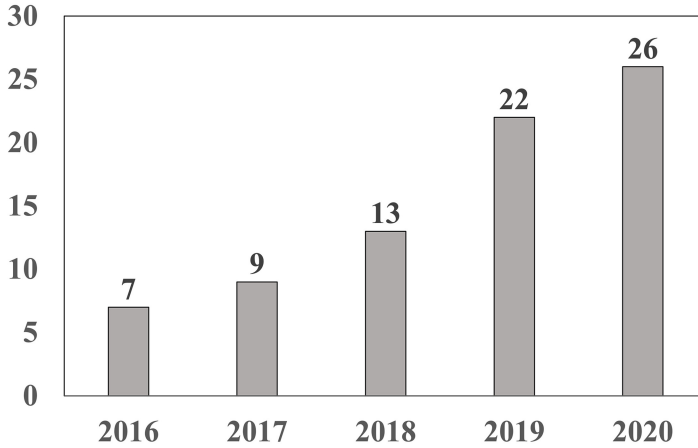


Fig. 2. Number of publications per year.

Table 2 lists the included studies by prediction task. Note that mortality, heart failure, readmission, and patient next-visit diagnosis predictions are the most studied prediction tasks, and a publicly available online dataset, the **Medical Information Mart for Intensive Care (MIMIC)** [35], is the most popular data source for the studies. A complete list of the included studies and their characteristics as delineated in Table 1 is available in the online supplement (Tables S2 and S3).

After reviewing the included studies, we found that the asserted contributions of researchers within the deep time series prediction literature can be distinguished and classified under the following 10 categories: patient representation, missing value handling, DL models, addressing temporal irregularity, attention mechanisms, (6) incorporation of medical ontologies, (7) static

Table 2. List of Reviewed Studies Per Prediction Task

Prediction Task	Reference
Mortality	Che et al. [36], Sun et al. [9], Yu et al. [37], Ge et al. [38], Caicedo-Torres et al. [39], Sha et al. [2], Harutyunyan et al. [12], Rajkomar et al. [13], Zhang et al. [40], Shickel et al. [41], Purushotham et al. [42], Gupta et al. [43], Baker et al. [44], Yu et al. [45]
Heart failure	Cheng et al. [18], Choi et al. [46], Yin et al. [47], Wang et al. [48], Ju et al. [49], Rasmy et al. [50], Choi et al. [22], Maragatham et al. [51], Zhang et al. [52], Choi et al. [53], Ma et al. [54], Choi et al. [55], Solares et al. [56]
Readmission	Zhang et al. [57], Wang et al. [58], Lin et al. [59], Barbieri et al. [60], Min et al. [1], Ashfaq et al. [61], Rajkomar et al. [13], Reddy and Dellen [62], Nguyen et al. [63], Zhang et al. [40], Solares et al. [56]
Next visit diagnosis	Lipton et al. [64], Choi et al. [7], Pham et al. [65], Wang et al. [20], Yang et al. [66], Guo et al. [67], Wang et al. [68], Ma et al. [69], Ma et al. [70], Harutyunyan et al. [12], Pham et al. [71], Lee and Hauskrecht [72], Rajkomar et al. [13], Lee et al. [73], Choi et al. [53], Purushotham et al. [42], Gupta et al. [43], Lipton et al. [74], Bai et al. [75], Liu et al. [76], Zhang et al. [77], Qiao et al. [78]
Cardiovascular disease	Che et al. [36], Park et al. [79], An et al. [80], Duan et al. [81], Park et al. [82]
Length-of-stay	Che et al. [36], Harutyunyan et al. [12], Rajkomar et al. [13], Zhang et al. [40], Purushotham et al. [42]
Septic shock	Zhang et al. [83], Zhang et al. [84], Wickramaratne and Mahmud [85], Svenson et al. [86], Fagerström et al. [87]
Hypertension	Mohammadi et al. [88], Ye et al. [89]
Decompensation	Harutyunyan et al. [12], Purushotham et al. [42], Thorsen-Meyer et al. [90]
Illness severity	Chen et al. [17], Zheng et al. [91], Suo et al. [92]
Acute kidney injury	Tomašev et al. [93]
Joint replacement surgery risk	Qiu et al. [94]
Post-stroke pneumonia	Ge et al. [95]
Renal disease	Razavian et al. [96]
Adverse drug event	Rebane et al. [97]
Cost	Morid et al. [98]
Chronic obstructive pulmonary disorder	Cheng et al. [18]
Kidney transplantation endpoint	Esteban et al. [3]
Surgery recovery	Che et al. [36]
Diabetes	Ju et al. [49]
Asthma	Xiang et al. [99]
Neonatal encephalopathy	Gao et al. [100]

data inclusion, (8) learning strategies, (9) interpretation strategies, and (10) scalability. The rest of Section 3 devotes one section for each of these categories to describe the associated findings by category. Figure 3 gives a general overview of the focal approaches adopted by the included studies.

3.1 Patient Representation

Patient representations employed for deep time series prediction in healthcare can broadly be classified into one of two categories: sequence representation and matrix representation [1]. In the former approach, each patient is represented as a sequence of medical event codes (e.g., diagnosis code, procedure code, or medication code), and the additional input may or may not include the

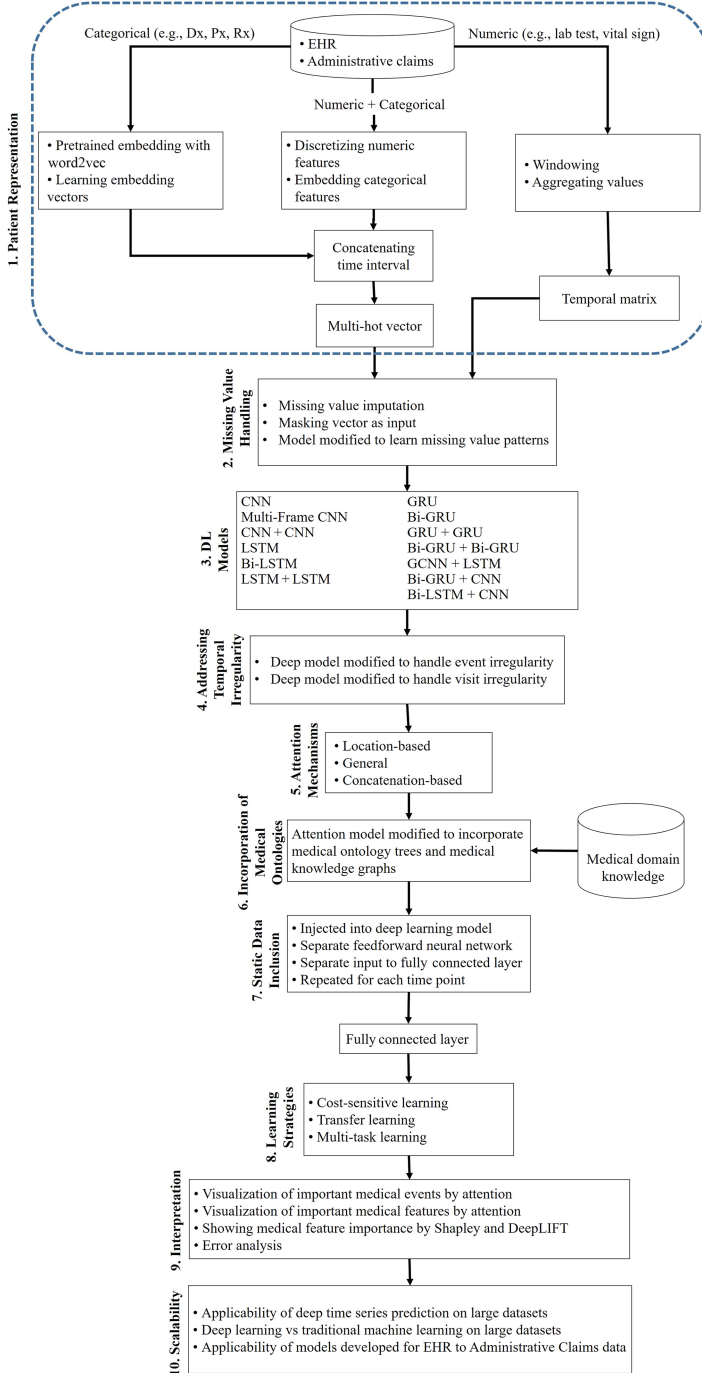


Fig. 3. Summary of deep patient time series prediction designs.

time interval between the events (Section 3.3). Since a complete list of medical codes is generally quite long, various embedding techniques are commonly used to shorten it or combine similar medical codes with comparable values. In the latter approach, each patient is represented as a longitudinal matrix, where columns correspond to different medical events and rows correspond to regular time intervals. As a result, a cell in a patient matrix provides the code of the patient's medical or claims event at a particular time point. Zhang et al. [57] followed a hybrid approach that splits the overall patient sequence of visits into multiple subsequences of equal length, then embeds the medical codes in each subsequence as a multi-hot vector.

As seen in Table S3, sequence representation is a slightly more prevalent approach employed by researchers (57%). Generally, for prediction tasks with numeric inputs, such as lab tests or vital signs, sequence representation is more commonly used, and for those with categorical inputs, like diagnosis codes or procedure codes, matrix representation is the trend. Nevertheless, there are some exceptions. Rajkomar et al. [13] converted patient lab test results from numeric values to categories by assigning a unique token to each lab test name, value, and unit (e.g., "Hemoglobin 12 g/dL") for predicting mortality, length-of-stay, and readmission in **intensive care units (ICUs)**. Ashfaq et al. [61] included the lab test code with a value if the value was designated to be abnormal (determined according to medical domain knowledge), in addition to the typical inclusion of diagnosis and procedure codes. Several research groups [72, 80, 89] converted numerical lab test results into predesigned categories by encoding them as either missing, low, normal, or high when predicting hypertension and the associated onset of high-risk cardiovascular states. Similarly, Barbieri et al. [60] transformed vital signs into OASIS severity scores, then discretized these scores into categories of low, normal, and high. Of note, a singular study observed the superiority of matrix representation over sequence representation for readmission prediction of chronic obstructive pulmonary disease (COPD) patients using a large AC database [1]. This study and other matrix representations [44, 57, 96] found that integrating coarse time granularities such as weekly or monthly rather than finer time granularity measures can improve performance. This study also compared various embedding techniques, and the authors found no significant differences in their results. Finally, Qiao et al. [78] summarized each numerical time series in terms of temporal measures such as their self-correlation structure, data distribution, entropy, and stationarity. They found that these measures can improve the interpretability of the extracted temporal features without degrading prediction performance.

For embedding medical events in the sequence representation, a commonly observed technique was to augment the neural network with an embedding layer that can learn effective medical code representations. This technique has benefited the prediction of hospital readmission [58], patient next-visit diagnosis [66], and the onset of vascular diseases [82]. Another event embedding technique has been to use a pretrained embedding layer via probabilistic methods, especially word2vec [101] and Skip-gram [102], which have shown promising results for predicting an assortment of healthcare outcomes, such as patient next-visit diagnosis [7], heart failure [46, 51], and hospital readmission [57]. Choi et al. [7] demonstrated that pretrained embedding layers can outperform trainable layers by a 2% margin in recall for the next-visit diagnosis prediction problem. Instead of relying on individual medical codes for the next-visit diagnosis problem, several studies grouped medical codes using the first three digits of each diagnosis code, and other works implemented **Clinical Classification Software (CCS)** [103] to obtain groupings of medical codes [68, 73]. However, Maragatham and Devi [51] observed that pretrained embedding layers can outperform medical group coding methods by a 1.5% margin in **area under the curve (AUC)** for heart failure prediction. Finally, Min et al. [1] showed that, independent of the embedding approach, patient matrix representation generally outperformed sequence representation.

Table 3. DL Model Architectures of the Reviewed Studies

Model	Reference
CNN	Caicedo-Torres et al. [39], Nguyen et al. [63]
Multi-frame CNN	Cheng et al. [18], Ju et al. [49]
CNN + CNN	Razavian et al. [96], Wang et al. [58], Morid et al. [98]
LSTM	Pham et al. [65, 71], Zhang et al. [83], Rajkomar et al. [13], Wang et al. [20], Gao et al. [100], Qiu et al. [94], Mohammadi et al. [88], Park et al. [79], Ashfaq et al. [61], Maragatham and Devi [51], Yu et al. [37], Ye et al. [89], Reddy and Dellen [62], Lee and Hauskrecht [72], Zhang et al. [84], Xiang et al. [99], Thorsen-Meyer et al. [90]
Bi-LSTM	Yang et al. [66], Ye et al. [89], Bai et al. [75], Duan et al. [81], Yu et al. [45]
LSTM + LSTM	Lipton et al. [64], Lipton et al. [74], Yin et al. [47], Wang et al. [68], Zhang et al. [40], Fagerström et al. [87]
GRU	Esteban et al. [3], Choi et al. [46], Zheng et al. [91], Choi et al. [53], Choi et al. [22], Che et al. [36], Ma et al. [70], Purushotham et al. [42], Tomašev et al. [93], Rasmy et al. [50], Min et al. [1], Shickel et al. [41], Solares et al. [56], Choi et al. [55], Rebane et al. [97], Ge et al. [95], Suo et al. [92], Liu et al. [76], Zhang et al. [77]
Bi-GRU	Ma et al. [69], Wickramaratne and Mahmud [85], Zhang et al. [57], Barbieri et al. [60], Sun et al. [9], Qiao et al. [78]
GRU + GRU	Choi et al. [7], Wang et al. [48], Gupta et al. [43]
Bi-GRU + Bi-GRU	Sha et al. [2], Park et al. [82], Guo et al. [67] (concurrent)
GCNN + LSTM	Lee et al. [73]
Bi-GRU + CNN	Ma et al. [54]
Bi-LSTM + CNN	Lin et al. [59], Baker et al. [44]
One RNN per feature or feature type	Ge et al. [38], Harutyunyan et al. [12], An et al. [80], Chen et al. [17], Svenson et al. [86]

3.2 Missing Value Handling

Missing value imputation using methods such as zero [3, 40], median [58], forward-backward [64, 66], and domain-knowledge by experts [12, 38] has been the most common approach for handling missing values in patient time series data. The work of Lipton et al. [74] was the first study that used a masking vector to utilize the availability of values as a separate input to predict discharge diagnosis. Other studies adopted the same approach for predicting readmission [59], acute kidney injury [93], ICU mortality [37], and length-of-stay [12]. Last, Che et al. [36] utilized missing patterns as input for predicting mortality, length-of-stay, surgery recovery, and cardiac condition. Their approach outperformed the masking vector technique by approximately 2% margin in AUC.

3.3 DL Models

Table 3 shows the summary of model architectures adopted to learn a deep patient time series prediction model for each included study. **Recurrent neural networks (RNNs)** and their modern variants, including **long short-term memory (LSTM)** and **gated recurrent units (GRU)**, were by far the most frequently used models (84%). A few studies compared the GRU variant against the LSTM architecture. Overall, GRU achieved around 1% advantage in AUC metrics over LSTM for predicting heart failure [47], kidney transplantation endpoint [3], mortality in the ICU [36], and readmission prediction of chronic disease patients [1]. However, for predicting the diagnosis code group of a patient's next admission to the ICU [68], septic shock [83], and hypertension [89], researchers did not find significant differences between these two advanced RNN model types. Additionally, bidirectional variants of GRU and LSTM—so-called Bi-GRU and Bi-LSTM—consistently outperformed their unidirectional counterparts for predicting hospital readmission [57], diagnosis at hospital discharge [66], patient next-visit diagnosis [67, 69, 75], adverse cardiac events [81],

readmission after ICU discharge [59, 60], in-hospital mortality [2, 45], length-of-stay in hospital [12], sepsis [85], and heart failure [54]. Although most studies (63%) employed single-layered RNN, many other works used multi-layered RNN models with GRU [7, 48], LSTM [40, 64, 68, 74], and Bi-GRU [2, 67, 82]. However, despite the numerous studies employing these methods and their variants, multi-layered GRU is the only architecture that has been experimentally compared to its single-layered counterpart for the patient next-visit diagnosis [7] and heart failure prediction tasks [48]. Alternatively, researchers have extensively explored training separate network layers with the architectures of LSTM [12, 38], Bi-LSTM [77], and GRU layers [17] for each feature. These channel-like architectures per feature were reported as being more successful than the simpler RNN models. Finally, for tasks such as predicting in-hospital mortality or hospital discharge diagnosis code, some RNN models were supervised to make assessments at each timestep [12, 64, 74], a procedure known as *target replication*. Their successes provided evidence that it can be more effective to repeatedly make a prediction at multiple time points than merely performing supervised learning for the last time-stamped entry.

Several studies, particularly those from when deep time series prediction within the healthcare domain was in its nascency, utilized **convolutional neural network (CNN)** models for prediction tasks without benchmarking against other types of DL models [18, 39, 58]. These early CNN models have been consistently outperformed by recently developed RNN models for predicting heart failure [49, 52], readmission of patients diagnosed with chronic disease [1], in-hospital mortality [40], diabetes [49], readmission after ICU discharge [40, 59], and joint replacement surgery risk [94]. Nevertheless, Cheng et al. [18] showed that temporal slow fusion can enhance CNN performance, and Ju et al. [49] suggested using 3D-CNN and spatial pyramid pooling for outperforming RNN models for heart failure and diabetes prediction tasks. Alternatively, hybrid deployments of CNN/RNN models have been successful in outperforming pure CNN or RNN models for predicting readmission after ICU discharge [59], patient next-visit diagnosis [73], mortality [44], and heart failure [54].

3.4 Addressing Temporal Irregularity

Two types of temporal irregularities, visit and feature, generally exist in patient data. Visit irregularity indicates that the time interval between visits can vary for the same patient over time. Feature irregularity occurs when different features belonging to the same patient for the same visit are recorded at various time points and frequencies.

The work of Choi et al. [7] was the first study to make use of the time interval between patient visits as a separate input to a DL model for the patient next-visit diagnosis prediction task. This approach also proved to be efficacious in predicting heart failure [46], vascular diseases [82], hospital mortality [13] and hospital readmission [13]. Yin et al. [47] used a sinusoidal transformation of time interval for assessing heart failure. In addition, Pham et al. [65] and Wang et al. [20] modified the internal mechanisms of the LSTM architecture to handle visit irregularity by giving higher weights to recent visits. Their proposed modifications outperformed traditional LSTM architectures by 3% in AUC for the highly frequent benchmarking task of predicting the diagnosis code group of a patient's next visit.

Certain studies hypothesized that handling feature irregularity is more effective than handling visit irregularity [60, 91]. Zheng et al. [91] also modified GRU memory cell learning processes to extract different decay patterns for each input feature for predicting the Alzheimer's severity score in half a year. Their results demonstrated that capturing feature and visit irregularity decreases the mean squared error (MSE) by up to 5% compared to models that capture visit irregularity only. Barbieri et al. [60] and Liu et al. [76] used a similar approach when predicting readmission to ICU and for generating relevant medications from billing codes.

3.5 Attention Mechanisms

Attention mechanisms, originally inspired by the visual attention system found in human physiology, have recently become quite popular among many domains, including deep time series prediction for healthcare [57]. The core underlying idea is that patient visits and their associated medical events should not carry an identical weight during the inference process. Rather, they are contingent on their relative importance for the prediction task at hand.

Most commonly, attention mechanisms initially assign a unique weight for each visit or each medical event, and subsequently optimize these weight parameters during network backpropagation [2, 13, 22, 37]. Also called *location-based attention* [69], this strategy has been incorporated into a variety of RNNs and learning tasks, such as GRU for heart failure [22] and Bi-GRU for mortality [51], as well as LSTM for hospital readmission, diagnosis, length-of-stay [13], and asthma exacerbation [99]. Other commonly used attention mechanisms include a concatenation-based attention device that has been employed for hospital readmission [60] as well as next-visit diagnosis prediction [69], and general attention models that are used primarily for hospital readmission [57] and mortality prediction [41]. Ma et al. [69] benchmarked these three attention mechanisms for predicting medical codes by using a large AC database, and Suo et al. [92] performed a similar benchmarking procedure for illness severity score prediction on EHR data. Both studies reported location-based attention as optimal.

With few exceptions, studies employing an attention mechanism tended not to report any differential prediction performance improvements enabled by attention. Those few studies that did distinguish a particular performance improvement reported that location-based attention mechanisms improved patient next-visit diagnosis by 4% in AUC [65], increased hospital readmission F1-score by 2.4%, and also saw a 13% boost in F1-score for mortality prediction [2]. Zhang et al. [57] was the sole work reporting contributions of visit-level attention and medical code attention separately for hospital readmission, observing that each technique provided an approximate 4% increase in F2-score. An innovative study by Guo et al. [67] argued that all medical codes should not go through the same weight allocation path during attention calculation. Instead, they proposed a crossover attention model with distinct bidirectional GRUs and attention weights for both diagnosis and medication codes. On the whole, we found that most studies utilized attention mechanisms to improve the interpretability of their proposed DL models by highlighting important visits or medical codes, at either a patient or population level. Section 3.9 further elaborates on patient- and population-level properties.

3.6 Incorporation of Medical Ontologies

Another facet of these research streams was the incorporation of medical domain knowledge into DL models to enhance their prediction performance. Standard CCS has the ability to establish a hierarchy of various medical concepts in the form of successive parent-child relationships. Based on this concept, Choi et al. [53] employed CCS to create a medical ontology tree for use in a network embedding layer. These encoded medical ontologies were better able to represent abstract medical concepts when predicting heart failure. Zhang et al. [77] later enhanced this initial ontological strategy by considering more than one parent for each node and also by providing an ordered set of ancestors for each medical concept. Separately, Ma et al. [70] showed that medical ontology trees can be leveraged when calculating attention weights in GRU models, achieving a 3% accuracy increase over Choi et al. [53] for the same prediction task. Following this, Yin et al. [47] demonstrated that causal medical knowledge graphs like KnowLife [104], which contain both “cause” and “is-caused-by” relationships between diseases, outperform both Choi et al. [53] and Ma et al. [70] with an approximate 2% AUC margin for heart failure prediction. Wang et al. [20],

however, enhanced Skip-gram embeddings by adding n-gram tokens from medical concept information, such as disease or drug name, to EHR data. These embedded tokens captured ancestral information for a medical concept similar to ontology trees, and they were applied to the patient next-visit diagnosis task.

3.7 Static Data Inclusion

RNNs are particularly apt at learning from sequential data, although leveraging static data into these types of models has been challenging. The hybrid combination of static along with temporal input is particularly important in a healthcare context, since static features like patient demographic information and prior history can be essential for achieving accurate predictions. Appending patient static data to the input of a final fully connected layer has been the most common approach for integrating these features. It has been applied to hospital readmission [57, 58], length-of-stay [40], and mortality [38, 40] tasks. Alternatively, Esteban et al. [3] fed 342 static features into an entirely independent feedforward neural network before combining the output with temporal data in a typical GRU layer for learning kidney transplant endpoints. Other studies also adopted this approach for predicting mortality [42], phenotyping [42], length-of-stay [42], and the risk of cardiovascular diseases [80]. Moreover, Pham et al. [65] modified the internal processes of LSTM networks to specifically incorporate the effects of unplanned hospital admissions, which involve higher risks than planned admissions. They employed this approach for predicting patient next-visit diagnosis codes in mental health and diabetes cohorts. Finally, Maragatham and Devi [51] converted static data into a temporal format by repeating it as input to every time point. Together, they used static demographic data, vascular risk factors, and a scored assessment of nursing levels for heart failure prediction. We found no study comparing the aforementioned static data inclusion methods against solid benchmarks.

3.8 Learning Strategies

We identified three principal learning strategies that differ from the basic supervised learning scenario: (1) cost-sensitive learning, (2) multi-task learning, and (3) transfer learning. When handling imbalanced datasets, cost-sensitive learning has frequently been implemented by modifying the cross-entropy loss function [58, 61, 93, 100]. In particular, two studies convincingly demonstrated the performance improvement achieved by cost-sensitive learning. Gao et al. [100] found a 3.7% AUC increase for neonatal encephalopathy prediction, and Ashfaq et al. [61] observed a 4% increase for the hospital readmission task. The latter study further calculated cost-saving outcomes by estimating the potential annual cost savings if an intervention is selectively offered to patients at high risk for readmission. Instead, multi-task learning was implemented to jointly predict mortality, length-of-stay, and phenotyping with LSTM [13, 40], Bi-LSTM [12], and GRU [42] architectures. Harutyunyan et al. [12] was a seminal study that reported a significant contribution of multi-task learning over state-of-the-art traditional learning, with a solid 2% increase in AUC. Last, transfer learning, originally used as a benchmark evaluation by Che et al [36], was recently adopted by Gupta et al. [43] to study both task adaptation and domain adaptation utilizing a non-healthcare model, TimeNet. They found that domain adaptation outperforms task adaptation when the data size is small, but otherwise task adaptation is superior. Moreover, they found that for task adaption on medium-sized data, fine-tuning is a better approach than learning from scratch with feature extraction.

3.9 Interpretation

By far, the most common DL interpretation method is to show visualized examples of selected patient records to highlight which visits and medical codes most influence the prediction task

[2, 13, 22, 41, 47, 49, 54, 57, 60, 66, 67, 69, 75, 82, 95, 97]. Specific contributions by feature are extracted from the calculated weight parameters of an attention mechanism (Section 3.6). Visualizations can also be implemented through a global average pooling layer [65, 82] or a one-sided convolution layer within the neural network [57]. Another interpretation approach is to report the top medical codes with the highest attention weights for all patients together [2] or for different patient groups by disease [47, 57, 69, 80]. Specifically, Nguyen et al. [63] extracted the most frequent patterns in medical codes by disease type, and Caicedo-Torres et al. [39] identified important temporal features for mortality prediction using both DeepLIFT [105] and Shapley [106] values. The technique of using Shapley values for interpretation was also employed for continuous mortality prediction within the ICU setting [90]. Finally, Choi et al. [46] performed error analysis on false-positive and false-negative predictions to differentiate the contexts in which their DL models are more or less accurate.

3.10 Scalability

Although most review studies evaluated their proposed models on a single dataset—usually a publicly available resource such as MIMIC and its updates [35]—certain studies focused on assessing the scalability of their models to a wider variety of data. Rasmy et al. [50] evaluated one of the most popular deep time series prediction models with two GRU layers, called *RETAIN*, which was first proposed in a study by Choi et al. [22], on a collection of 10 hospital EHR datasets for heart failure prediction. Overall, they achieved a similar AUC compared to the original study, although a higher dimensionality did further improve prediction performance. Using the same *RETAIN* model, Solares et al. [56] conducted a scalability study on approximately 4 million patients in the UK National Health Service, and they reported an identical observation to that of Ju et al. [49]. Another large dataset was explored by Rajkomar et al. [13], who demonstrated the power of LSTM models on a variety of healthcare prediction tasks for 216,000 hospitalizations involving 114,000 unique patients. Finally, we found a singular study [1] investigating the scalability of deep time series prediction methods for AC data, as opposed to EHR sequences. Min et al. [1] observed that DL models are effective for readmission prediction with patient EHR data, but they tend not to be superior to traditional ML models using AC data.

Studies on the MIMIC database have consistently used the same 17 features in the dataset, which have a low missing rate [107]. To address dimensional scalability, Purushotham et al. [42] attempted using as many as 136 features for mortality, length-of-stay, and phenotype prediction with a standard GRU architecture. Compared to an ensemble model constructed from several traditional ML models, they found that for lower-dimensional data, traditional ML performance is comparable to DL performance, whereas for high-dimensional data, DL's advantage is more pronounced. On a similar note, Min et al. [1] evaluated a GRU architecture against traditional supervised learning methods on around 103 million medical claims and 17 million pharmacy claims for 111,000 patients. Again, they found that strong traditional supervised ML techniques have a comparable performance to that of their DL competitors.

4 DISCUSSION

4.1 Patient Representation

Out of the commonly used sequential and matrix patient representations, prediction tasks with predominantly numeric inputs, such as lab tests and vital signs, often rely on sequence representations, whereas those studies utilizing mainly categorical inputs, like diagnosis codes or procedure codes, commonly incorporate a matrix representation. Other than a lone study [1] that documented the superiority of the matrix approach on AC data, we found no consistent comparison between

these two approaches in our systematic review. In addition, while a coarse-grain abstraction has not been suggested for each of these approaches, changing the granularity level to find the optimal level would be highly suggested to further ascertain their respective efficacy. The rationale is that the sparsity of temporal patient data is typically high, and considering every individual visit for an embedded patient representation may not be the optimal approach when factoring in the corresponding increase in computational complexity.

To combine numeric and categorical input features, researchers have generally employed three distinct methods. One method involves converting patient numeric quantities to categorical ones by assigning a unique token to each measure. Thus, each specific lab test code, value, and unit will have its own identifying marker. Using a second method, researchers encode numeric measures with clinically meaningful names, such as *missing*, *low*, *high*, *normal*, and *abnormal*. A third alternative requires the conversion of numeric measures to severity scores, to discretize them into *low*, *normal*, and *high* categories. The second approach was quite common in our selected studies, likely due to its implementation simplicity and effectiveness for a wide variety of clinical healthcare applications. We therefore report it to be the most dominant strategy for combining numeric and categorical inputs for deep time series prediction tasks.

When embedding medical events into a sequence representation, we again found three prevalent techniques. Using the first technique, researchers commonly added a separate embedding layer, prefacing the bulwark of the recurrent network, to optimize medical code representation. Alternatively, pretrained embedding layers with established methods such as word2vec were adopted in lieu of learning embeddings from scratch. Last, researchers often utilized medical code groups instead of the atomized medical code. Among the three practices, pretrained embedding layers have consistently outperformed naive embedding layers and medical code groupings for EHR data, whereas no significant difference in model performance has been observed for AC data. In addition, researchers have shown that temporal matrix representation is the most effective approach for AC data. The rationale is that the temporal granularity of EHR data is usually at the level of an *hour* or even *minute*, whereas the granularity of AC data is at the *day* level. As a result, the order of medical codes within a day is ordinarily lost for the embedding algorithms such as word2vec. Combining our findings, a sequence representation with a pretrained embedding layer is highly recommended for learning tasks on EHR data, whereas a matrix representation seems to be more effective for AC data.

Several important gaps exist regarding the specific representation of longitudinal patient data. Sequence and matrix methodologies should be compared in a sufficient variety of healthcare settings for EHR data. If extensive comparisons could confirm the relative performance of matrix representation, then it would further enhance its desirability, as it is easier to implement and has a faster runtime than sequences of EHR codes. Moreover, to improve patient similarity measures, researchers should analyze the effect of different representation approaches under various DL model architectures. Last, we found that few reviewed studies included both numerical and categorical measures as feature input. A superior approach that synergistically combines their relative strengths has not yet been sufficiently studied and thus requires the attention of future research. Further investigation of novel DL architectures with a variety of possible input measures is therefore recommended.

4.2 Missing Value Handling

The most common missing value handling approach found in the deep time series prediction literature was imputation by predetermined measures, such as zero or the median—also a common practice in non-healthcare domains [108]. However, missing values in healthcare data typically do not occur at random, as they can reflect specific decisions by caregivers [74]. These missing

values thus represent *informative missingness*, providing rich information about target labels [36]. To capture this correspondence, researchers have implemented two primary approaches. The first approach involves creating a binary (masking) vector for each temporal variable, indicating the availability of data at each time point. This approach has been evaluated in various applications, and it seems to be an effective way of handling missing values. Second, missing patterns can be learned by directly training the imputation value as a function of either the latest observation or the empirical mean prior to variable observations. This latter approach is more effective when there is a high missing rate and a high correlation between missing values and the target variable. For instance, Che et al. [36] found that learning missing values was more effective when the average Pearson correlation between lab tests with a high rate of missingness and the dependent variable, mortality, was above 0.5. Despite this, since masking vectors have been evaluated on a wider variety of healthcare applications, and with different degrees of missingness, they should remain as the suggested missing value handling strategy for deep time series prediction.

Interestingly, there was no study assessing the differential impact of missingness for individual features on a given learning task. The identification of features whose exclusion or missingness most harms the prediction process informs practitioners about how to focus their data collection and imputation strategies. Furthermore, although *informative missingness* applies to many temporal features, *missing-at-random* can still be the case for other feature types. As a direction for future study, we recommend a comprehensive analysis of potential sources of missingness, for each feature and its type, along with assistance from domain experts. This would better inform a missing value handling approach within the healthcare domain and, as a consequence, enhance prediction performance accordingly.

4.3 DL Models

Rooted in their ability to efficiently represent sequential data and extract its temporal patterns [64], RNN-based DL models and their variants were found to be the most prevalent architecture for deep time series prediction on healthcare data. Patient data naturally has a sequential nature, where hospital visits or medical events occur chronologically. Lab test orders or vital sign records, for example, take place at specific timestamps during a hospital visit. However, vanilla RNN architectures are not sophisticated enough to sufficiently capture temporal dependencies when EHR sequences are relatively long, due to the vanishing gradient problem [109]. To address this issue, LSTM and GRU recurrent networks, with their memory cells and elaborate gating mechanisms, have been habitually employed by researchers, with improved outcomes on a variety of healthcare prediction tasks. Although some studies display a slight superiority of GRU architectures versus LSTM networks (around 1% increase in AUC), other studies did not find significant differences between them. Overall, LSTM and GRU have similar memory-retention mechanisms, although GRU implementations are less complex and have faster runtimes [89]. Due to this similarity, most works have used one without benchmarking it against the other. In addition, for very long EHR sequences, such as ICU admissions with a high rate of recorded medical events, bidirectional GRU and LSTM networks consistently outperformed their unidirectional counterparts. This is likely because bidirectional recurrent networks simultaneously learn from both past and future values in a temporal sequence, so they retain additional trend information [69]. This is particularly important in the healthcare context, since patient health status patterns change rapidly or gradually over time [12]. For example, an ICU patient with a rapidly fluctuating health status over the past week may eventually die, even if the patient is currently in a good condition. Another patient, initially admitted to the ICU within the past week in a very bad condition, may gradually improve and survive. Therefore, bidirectional recurrent networks are the most state-of-the-art DL models for time series prediction in healthcare. GRU, which has lower complexity and comparable performance to

LSTM, is the preferred model variant, although additional comparative studies are recommended by this review to affirm this conclusion.

Most RNN studies employed single-layer architectures; however, some studies chose an increased complexity with multi-layered GRU [7, 48], LSTM [40, 64, 68, 74], and Bi-GRU [2, 67, 82] networks. Other than two earlier works [7, 48], multi-layered architectures were not consistently tested against their single-layered counterparts. Consequently, it is difficult to decipher if adding additional RNN layers, whether they are bidirectional or not, improves learning performance. However, *channel-wise learning*, a technique that trains a separate RNN layer per feature or feature type, successfully enhanced traditional RNN models that contain network layers that learn all feature parameters simultaneously. There are two underlying ideas behind this development. First, it helps identify unique patterns within each individual time series (e.g., body organ system status) [17] prior to integration with patterns found in multivariate data. Second, channel-wise learning facilitates the identification of patterns related to *informed missingness*, by discovering which of the masked variables correlates strongly with other variables, target or otherwise [12]. Nevertheless, channel-wise learning needs further benchmarking against vanilla RNN models to learn the conditions under which it is most beneficial. Additionally, certain works enhanced upon the supervised learning process of RNN models. For prediction tasks with a static target, such as in-hospital mortality, RNN models were supervised at multiple timesteps instead of merely the final time point. This so-called *target replication* has been shown to be quite efficient during backpropagation [64]. Specifically, instead of passing patient target information across many timesteps, the prediction targets are replicated at each time point within the sequence, thus providing additional local error signals that can be individually optimized. Moreover, target replication can improve model predictions even when the temporal sequence is perturbed by small, yet significant, truncations.

As noted in Section 3.3, convolutional network models were more commonly used in the early stages of deep time series prediction for healthcare. Eventually, they were shown to be consistently outperformed by recurrent models. However, recent architectural trends have been using convolutional layers as a complement to GRU and LSTM [44, 54, 59, 73]. The underlying idea is that RNN layers capture the global structure of the data via modeling interactions between events, whereas the CNN layers, using their temporal convolution operators [54], capture local structures of the data occurring at various abstraction levels. Therefore, our systematic review suggests using CNNs to enhance RNN prediction performance instead of employing either in a stand-alone setting. Another recent trend in the literature is the splitting of entire temporal sequences into subsequences for various time periods—before applying convolutions of different filter size—to capture temporal patterns within each time period [49]. For optimal local pattern (motif) detection, slow-fusion CNN that considers both individual patterns of the time periods as well as their interactions has been shown to be the most effective convolutional approach [18].

Several important research gaps were identified in the models used for deep time series prediction in healthcare. First, there is no systematic comparison among state-of-the-art models in different healthcare settings, such as rare versus common diseases, chronic versus nonchronic maladies, and inpatient versus outpatient visits. These different healthcare settings have identifiable heterogeneous temporal data characteristics. For instance, outpatient EHR data contains large numbers of visits with few medical events recorded during each visit, whereas inpatient visits contain relatively few visit records but documented long sequences of events for each visit. Therefore, the effectiveness of a given DL architecture will vary over these different clinical settings. Second, it is not clear whether adding multiple layers of RNN or CNN within a given architecture can further improve model performance. The maximum number of layers observed within the reviewed and selected studies was two. Given enough training samples, the addition of more layers may further

improve performance by allowing for the learning of increasingly sophisticated temporal patterns. Third, most of the reviewed studies (92%) targeted a prediction task on EHR data, whereas the generalizability of the models to AC data needs more investigation. For example, although many studies reported promising outcomes for EHR-based hospital readmission predictions using GRU models, Min et al. [1] found that similar DL architectures are ineffective for claims data. Finding novel models that can extract temporal patterns from EHR data—which are simultaneously applicable to claims data—can be an interesting future direction for transfer learning projects. Fourth, although channel-wise learning seems to be a promising new trend, it needs researchers to further investigate the precise temporal patterns detected by this approach. DL methods focused on interpretability would be ideal for such an application. Fifth, many studies compared their DL methods against expert domain knowledge, but a hybrid approach that leverages expert domain knowledge within the embeddings should help improve representation performance. Last, the prediction of medications, either by code or group, has been a well-targeted task. However, a more aggressive approach, such as predicting medications along with their appropriate dosage and frequency, would be a more realistic and useful target for clinical decision making in practice.

4.4 Addressing Temporal Irregularity

The most common approach for handling visit irregularity is to treat the time interval between adjacent events as an independent variable and concatenating it to the input embedding vectors. Although this technique is easy to implement, it does not consider contextual differences between recent and earlier visits. Addressing this limitation, researchers modified the internal memory cells of RNN networks by giving higher weights to recent visits [20, 65]. However, a systematic comparison between the two approaches has not been explored. Therefore the time interval approach, which has been shown to be effective in various applications, remains the most efficient, tested strategy to handle visit irregularity. It is noteworthy that tokenizing time intervals is also considered the most effective method of capturing duration in the context of natural language processing [110, 111], a field of study that inspires many of the deep time series prediction methods in healthcare.

Although most works addressing irregularity focus on visit irregularity, a few studies concentrated on feature irregularity [60, 91]. A fundamental concept underpinning the difference between the two is that fine-grained temporal information is more complex and yet important to learn at the feature level than visit level. Specifically, different features expose different temporal patterns, such as when certain features decay faster than others. Paralleling the work on visit irregularity and time intervals, these studies [60, 91] modified the internal processes of RNN networks to learn unique decay patterns for each individual input feature. Again, this research direction is relatively new and boasts few published works, so it is difficult to make a general suggestion for unilaterally handling feature irregularity in deep time series learning tasks.

Overall, we can stipulate that adjusting the memory mechanisms of recurrent networks when addressing either visit or feature irregularity needs additional benchmarking experiments to make the arguments robust. Currently, it has been evaluated in a single hospital setting for each case. Therefore, optimal synergies among patient types (inpatient vs. outpatient), sequence lengths (long vs. short), and irregularity approaches (time interval vs. modifying RNN memory cells) are not entirely conclusive, but time interval approaches have been most commonly published.

4.5 Attention Mechanisms

Attention mechanisms have been employed by researchers with the premise that neither patient visits nor medical codes should contribute equally when performing a target prediction task. As

such, learning attention weights for visits and codes have been the subject of many deep time series prediction studies. The three most commonly used attention mechanisms are (1) location-based, (2) general attention, and (3) concatenation-based frameworks. The methods differ primarily on how the learned weight parameters are connected to the model's hidden states [69]. Location-based attention schemes calculate weights from the most current hidden state. Alternatively, general attention calculations are based on a linear combination connecting the current hidden states to the previous hidden states, with weight parameters being the linear coefficients. Most complex is the concatenation-based attention framework, which trains a multi-layer perceptron to learn the relationship between parameter weights and hidden states. Location-based attention systems have been the most commonly used attention mechanisms for deep time series prediction in healthcare.

We found several research gaps regarding attention. Most studies relied on attention mechanisms to improve the interpretability of their proposed DL model by highlighting important visits or medical codes, without evaluating the differential effect of attention on prediction performance. This is an important issue, as incorporating attention into a model may improve interpretability, but it does not have an established effect on performance in the DL for the healthcare time series domain. Furthermore, with only a single exception [57], we did not find studies reporting the separate contributions of visit-level attention and medical code level attention. Last, and again with only a single exception [69], no study compared the performance or interpretability of different attention mechanisms. All of these research gaps should be investigated in a comprehensive manner in future studies, particularly for EHR data, as most prior attention studies have focused on the clinical histories of individual patients.

4.6 Incorporation of Medical Ontologies

When incorporating medical domain knowledge into deep time series prediction models, researchers have mainly utilized medical ontology trees and knowledge graphs within embedding layers of recurrent networks. Some of the success of these approaches is due to the enhancement they provide when addressing rare diseases. Being less frequent in the data, a proper representation and pattern extraction for rare diseases is challenging for simple RNN models. Medical domain knowledge graphs provide rare disease information to the model through ancestral node embeddings that contain hierarchical information of the disease. However, this advantage is not as exceptional when sufficient data is available for all patients over a long record history [53, 70]. Continuing research is needed to expand the innovative architectures that incorporate medical ontologies for a broad variety of prediction tasks and case studies.

4.7 Static Data Inclusion

There are four published approaches for integrating static patient data with their temporal data. By far, the most common approach is to feed a vector of static features as additional input to the final fully connected layer of a DL network. Another strategy trains a separate feedforward neural network on the static features, then adds the encoded output of this separate network to the final dense layer in the principal neural network for target prediction. Researchers have also injected static data vectors as input to each time point of the recurrent network, effectively treating the patient demographic and historical data as quasi-dynamic. Last, similar to those strategies that handle visit and feature irregularities, researchers have modified the internal memory processes of recurrent networks to incorporate specific static features as input.

The most important research gap regarding static data inclusion is that we have found no study evaluating the differential effects of static data on prediction performance. Moreover, comparing these four approaches in a meaningful benchmarking setting, with the expressed goal of finding the

most optimal technique, could be an interesting future research direction. Finally, since DL models may not learn the same representation for every subpopulation of patients (e.g., male vs. female, chronic vs. nonchronic, or young vs. old), significant research gaps exist in the post analysis of static feature performance as input. Such analyses could inform decision makers of crucial insights into model fairness and would also stimulate future research on predictive models that better balances fairness with accuracy.

4.8 Learning Strategies

Recent literature has investigated three new DL strategies: (1) cost-sensitive learning, (2) multi-task learning, and (3) transfer learning. Although many reviewed studies used an imbalanced dataset for their experiments, a select few embedded cost information as a learning strategy that incorporated additional cost-sensitive loss. Specifically, each of these studies changed the loss function of the DL model to increasingly penalize for misclassification of the minority class. In the healthcare domain, imbalanced datasets are very common, and patients with diseases are less common than healthy patients. Moreover, most of the prediction tasks on the minority class lead to critical care decisions, such as identifying those patients who are likely to die in the next 48 hours or those who will become diabetic in the relatively near future. Devising cost-sensitive learning components into DL networks thus needs further attention and is a wide open research gap for future inquiry. As an example, exploring cost-sensitive methods in tandem with the traditional ML techniques of oversampling or undersampling could lead to significant performance increases in model prediction rates for the minority class. In addition, calculating the precise cost savings when correctly identifying the minority class of patients, similar to Ashfaq et al. [61], can further underline the importance of the cost-sensitive learning strategy.

Researchers have reported the benefit of multi-task learning by documenting its performance in a significant variety of healthcare outcome prediction tasks. However, the cited works do not distinguish the model components that exemplify why learning a single, multi-task deep model is preferable to simultaneously training multiple DL models for respective individualized prediction tasks. More specifically, we ask which layers, components, or learned temporal patterns in a DL network should be shared among different tasks, and in which healthcare applications might this strategy be most efficient? These research questions are straightforward and could be fruitfully studied in the near future with explainable DL models.

Among the three noted, transfer learning was the least studied learning strategy found within our systematic review of the literature with just a single citation [43], displaying the effectiveness of the method for both task and domain adaptation. It is commonly assumed that, with sufficient data, trained DL models can be effective for a wider variety of prediction tasks and domains. However, in many healthcare settings, such as those with rural patients, sufficient data is difficult to collect [112]. Transfer learning methods have the potential to make a huge impact on deep time series prediction in healthcare by making pretrained models applicable to essentially any healthcare setting. Still, further research is recommended to ascertain which pathological prediction tasks are most transferable, which network architectures are most flexible, and which model parameters require the least tuning when transferring to different domains.

4.9 Interpretation

One of the most common critiques of DL models is the difficulty of their interpretation, and researchers have attempted to alleviate this issue with five different approaches. The first approach uses feature importance measures such as Shapley and DeepLIFT. A Shapley value of a feature is the average of its contribution across all possible coalitions with other features, whereas DeepLIFT compares the activation of each neuron in the deep model inputs to its default reference

activation value and assigns contribution scores according to the difference [113]. Although both of these measures cannot illuminate the internal procedure of DL models, they can identify which features have been most frequently used to make final predictions. A second approach visualizes what input data the model focused on for each individual patient [13] through the implementation of interpretable attention mechanisms. In particular, some studies investigated which medical visits and features contributed most to prediction performance with a network attention layer. As a clinical decision support tool, this raises clinician awareness of which medical visits deserve careful human examination. In addition to individual patient visualization, a third interpretation tactic aggregated model attention weights to calculate the most important medical features for specific diseases or patient groups. Additionally, error analysis of final prediction results allowed for consideration of the medical conditions or patient groups for which a DL model might be more accurate. This fourth interpretation approach is also popular in non-healthcare domains [114]. Finally, considering each set of medical events as a basket of items and each target disease as the label, researchers extracted frequent patterns of medical events most predictive of the target disease.

Overall, this review found explainable attention to be the most commonly used strategy for interpreting deep time series prediction models evaluated on healthcare applications. Indeed, individual patient exploration can help make DL models more trustworthy to clinicians and facilitate subsequent clinical actions. Nevertheless, because implementing feature importance measures is much less complex, this study recommends consistently reporting them on most healthcare deep times series prediction studies, providing useful clinical implication with little added effort. Although individual-level interpretation is important, extracting general patterns and medical events associated with target healthcare outcomes is also beneficial for clinical decision makers, thereby contributing to clinical practice guidelines. We found just one study implementing a population-level interpretation [63], extracting frequent CNN motifs of medical codes associated with different diseases. Otherwise, researchers broadly have reported the top medical codes with the highest attention weights for all patients [2] or different patient groups, to provide a population-level interpretation. This current limitation can be an essential direction for future research involving network interpretability.

4.10 Scalability

We identified two main findings regarding the scalability of deep time series prediction methods in healthcare. First, although DL models are usually evaluated on a single dataset with a limited number of features, some studies confirmed their scalability to large hospital EHR datasets with high dimensionality. The fundamental observation is that higher dimensionality and larger amounts of data can further enhance model performance by raising their representational learning power [42]. Such studies have typically used single-layered GRU or LSTM architectures, but analyzing more advanced neural network schemas, such as those proposed in recent studies (Section 3.1), is a venue for future research. In addition, one scalability study observed that models which are primarily purposed for EHR data may not be as effective with AC data [1]. This is mainly because potent predictive features available in EHR data, such as lab test results, tend to be missing in AC datasets. Therefore, scalability studies on AC data merits further inquiry. Second, DL models are typically compared against traditional supervised ML methods on a singular method only (Table S3). However, two studies [1, 42] compared DL methods against ensembled traditional supervised learning models, both on EHR and AC data, and found that their performances are comparable. This shows an important research gap for proper comparison between DL and traditional supervised learning models to identify data settings, such as feature types, dimensionality, and missingness, in which DL models either perform comparably or excel against their traditional ML counterparts.

5 CONCLUSION

In this work, we systematically reviewed studies focused on deep time series prediction to leverage structured patient time series data for healthcare prediction tasks from a technical perspective. The following is a summary of our main findings and suggestions:

- *Patient representation*: There are two common approaches—sequence representation and matrix representation. For prediction tasks in which inputs are numeric, such as lab tests or vital signs, sequence representations have typically been used. For those with categorical inputs, such as diagnosis codes or procedure codes, matrix representation is the premiere choice. To combine numeric and categorical inputs, researchers have employed three distinct methods: (1) assigning a unique token to each combination of measure name, value, and unit; (2) encoding the numeric measures categorically as missing, low, normal, or high; and (3) converting the numeric measures to severity scores to further discretize them as low, normal, or high. Moreover, embedding medical events in a sequence representation involved an additional three prevailing techniques: (1) adding a separate embedding layer to learn an optimal medical code representation from scratch, (2) adopting a pretrained embedding layer such as with word2vec, or (3) using a medical code grouping strategy, sometimes involving CCS. Comparing these diverse approaches and techniques in a solid benchmarking setting needs further investigation.
- *Missing value handling*: Missing values in healthcare data are generally not missing at random but often reflect decisions by caregivers. Capturing missing values as a separate input masking vector or learning the missing patterns with a neural network have been the most effective methods to date. Identifying impactful missing features will help healthcare providers implement optimal data collection strategies and better inform clinical decision making.
- *DL models*: RNN architectures, particularly their single-layered GRU and LSTM versions, were identified as the most prevalent networks in extant literature. These models excel at large sequences of input data representing longitudinal patient history. RNN models extract global temporal patterns; however, CNNs are proficient at detecting local patterns and motifs. Combining RNN and CNN in a hybrid structure for capturing both types of patterns has become a trend in recent studies. More investigation is required to understand optimal network architecture for various hospital settings and learning tasks.
- *Addressing temporal irregularity*: For handling visit irregularity, the time interval between visits is given as an additional independent input, or alternatively, the internal memory processes of recurrent networks are slightly modified to assign differing weights to earlier versus more recent visits. When addressing feature irregularities, the memory and gating activities of RNN networks are similarly modified to learn individualized decay patterns for each feature or feature type. Overall, temporal irregularity handling methods need more robust benchmarking experiments in an assortment of hospital settings, including variations in patient type (inpatient vs. outpatient) and visit length (long-sequence vs. short-sequence).
- *Attention mechanisms*: Location-based attention is by far the most commonly used means of differentiating importance in portions of the input data and network nodes. Most studies used attention mechanisms to improve the interpretability of their proposed DL models by highlighting important visits or medical codes but without evaluating the differential effect of attention mechanisms on prediction performance. Furthermore, we found that further inquiry is warranted to separately evaluate contributions of visit-level and medical code attention.
- *Incorporation of medical ontologies*: Researchers have incorporated medical ontology trees and knowledge graphs in the embedding layers of recurrent networks to compensate for

lack of sufficient data when regarding rare diseases for prediction tasks. Using these medical domain knowledge resources, the information for such rare diseases is captured through the ancestral nodes and pathways in the tree or graph for input into network embeddings.

- *Static data inclusion:* We found four basic approaches followed by researchers to merge demographic and patient history data with the dynamic longitudinal input of EHR or AC data: (1) feeding static features to the final fully connected layer of the neural network, (2) training a separate feedforward network for the subsequent inclusion of encoded output into the main network, (3) the repetition of static feature input at each time point in a quasi-static manner, and (4) modifying the internal processes of recurrent networks. We found no study evaluating the effects of static data on prediction performance, especially post analysis of performance results for static features.
- *Learning strategies:* Three learning strategies have been investigated by the authors included in this review: (1) cost-sensitive, (2) multi-task, and (3) transfer learning. Devising cost-sensitive learning components into DL networks is a wide open research gap for future study. Regarding multi-task learning, researchers have reported its benefit by citing increased performance levels in a variety of healthcare outcome prediction tasks. However, multi-task learning does not make clear which network layers, components, or types of extracted temporal patterns within the architectural design should be shared among the different tasks—as well as in which healthcare scenarios the multi-task strategy is most efficient. Transfer learning was the least studied method found in our systematic review, but it has promising prospects for further inquiry, as the scale of data and number of external data sets in published works increases.
- *Interpretation:* The most common approach to visualize important visits or medical codes on individual patients was the use of an attention mechanism in the neural network. Although individual-level interpretation is indeed important, as a future research direction, the use of population-level interpretation techniques to extract general patterns and identify specific medical events associated with target healthcare outcomes will be a boon for clinical decision makers.
- *Scalability:* Several studies confirm the generalizability of well-known deep time series prediction models to large hospital EHR datasets, even with high input dimensionality. However, analyzing advanced network architectures that have been proposed in recent works is a suggested venue for future research. Furthermore, some studies found that ensembles of traditional supervised learning methods have comparable performance to DL models, both on EHR and AC data. Important research gaps remain for establishing a proper comparison of DL against single or ensembled traditional ML models. In particular, it would be useful to identify patient, dimensionality, and missing value conditions in which DL models, with their higher complexity and runtimes, might be superfluous. This is a continual concern when considering the need for implementing real-time information systems that can better inform clinical decision makers.

A potential limitation of this systematic review is a possible incomplete retrieval of relevant studies on deep time series prediction. Although we included a wide set of keywords, it remains challenging to conduct an inclusive search strategy with an automatic query by keyword searching. We alleviated this concern by applying snowballing search strategies from the originally included publications. In other words, we assumed that any newer publication should reference one of the former included studies within their paper, especially well-known benchmarking models such as Doctor AI [7], RETAIN [22], DeepCare [65], and Deepr [63]. Another challenge was selectively differentiating the included studies from numerous other adjacent works when predicting

a single clinical outcome with a DL methodology. To achieve this, we implemented a full-text review step that included all papers that specifically mention patient representations or embedding strategies. In addition, we ensured that the authors' stated goals involved learning these representations at a patient level, not merely devising models to maximize performance on a specific disease prediction task. The aforementioned limitations pose a potential threat to selective bias in publication trends for any systematic review, but particularly one in which publication rates increase with recency, such as seen in the ever-increasing popularity of utilizing DL models on myriad applications, healthcare or otherwise.

REFERENCES

- [1] X. Min, B. Yu, and F. Wang. 2019. Predictive modeling of the hospital readmission risk from patients' claims data using machine learning: A case study on COPD. *Sci. Rep.* 9 (2019), 1–10. DOI: [10.1038/s41598-019-39071-y](https://doi.org/10.1038/s41598-019-39071-y)
- [2] Y. Sha, and M. D. Wang. 2017. Interpretable predictions of clinical outcomes with an attention-based recurrent neural network. In *Proceedings of the 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, New York, NY, 233–240. DOI: [10.1145/3107411.3107445](https://doi.org/10.1145/3107411.3107445)
- [3] C. Esteban, O. Staeck, S. Baier, Y. Yang, and V. Tresp. 2016. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *Proceedings of the 2016 IEEE International Conference on Healthcare Informatics (ICHI'16)*. IEEE, Los Alamitos, CA, 93–101. DOI: [10.1109/ICHI.2016.16](https://doi.org/10.1109/ICHI.2016.16)
- [4] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu. 2017. Boosting deep learning risk prediction with generative adversarial networks for electronic health records. In *Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM'17)*. 787–792. DOI: [10.1109/ICDM.2017.93](https://doi.org/10.1109/ICDM.2017.93)
- [5] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi. 2020. BEHRT: Transformer for electronic health records. *Sci. Rep.* 10 (2020), 1–12. DOI: [10.1038/s41598-020-62922-y](https://doi.org/10.1038/s41598-020-62922-y)
- [6] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. 2019. Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv:1602.03686 (2016). <http://arxiv.org/abs/1602.03686>.
- [7] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. 2020. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*. 301–318. <http://www.ncbi.nlm.nih.gov/pubmed/28286600>.
- [8] T. Tran, T. D. Nguyen, D. Phung, and S. Venkatesh. 2015. Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). *J. Biomed. Inform.* 54 (2015), 96–105. DOI: [10.1016/j.jbi.2015.01.012](https://doi.org/10.1016/j.jbi.2015.01.012)
- [9] Z. Sun, S. Peng, Y. Yang, X. Wang, and F. Li. 2019. A general fine-tuned transfer learning model for predicting clinical task acrossing diverse EHRs datasets. In *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'19)*. IEEE, Los Alamitos, CA, 490–495. DOI: [10.1109/BIBM47256.2019.8983098](https://doi.org/10.1109/BIBM47256.2019.8983098)
- [10] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar. 2014. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* 33 (2014), 1123–1131. DOI: [10.1377/hlthaff.2014.0041](https://doi.org/10.1377/hlthaff.2014.0041)
- [11] S. Saria and A. Goldenberg. 2015. Subtyping: What it is and its role in precision medicine. *IEEE Intell. Syst.* 30 (2015) 70–75. DOI: [10.1109/MIS.2015.60](https://doi.org/10.1109/MIS.2015.60)
- [12] Hrayr Harutyunyan, Hrant Khachatryan, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2020. Multitask learning and benchmarking with clinical time series data. *Sci. Data.* 6 (2019), 1–18. <https://www.nature.com/articles/s41597-019-0103-9>.
- [13] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, et al. 2018. Scalable and accurate deep learning with electronic health records. *npj Digit. Med.* 1 (2018), 1–10. DOI: [10.1038/s41746-018-0029-1](https://doi.org/10.1038/s41746-018-0029-1)
- [14] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah. 2017. Improving palliative care with deep learning. In *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'17)*. IEEE, Los Alamitos, CA, 311–316. DOI: [10.1109/BIBM.2017.8217669](https://doi.org/10.1109/BIBM.2017.8217669)
- [15] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis. 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J. Am. Med. Inform. Assoc.* 24 (2017), 198–208. DOI: [10.1093/jamia/ocw042](https://doi.org/10.1093/jamia/ocw042)
- [16] C. Lin, Y. Zhangy, J. Ivy, M. Capan, R. Arnold, J. M. Huddleston, and M. Chi. 2018. Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. In *Proceedings of the 2018 IEEE International Conference on Healthcare Informatics (ICHI'18)*. IEEE, Los Alamitos, CA, 219–228. DOI: [10.1109/ICHI.2018.00032](https://doi.org/10.1109/ICHI.2018.00032)

- [17] W. Chen, S. Wang, G. Long, L. Yao, Q. Z. Sheng, and X. Li. 2018. Dynamic illness severity prediction via multi-task RNNs for intensive care unit. In *Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM'18)*. IEEE, Los Alamitos, CA, 917–922. DOI: [10.1109/ICDM.2018.00111](https://doi.org/10.1109/ICDM.2018.00111)
- [18] Y. Cheng, F. Wang, P. Zhang, and J. Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. 432–440. DOI: [10.1137/1.9781611974348.49](https://doi.org/10.1137/1.9781611974348.49)
- [19] J. Zhang, J. Gong, and L. Barnes. 2017. HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In *Proceedings of the 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems, and Engineering Techniques (CHASE'17)*. 214–221. DOI: [10.1109/CHASE.2017.80](https://doi.org/10.1109/CHASE.2017.80)
- [20] L. Wang, H. Wang, Y. Song, and Q. Wang. 2019. MCPL-Based FT-LSTM: Medical representation learning-based clinical prediction model for time series events. *IEEE Access* 7 (2019), 70253–70264. DOI: [10.1109/ACCESS.2019.2919683](https://doi.org/10.1109/ACCESS.2019.2919683)
- [21] T. Zebin and T. J. Chaussalet. 2019. Design and implementation of a deep recurrent model for prediction of readmission in urgent care using electronic health records. In *Proceedings of the 16th IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'19)*. DOI: [10.1109/CIBCB.2019.8791466](https://doi.org/10.1109/CIBCB.2019.8791466)
- [22] E. Choi, M. Taha Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, J. Sun, and S. Health. 2020. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. ACM, New York, NY, 3512–3520. <http://papers.nips.cc/paper/6321-retain-an-interpretable-predictive-model-for-healthcare-using-reverse-time-attention-mechanism>.
- [23] E. Xu, S. Zhao, J. Mei, E. Xia, Y. Yu, and S. Huang. 2019. Multiple MACE risk prediction using multi-task recurrent neural network with attention. In *Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI'19)*. IEEE, Los Alamitos, CA. DOI: [10.1109/ICHI.2019.8904675](https://doi.org/10.1109/ICHI.2019.8904675)
- [24] B. L. P. Cheung and D. Dahl. 2018. Deep learning from electronic medical records using attention-based cross-modal convolutional neural networks. In *Proceedings of the 2018 IEEE International Conference on Biomedical Health Informatics (BHI'18)*. IEEE, Los Alamitos, CA, 222–225. DOI: [10.1109/BHI.2018.8333409](https://doi.org/10.1109/BHI.2018.8333409)
- [25] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, and A. Kronzer. 2018. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15 (2018), 1968–1978. DOI: [10.1109/TCBB.2018.2827029](https://doi.org/10.1109/TCBB.2018.2827029)
- [26] R. Amirkhan, M. Hoogendoorn, M. E. Numans, and L. Moons. 2017. Using recurrent neural networks to predict colorectal cancer among patients. In *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI'17)*. 1–8. DOI: [10.1109/SSCI.2017.8280826](https://doi.org/10.1109/SSCI.2017.8280826)
- [27] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. 2018. Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* 22 (2018), 1589–1604. DOI: [10.1109/JBHI.2017.2767063](https://doi.org/10.1109/JBHI.2017.2767063)
- [28] S. K. Pandey and R. R. Janghel. 2019. Recent deep learning techniques, challenges and its applications for medical healthcare system: A review. *Neural Process. Lett.* 50 (2019), 1907–1935. DOI: [10.1007/s11063-018-09976-2](https://doi.org/10.1007/s11063-018-09976-2)
- [29] C. Xiao, E. Choi, and J. Sun. 2018. Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *J. Am. Med. Inform. Assoc.* 25 (2018), 1419–1428. DOI: [10.1093/jamia/ocy068](https://doi.org/10.1093/jamia/ocy068)
- [30] K. Yazhini and D. Loganathan. 2019. A state of art approaches on deep learning models in healthcare: An application perspective. In *Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI'19)*. 195–200. DOI: [10.1109/ICOEI.2019.8862730](https://doi.org/10.1109/ICOEI.2019.8862730)
- [31] S. Srivastava, S. Soman, A. Rai, and P. K. Srivastava. 2017. Deep learning for health informatics: Recent trends and future directions. In *Proceedings of the 2017 International Conference on Advances in Computing, Communications, and Informatics (ICACCI'17)*. 1665–1670. DOI: [10.1109/ICACCI.2017.8126082](https://doi.org/10.1109/ICACCI.2017.8126082)
- [32] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny. 2021. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *J. Biomed. Inform.* 113 (2021), 103627. DOI: [10.1016/j.jbi.2020.103627](https://doi.org/10.1016/j.jbi.2020.103627)
- [33] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. Jim Zheng, and K. Roberts. 2021. Deep representation learning of patient data from electronic health records (EHR): A systematic review. *J. Biomed. Inform.* 115 (2021), 103671. DOI: [10.1016/j.jbi.2020.103671](https://doi.org/10.1016/j.jbi.2020.103671)
- [34] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ* 339 (2009), 332–336. DOI: [10.1136/BMJ.B2535](https://doi.org/10.1136/BMJ.B2535)
- [35] A. E. W. Johnson, T. J. Pollard, L. Shen, L. W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data.* 3 (2016), 1–9. DOI: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35)
- [36] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. 2018. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* 8 (2018), 1–12. DOI: [10.1038/s41598-018-24271-9](https://doi.org/10.1038/s41598-018-24271-9)

- [37] R. Yu, R. Zhang, Y. Jiang, and C. C. Y. Poon. 2020. Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients. *IEEE J. Biomed. Health Inform.* 24 (2020), 486–492. DOI: [10.1109/JBHI.2019.2916667](https://doi.org/10.1109/JBHI.2019.2916667)
- [38] W. Ge, J. W. Huh, Y. R. Park, J. H. Lee, Y. H. Kim, and A. Turchin. 2020. An interpretable ICU mortality prediction model based on logistic regression and recurrent neural networks with LSTM units. In *Proceedings of the AMIA Annual Symposium*. 460–469. [/pmc/articles/PMC6371274/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/3471274/).
- [39] W. Caicedo-Torres and J. Gutierrez. 2019. ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU. *J. Biomed. Inform.* 98 (2019), 103269. DOI: [10.1016/j.jbi.2019.103269](https://doi.org/10.1016/j.jbi.2019.103269)
- [40] D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang. 2020. Combining structured and unstructured data for predictive models: A deep learning approach. *BMC Med. Inform. Decis. Mak.* 20, 1 (2020), 280. DOI: [10.1186/s12911-020-01297-6](https://doi.org/10.1186/s12911-020-01297-6)
- [41] B. Shickel, T. J. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, and P. Rashidi. 2019. DeepSOFA: A continuous acuity score for critically ill patients using clinically interpretable deep learning. *Sci. Rep.* 9 (2019), 1–12. DOI: [10.1038/s41598-019-38491-0](https://doi.org/10.1038/s41598-019-38491-0)
- [42] S. Purushotham, C. Meng, Z. Che, and Y. Liu. 2018. Benchmarking deep learning models on large healthcare datasets. *J. Biomed. Inform.* 83 (2018), 112–134. DOI: [10.1016/j.jbi.2018.04.007](https://doi.org/10.1016/j.jbi.2018.04.007)
- [43] P. Gupta, P. Malhotra, J. Narwariya, L. Vig, and G. Shroff. 2020. Transfer learning for clinical time series analysis using deep neural networks. *J. Healthc. Inform. Res.* 4 (2020), 112–137. DOI: [10.1007/s41666-019-00062-3](https://doi.org/10.1007/s41666-019-00062-3)
- [44] S. Baker, W. Xiang, and I. Atkinson. 2020. Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: A hybrid neural network approach. *Sci. Rep.* 10 (2020), 1–12. DOI: [10.1038/s41598-020-78184-7](https://doi.org/10.1038/s41598-020-78184-7)
- [45] K. Yu, M. Zhang, T. Cui, and M. Hauskrecht. 2020. Monitoring ICU mortality risk with a long short-term memory recurrent neural network. In *Proceedings of the Pacific Symposium on Biocomputing*. 103–114. DOI: [10.1142/9789811215636_0010](https://doi.org/10.1142/9789811215636_0010)
- [46] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2017. Using recurrent neural network models for early detection of heart failure onset. *J. Am. Med. Inform. Assoc.* 24 (2017), 361–370.
- [47] C. Yin, R. Zhao, B. Qian, X. Lv, and P. Zhang. 2019. Domain knowledge guided deep learning with electronic health records. In *Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM'19)*. IEEE, Los Alamitos, CA, 738–747. DOI: [10.1109/ICDM.2019.00084](https://doi.org/10.1109/ICDM.2019.00084)
- [48] W. W. Wang, H. Li, L. Cui, X. Hong, and Z. Yan. 2018. Predicting clinical visits using recurrent neural networks and demographic information. In *Proceedings of the IEEE 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD'18)*. 785–789. DOI: [10.1109/CSCWD.2018.8465194](https://doi.org/10.1109/CSCWD.2018.8465194)
- [49] R. Ju, P. Zhou, S. Wen, W. Wei, Y. Xue, X. Huang, and X. Yang. 2020. 3D-CNN-SPP: A patient risk prediction system from electronic health records via 3D CNN and spatial pyramid pooling. *IEEE Trans. Emerg. Topics Comput. Intell.* 5 (2020), 247–261. DOI: [10.1109/tetci.2019.2960474](https://doi.org/10.1109/tetci.2019.2960474)
- [50] L. Rasmy, W. J. Zheng, H. Xu, D. Zhi, Y. Wu, N. Wang, H. Wu, X. Geng, and F. Wang. 2018. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set. *J. Biomed. Inform.* 8 (2018), 11–16. DOI: [10.1016/j.jbi.2018.06.011](https://doi.org/10.1016/j.jbi.2018.06.011)
- [51] G. Maragatham and S. Devi. 2019. LSTM model for prediction of heart failure in big data. *J. Med. Syst.* 3 (2019), 1–13. DOI: [10.1007/s10916-019-1243-3](https://doi.org/10.1007/s10916-019-1243-3)
- [52] X. Zhang, B. Qian, Y. Li, C. Yin, X. Wang, and Q. Zheng. 2019. KnowRisk: An interpretable knowledge-guided model for disease risk prediction. In *Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM'19)*. IEEE, Los Alamitos, CA, 1492–1497. DOI: [10.1109/ICDM.2019.00196](https://doi.org/10.1109/ICDM.2019.00196)
- [53] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. GRAM: Graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM Conference on Knowledge Discovery and Data Mining (KDD'17)*. 787–795.
- [54] T. Ma, C. Xiao, and F. Wang. 2018. Health-ATM: A deep architecture for multifaceted patient health record representation and risk prediction. In *Proceedings of the SIAM International Conference on Data Mining (SDM'18)*. 261–269. DOI: [10.1137/1.9781611975321.30](https://doi.org/10.1137/1.9781611975321.30)
- [55] Edward Choi, Cao Xiao, Walter F. Stewart, and Jimeng Sun. 2021. MiME: Multilevel medical embedding of electronic health records for predictive healthcare. In *Proceedings of the 2018 Conference on Neural Information Processing Systems (NIPS'18)*. 4552–4562. <https://dl.acm.org/doi/10.5555/3327345.3327366>.
- [56] J. R. Ayala Solares, F. E. Diletta Raimondi, Y. Zhu, F. Rahimian, D. Canoy, J. Tran, A. C. Pinho Gomes, et al. 2020. Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* 101 (2020), 103337. DOI: [10.1016/j.jbi.2019.103337](https://doi.org/10.1016/j.jbi.2019.103337)
- [57] J. Zhang, K. Kowsari, J. H. Harrison, J. M. Lobo, and L. E. Barnes. 2018. Patient2Vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* 6 (2018), 65333–65346. DOI: [10.1109/ACCESS.2018.2875677](https://doi.org/10.1109/ACCESS.2018.2875677)

- [58] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, A. Kronzer, and S. Louis. 2017. Cost-sensitive deep learning for early readmission prediction at a major hospital. In *Proceedings of the 8th International Workshop on Biological Knowledge Discovery from Data (BioKDD'17)*. ACM, New York, NY.
- [59] Y. W. Lin, Y. Zhou, F. Faghri, M. J. Shaw, and R. H. Campbell. 2019. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *PLoS One* 14 (2019), e0218942. DOI: [10.1371/journal.pone.0218942](https://doi.org/10.1371/journal.pone.0218942)
- [60] S. Barbieri, J. Kemp, O. Perez-Concha, S. Kotwal, M. Gallagher, A. Ritchie, and L. Jorm. 2020. Benchmarking deep learning architectures for predicting readmission to the ICU and describing patients-at-risk. *Sci. Rep.* 10 (2020), Article 1111, 10 pages. DOI: [10.1038/s41598-020-58053-z](https://doi.org/10.1038/s41598-020-58053-z)
- [61] A. Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk. 2019. Readmission prediction using deep learning on electronic health records. *J. Biomed. Inform.* 97 (2019), 103256. DOI: [10.1016/j.jbi.2019.103256](https://doi.org/10.1016/j.jbi.2019.103256)
- [62] B. K. Reddy and D. Delen. 2018. Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. *Comput. Biol. Med.* 10 (2018), 199–209. DOI: [10.1016/j.combiomed.2018.08.029](https://doi.org/10.1016/j.combiomed.2018.08.029)
- [63] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh. 2017. Deep: A convolutional net for medical records. *IEEE J. Biomed. Health Inform.* 21, 1 (2017), 22–30. DOI: [10.1109/JBHI.2016.2633963](https://doi.org/10.1109/JBHI.2016.2633963)
- [64] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. 2015. Learning to diagnose with LSTM recurrent neural networks. arXiv:1511.03677 (2015). <http://arxiv.org/abs/1511.03677>.
- [65] T. Pham, T. Tran, D. Phung, and S. Venkatesh. 2016. DeepCare: A deep dynamic memory model for predictive medicine. In *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science, Vol. 9652. Springer, 30–41. DOI: [10.1007/978-3-319-31750-2_3](https://doi.org/10.1007/978-3-319-31750-2_3)
- [66] Y. Yang, X. Zheng, and C. Ji. 2019. Disease prediction model based on BiLSTM and attention mechanism. In *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'19)*. IEEE, Los Alamitos, CA, 1141–1148. DOI: [10.1109/BIBM47256.2019.8983378](https://doi.org/10.1109/BIBM47256.2019.8983378)
- [67] W. Guo, W. Ge, L. Cui, H. Li, and L. Kong. 2019. An interpretable disease onset predictive model using crossover attention mechanism from electronic health records. *IEEE Access* 7 (2019), 134236–134244. DOI: [10.1109/ACCESS.2019.2928579](https://doi.org/10.1109/ACCESS.2019.2928579)
- [68] T. Wang, Y. Tian, and R. G. Qiu. 2020. Long short-term memory recurrent neural networks for multiple diseases risk prediction by leveraging longitudinal medical records. *IEEE J. Biomed. Health Inform.* 24, 8 (2020), 2337–2346. DOI: [10.1109/JBHI.2019.2962366](https://doi.org/10.1109/JBHI.2019.2962366)
- [69] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2020. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 2020 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'20)*. ACM, New York, NY, 1903–1911. <https://dl.acm.org/doi/abs/10.1145/3097983.3098088>
- [70] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. KAME: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 2018 ACM International Conference on Information and Knowledge Management (CIKM'18)*. 743–752. <https://dl.acm.org/doi/abs/10.1145/3269206.3271701>
- [71] T. Pham, T. Tran, D. Phung, and S. Venkatesh. 2017. Predicting healthcare trajectories from medical records: A deep learning approach. *J. Biomed. Inform.* 69 (2017), 218–229. DOI: [10.1016/j.jbi.2017.04.001](https://doi.org/10.1016/j.jbi.2017.04.001)
- [72] J. M. Lee and M. Hauskrecht. 2019. Recent context-aware LSTM for clinical event time-series prediction. In *Proceedings of the 17th Conference on Artificial Intelligence in Medicine (AIME'19)*. 13–23. DOI: [10.1007/978-3-030-21642-9_3](https://doi.org/10.1007/978-3-030-21642-9_3)
- [73] D. Lee, X. Jiang, and H. Yu. 2020. Harmonized representation learning on dynamic EHR graphs. *J. Biomed. Inform.* 106 (2020), 103426. DOI: [10.1016/j.jbi.2020.103426](https://doi.org/10.1016/j.jbi.2020.103426)
- [74] Z. C. Lipton, D. C. Kale, R. Wetzel, and L. K. Whittier. 2021. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*. 253–270. <http://proceedings.mlr.press/v56/Lipton16.html>.
- [75] T. Bai, A. K. Chanda, B. L. Egleston, and S. Vucetic. 2017. Joint learning of representations of medical concepts and words from EHR data. In *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM'17)*. IEEE, Los Alamitos, CA, 764–769. DOI: [10.1109/BIBM.2017.8217752](https://doi.org/10.1109/BIBM.2017.8217752)
- [76] D. Liu, Y. L. Wu, X. Li, and L. Qi. 2020. Medi-Care AI: Predicting medications from billing codes via robust recurrent neural networks. *Neural Netw.* 124 (2020), 109–116. DOI: [10.1016/j.neunet.2020.01.001](https://doi.org/10.1016/j.neunet.2020.01.001)
- [77] M. Zhang, C. R. King, M. Avidan, and Y. Chen. 2021. Hierarchical attention propagation for healthcare representation learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 249–256. <https://dl.acm.org/doi/abs/10.1145/3394486.3403067>
- [78] Z. Qiao, Z. Zhang, X. Wu, S. Ge, and W. Fan. 2020. MHM: Multi-modal clinical data based hierarchical multi-label diagnosis prediction. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20)*. 1841–1844. DOI: [10.1145/3397271](https://doi.org/10.1145/3397271)

- [79] J. Park, J. W. Kim, B. Ryu, E. Heo, S. Y. Jung, and S. Yoo. 2019. Patient-level prediction of cardio-cerebrovascular events in hypertension using nationwide claims data. *J. Med. Internet Res.* 21, 2 (2019), e11757. DOI: [10.2196/11757](https://doi.org/10.2196/11757)
- [80] Y. An, N. Huang, X. Chen, F. Wu, and J. Wang. 2019. High-risk prediction of cardiovascular diseases via attention-based deep neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 18 (2019), 1093–1105. DOI: [10.1109/tcbb.2019.2935059](https://doi.org/10.1109/tcbb.2019.2935059)
- [81] H. Duan, Z. Sun, W. Dong, and Z. Huang. 2019. Utilizing dynamic treatment information for MACE prediction of acute coronary syndrome. *BMC Med. Inform. Decis. Mak.* 19 (2019), 1–11. DOI: [10.1186/s12911-018-0730-7](https://doi.org/10.1186/s12911-018-0730-7)
- [82] S. Park, Y. J. Kim, J. W. Kim, J. J. Park, B. Ryu, and J. W. Ha. 2018. Interpretable prediction of vascular diseases from electronic health records via deep attention networks. In *Proceedings of the 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE'18)*. IEEE, Los Alamitos, CA, 110–117. DOI: [10.1109/BIBE.2018.00028](https://doi.org/10.1109/BIBE.2018.00028)
- [83] Y. Zhang, C. Lin, M. Chi, J. Ivy, M. Capan, and J. M. Huddleston. 2017. LSTM for septic shock: Adding unreliable labels to reliable predictions. In *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data'17)*. IEEE, Los Alamitos, CA, 1233–1242. DOI: [10.1109/BigData.2017.8258049](https://doi.org/10.1109/BigData.2017.8258049)
- [84] Y. Zhang, X. Yang, J. Ivy, and M. Chi. 2019. Time-aware adversarial networks for adapting disease progression modeling. In *Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI'19)*. IEEE, Los Alamitos, CA. DOI: [10.1109/ICHI.2019.8904698](https://doi.org/10.1109/ICHI.2019.8904698)
- [85] S. D. Wickramaratne and M. D. Shaad Mahmud. 2020. Bi-directional gated recurrent unit based ensemble model for the early detection of sepsis. In *Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC'20)*. 70–73. DOI: [10.1109/EMBC44109.2020.9175223](https://doi.org/10.1109/EMBC44109.2020.9175223)
- [86] P. Svenson, G. Haralabopoulos, and M. Torres Torres. 2020. Sepsis deterioration prediction using channelled long short-term memory networks. In *Proceedings of the 18th International Conference on Artificial Intelligence in Medicine (AIME'20)*. 359–370. DOI: [10.1007/978-3-030-59137-3_32](https://doi.org/10.1007/978-3-030-59137-3_32)
- [87] J. Fagerström, M. Bång, D. Wilhelms, and M. S. Chew. 2019. LiSep LSTM: A machine learning algorithm for early detection of septic shock. *Sci. Rep.* 9, 9 (2019), 1–8. DOI: [10.1038/s41598-019-51219-4](https://doi.org/10.1038/s41598-019-51219-4)
- [88] R. Mohammadi, S. Jain, S. Agboola, R. Palacholla, S. Kamarthi, and B. C. Wallace. 2019. Learning to identify patients at risk of uncontrolled hypertension using electronic health records data. *AMIA Jt. Summits Transl. Sci. Proc.* 2019 (2019), 533–542. <http://www.ncbi.nlm.nih.gov/pubmed/31259008>.
- [89] X. Ye, Q. T. Zeng, J. C. Facelli, D. I. Brixner, M. Conway, and B. E. Bray. 2020. Predicting optimal hypertension treatment pathways using recurrent neural networks. *Int. J. Med. Inform.* 139 (2020), 104122. DOI: [10.1016/j.ijmedinf.2020.104122](https://doi.org/10.1016/j.ijmedinf.2020.104122)
- [90] H. C. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen, B. S. Kaas-Hansen, P. Toft, J. Schierbeck, T. Strøm, et al. 2020. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records. *Lancet Digit. Health* 2 (2020), e179–e191. DOI: [10.1016/S2589-7500\(20\)30018-2](https://doi.org/10.1016/S2589-7500(20)30018-2)
- [91] Kaiping Zheng, Wei Wang, Jinyang Gao, Kee Yuan Ngiam, Bengchin Ooi, and Weiluen Yip. 2017. Capturing feature-level irregularity in disease progression modeling. In *Proceedings of the 2017 ACM International Conference on Information and Knowledge Management (CIKM'17)*. 1579–1588. <https://dl.acm.org/doi/abs/10.1145/3132847.3132944>
- [92] Q. Suo, F. Ma, G. Canino, J. Gao, A. Zhang, P. Veltri, and G. Agostino. 2017. A multi-task framework for monitoring health conditions via attention-based recurrent neural networks. *AMIA Annu. Symp. Proc.* 2017 (2017), 1665.
- [93] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, et al. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 572 (2019), 116–119. DOI: [10.1038/s41586-019-1390-1](https://doi.org/10.1038/s41586-019-1390-1)
- [94] R. Qiu, Y. Jia, F. Wang, P. Divakarmurthy, S. Vinod, B. Sabir, and M. Hadzikadic. 2020. Predictive modeling of the total joint replacement surgery risk: A deep learning based approach with claims data. *AMIA Jt. Summits Transl. Sci. Proc.* 2019 (2019), 562–571. <https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp>.
- [95] Y. Ge, Q. Wang, L. Wang, H. Wu, C. Peng, J. Wang, Y. Xu, G. Xiong, Y. Zhang, and Y. Yi. 2019. Predicting post-stroke pneumonia using deep neural network approaches. *Int. J. Med. Inform.* 132 (2019), 103986. DOI: [10.1016/j.ijmedinf.2019.103986](https://doi.org/10.1016/j.ijmedinf.2019.103986)
- [96] N. Razavian, J. Marcus, and D. Sontag. 2021. Multi-task prediction of disease onsets from longitudinal lab tests. In *Proceedings of the 1st Machine Learning for Healthcare Conference*. 73–100.
- [97] J. Rebane, I. Karlsson, and P. Papapetrou. 2019. An investigation of interpretable deep learning for adverse drug event prediction. In *Proceedings of the 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS'19)*. 337–342. DOI: [10.1109/CBMS.2019.00075](https://doi.org/10.1109/CBMS.2019.00075)
- [98] M. A. Morid, O. R. L. Sheng, K. Kawamoto, and S. Abdelrahman. 2020. Learning hidden patterns from patient multi-variate time series data using convolutional neural networks: A case study of healthcare cost prediction. *J. Biomed. Inform.* 111 (2020), 103565. DOI: [10.1016/j.jbi.2020.103565](https://doi.org/10.1016/j.jbi.2020.103565)

- [99] Y. Xiang, H. Ji, Y. Zhou, F. Li, J. Du, L. Rasmy, S. Wu, et al. 2020. Asthma exacerbation prediction and risk factor analysis based on a time-sensitive, attentive neural network: Retrospective cohort study. *J. Med. Internet Res.* 22 (2020), e16981. DOI: [10.2196/16981](https://doi.org/10.2196/16981)
- [100] C. Gao, C. Yan, S. Osmundson, B. A. Malin, and Y. Chen. 2019. A deep learning approach to predict neonatal encephalopathy from electronic health records. In *Proceedings of the 2019 IEEE International Conference on Healthcare Informatics (ICHI'19)*. IEEE, Los Alamitos, CA. DOI: [10.1109/ICHI.2019.8904667](https://doi.org/10.1109/ICHI.2019.8904667)
- [101] L. Ma and Y. Zhang. 2015. Using Word2Vec to process big text data. In *Proceedings of the 2015 IEEE International Conference on Big Data (Big Data'15)*. 2895–2897. DOI: [10.1109/BigData.2015.7364114](https://doi.org/10.1109/BigData.2015.7364114)
- [102] W. Cheng, C. Greaves, and M. Warren. 2006. From n-gram to skipgram to conngam. *Int. J. Corpus Linguistics* 11 (2006), 411–433. DOI: [10.1075/ijcl.11.4.04che](https://doi.org/10.1075/ijcl.11.4.04che)
- [103] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. 2001. SNOMED clinical terms: Overview of the development process and project status. In *Proceedings of the AMIA Annual Symposium*. 662–666.
- [104] P. Ernst, A. Siu, and G. Weikum. 2015. KnowLife: A versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC Bioinform.* 16 (2015), 1–13. DOI: [10.1186/s12859-015-0549-5](https://doi.org/10.1186/s12859-015-0549-5)
- [105] A. Shrikumar, P. Greenside, and A. Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*. 3145–3153. <http://goo.gl/RM8jvH>.
- [106] E. Winter. 2002. The Shapley value. In *Handbook of Game Theory with Economic Applications*. Elsevier, 2025–2054. DOI: [10.1016/S1574-0005\(02\)03016-3](https://doi.org/10.1016/S1574-0005(02)03016-3)
- [107] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. 2012. Predicting in-hospital mortality of ICU patients: The PhysioNet/Computing in Cardiology Challenge 2012. *Comput. Cardiol.* 39 (2012), 245–248.
- [108] C. Fang and C. Wang. 2020. Time series data imputation: A survey on deep learning approaches. arXiv:2011.11347 (2020). <http://arxiv.org/abs/2011.11347>.
- [109] Y. Lecun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521 (2015), 436–444. DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539)
- [110] S. M. Boker, S. S. Tiberio, and R. G. Moulder. 2018. Robustness of time delay embedding to sampling interval misspecification. In *Continuous Time Modeling in the Behavioral and Related Sciences*. Springer, 239–258. DOI: [10.1007/978-3-319-77219-6_10](https://doi.org/10.1007/978-3-319-77219-6_10)
- [111] W. Liu, P. Zhou, Z. Wang, Z. Zhao, H. Deng, and Q. Ju. 2020. FastBERT: A self-distilling BERT with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6035–6044.
- [112] C. P. Rees, S. Hawkesworth, S. E. Moore, B. L. Dondeh, and S. A. Unger. 2016. Factors affecting access to healthcare: An observational study of children under 5 years of age presenting to a rural Gambian primary healthcare centre. *PLoS One* 11 (2016), e0157790. DOI: [10.1371/journal.pone.0157790](https://doi.org/10.1371/journal.pone.0157790)
- [113] X. Li, Y. Zhou, N. C. Dvornek, Y. Gu, P. Ventola, and J. S. Duncan. 2020. Efficient Shapley explanation for features importance estimation under uncertainty. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*. Lecture Notes in Computer Science, Vol. 12261. Springer, 792–801. DOI: [10.1007/978-3-030-59710-8_77](https://doi.org/10.1007/978-3-030-59710-8_77).
- [114] C. Beck, A. Jentzen, and B. Kuckuck. 2021. Full error analysis for the training of deep neural networks. arXiv: 1910.00121v2 (2021). <https://arxiv.org/abs/1910.00121v2>.

Received 2 August 2021; revised 19 March 2022; accepted 11 April 2022