Machine Learning for Clinical Outcome Prediction

Farah Shamout, Tingting Zhu, and David A. Clifton

(Clinical Application Review)

***Abstract*—Clinical decision-making in healthcare is already being influenced by predictions or recommendations made by data-driven machines. Numerous machine learning applications have appeared in the latest clinical literature, especially for outcome prediction models, with outcomes ranging from mortality and cardiac arrest to acute kidney injury and arrhythmia. In this review article, we summarize the state-of-the-art in related works covering data processing, inference, and model evaluation, in the context of outcome prediction models developed using data extracted from electronic health records. We also discuss limitations of prominent modeling assumptions and highlight opportunities for future research.**

***Index Terms*—Learning (artificial intelligence), machine learning, decision support systems, electronic medical records, big data applications.**

I. Introduction

RECENT artificial intelligence (AI) developments seek to positively impact medicine and clinical practice [1]. Machine learning (ML), an application of AI, recognizes patterns within large quantities of medical data to make future predictions, with many successful applications in natural language processing, computer vision applications, and automatic speech recognition [2]–[5]. Applications of ML have been successful across several medical domains, such as disease prediction [6] using various data modalities, including speech signals and medical imaging [7]–[10], as well as clinical outcome prediction to detect deterioration, such as cardiac arrest, mortality, or intensive care unit (ICU) admission [11]–[14]. The intention of this paper is to provide a technology survey of recent works on clinical outcome prediction models that illustrate the respective areas of the fields in which they are described.

In general, designing an ML system involves a multidisciplinary effort that extends from data engineering to training and evaluating a predictive model. We consider the general model as a mapping of an input to an output:

$$f : \mathbf{X} \rightarrow y \tag{1}$$

where $f(.)$ is a function consisting of parameters Θ, $\mathbf{X}$ is the *input* and y is the *output*. For example, $\mathbf{X}$ can consist of vital signs measurements of the patient, such as heart rate, blood pressure, and respiratory rate, and y can represent a binary label indicating the occurrence of ICU transfer or cardiac arrest during the patient's hospital stay [14].

Fig. 1 depicts the typical pipeline of a ML application, starting from the input $\mathbf{X}$, and ending with the corresponding output represented by y. The first task learns to extract intermediary features (Section IV) while the second task learns from patterns in the data to produce the predicted label (Section V). Such models are usually assessed based on clinical utility and interpretability (Section VI).

As we discuss related works throughout this review, we also provide an intuitive explanation of the ML techniques used for feature extraction or predictive inference. In general, 'learning' how to map the input to the output involves approximating the parameters of the model $f(.)$, a loss function $\mathcal{L}(y, \hat{y}|\Theta)$, and an optimization algorithm. The loss function $\mathcal{L}(y, \hat{y}|\Theta)$, also known as the cost function, measures the dissimilarity between the true labels y and the values $\hat{y}$ predicted by the approximated model (e.g., mean square error, binary cross-entropy, etc.). An optimization algorithm, such as gradient descent [15], minimizes $\mathcal{L}(y, \hat{y}|\Theta)$ in an iterative manner based on the examples present in the dataset.

II. Clinical Context and Frameworks of Outcome Prediction Models

Care pathways within hospitals vary largely due to the diversity of admitted patients. Thus, an understanding of the clinical context is key for developing machine learning models that can be incorporated within existing medical processes. As shown in Fig. 2, a patient may be hospitalized as an emergency or elective admission, where the latter constitutes a routine

Manuscript received January 4, 2020; accepted June 20, 2020. Date of publication July 7, 2020; date of current version January 22, 2021. *(Corresponding author: Farah Shamout.)* This research was supported by a Health Innovation Challenge Fund (HICF-R9-524; WT-103703/Z/14/Z), a parallel funding partnership between the Department of Health and Wellcome Trust, and in part by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. David A. Clifton is supported by an EPSRC Grand Challenge Award.

Farah Shamout is with the Engineering Division at NYU Abu Dhabi, Abu Dhabi 129 188, UAE (e-mail: fs999@nyu.edu).

Tingting Zhu and David A. Clifton are with the Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, U.K. (e-mail: tingting.zhu@eng.ox.ac.uk; davidc@robots.ox.ac.uk).

Digital Object Identifier 10.1109/RBME.2020.3007816

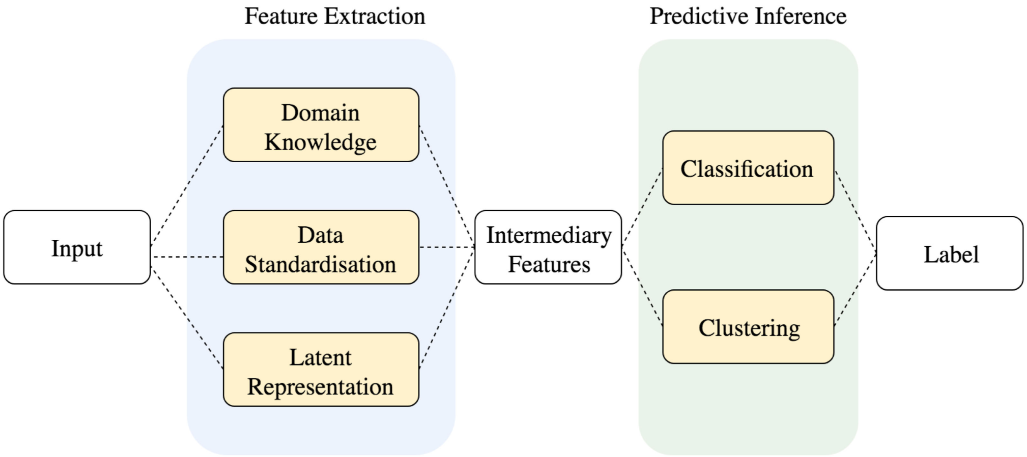


Fig. 1. General ML pipeline that maps an input to a label. The two main steps of the pipeline are (i) extraction of an intermediary feature space and (ii) label prediction using a classification or clustering algorithm.

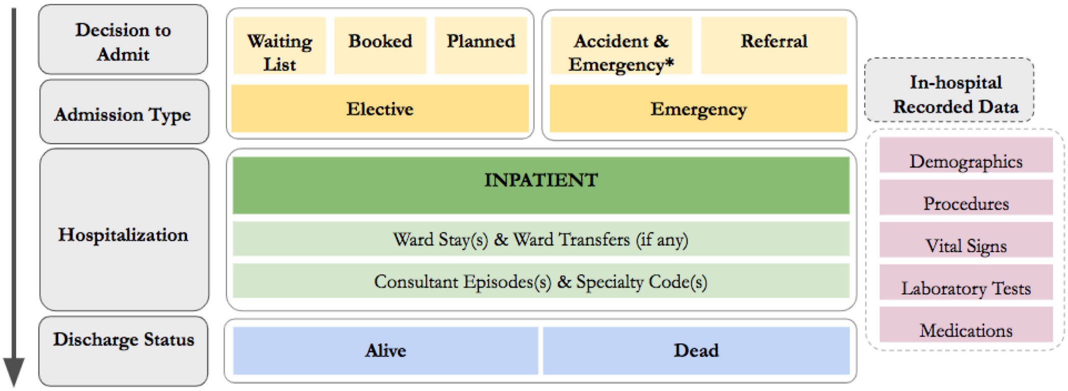


Fig. 2. Visualization of the patient flow in a hospital: Patient is either admitted as an elective or emergency admission, monitored in ward stay(s) during consultant episode(s). Patient may transfer from one ward to another, or may change the consultant during the in-hospital stay. ** Accident & Emergency patients may be admitted as inpatients or just discharged.*

procedure. During hospitalization, different types of data are routinely collected from the patient for monitoring purposes.

Patient monitoring tools, such as early warning systems [16], are widespread across different hospital wards to continuously assess for patient deterioration. The definition of what exactly constitutes clinical deterioration has evolved over time based on the data collection and processing techniques. Early attempts to define clinical deterioration focused on medical neglect and its end result of clinical complications [17]. Subsequent studies focused on more discrete clinical events, such as severe sepsis, unexpected cardiac arrest, ICU admission or mortality [18], [19], and tend to select one or more end-point measures of clinical deterioration. Such events incur high costs of prolonged hospital stays, litigation, staff time, impact on patients and staff, and broader economic consequences [20]. The latter definition is the most popular one, as it enables researchers to group patients into discrete classes, such as deteriorating (i.e., those who experience an outcome) and non-deteriorating (i.e., those who are discharged without experiencing any outcomes), and as such infer the y labels.

The framework of outcome prediction models also varies across the literature. Some studies predict the risk of an outcome only once using the patient's first N hours of data after admission, such as 24 or 48 hours [21]. Others choose to predict the risk of an outcome, such as ICU readmission, using the patient's last N hours of data prior to discharge. Another common methodology is to develop a real-time alerting score, which computes the risk of deterioration every time a set of clinical observations is collected [22], as in clinical early warning systems [23].

III. Electronic Health Records

Various types of data can be used to develop outcome prediction models, such as imaging, speech, or claims data [24]. Here, we focus on data extracted from electronic health records (EHR), which are being increasingly deployed in hospitals worldwide. EHRs are used in hospitals to store longitudinal information of patients collected in a care delivery setting. Such information includes patient demographics, vital signs, medications, laboratory data, and description of any outcomes that may have occurred to the patient during hospitalization, or shortly after discharge.

Data extracted from an EHR database can be used to develop and evaluate ML models. The dataset is typically split into a

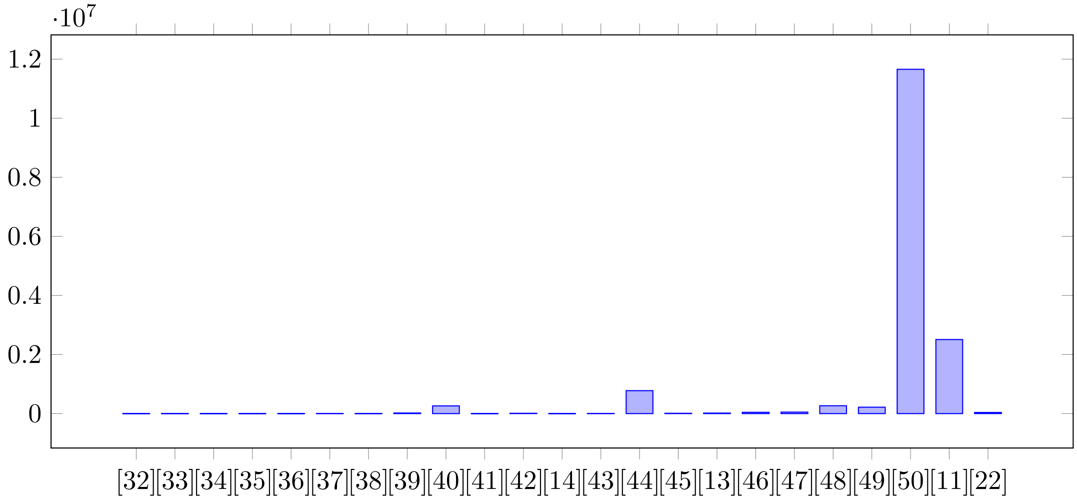


Fig. 3. Dataset sizes reported in the literature in ascending order from left (2008) to right (2019). The vertical axis represents the dataset size, in terms of the number of patient admissions, and the horizontal axis represents the reference number. There is an increase of six orders of magnitude between 2008 and 2019 in terms of dataset size, highlighting the increased availability of EHR data for ML research.

training set and a *test set*,[1] either by a random or a nonrandom split based on location or time. According to the *Transparent Reporting of a multi-variable prediction model for Individual Prognosis Or Diagnosis* (TRIPOD) statement, the nonrandom split by time is the strongest evaluation technique as it avoids random variations between the training and testing sets [25]. During model learning, the *training set* is used to optimize the parameters Θ of the model. The trained model is then evaluated on the held-out *test set* using various performance metrics.

Fig. 3 shows the overall dataset sizes, in terms of number of patient admissions, reported in studies published in the last decade (arranged in chronological order from left to right), extracted from EHRs. There is an increase of six orders of magnitude between 2008 and 2019, which highlights the increased accessibility to EHR data for research purposes. Most datasets are reported to be private, and there have only been a few notable efforts to release open access datasets, such as the MIMIC-III database [26]. Data and resource sharing is important for the advancements of the field.

It is also commonly agreed that data in EHRs may reflect the recording process present in the hospital rather than being a direct reflection of patient physiology [27]. First, EHRs are complex as they may include structured and unstructured data; an example of the latter is textual information which could require natural language processing techniques to process [28]. Additionally, categorical data, such as diagnostic coding, may adopt different coding systems across different institutions.

Another important dimension is data completeness, which may be defined as the proportion of observations that are actually recorded in the system [29]. Incompleteness of EHRs can be a result of health service fragmentation due to inefficient communication following patient transfer among institutions; the recording of data taking place only during healthcare episodes that correspond to illness, or the increased personalisation of attributes per patient [27], [30]. Completeness may also vary across institutions based on adopted protocols.

The third challenge is the accuracy of the data, or "the proportion of recorded observations in the system that are correct" [29]. Errors can occur while clinical staff observe a patient or record data, and their occurrence may be influenced by random and systematic errors such as billing requirements or avoidance of liability [27]. The accuracy of EHRs can be assessed by checking agreement between different elements within the EHR (such as assigned diagnosis and supplied medications), or by verifying whether values are within expected ranges [31].

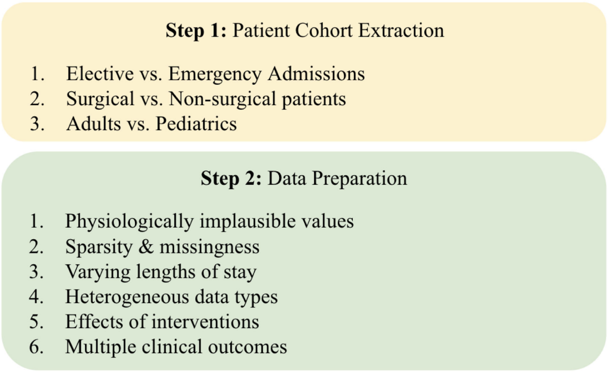
Finally, it is important to verify whether the data was recorded within a reasonable period of time [31]. For example, the recorded collection time of vital signs may precede the time of admission. Although this aspect of data quality is highly dependent on the efficiency of the clinical staff, it also depends on the work flow protocols adopted at different institutions. Timeliness of data must be assessed to evaluate the chronology of data elements in relation to admission or discharge decisions, for example laboratory results prior to admission may be considered as part of subsequent admission, or death within 24 hours of discharge can be considered as in-hospital mortality.

This imposes challenges on the usability of the data, which usually incurs preliminary data pre-processing as shown in Fig. 4. The first step is to define an inclusion and exclusion criteria to extract the patient cohort of interest. The second step involves setting assumptions to aid the analysis of the heterogeneous data, such as defining a minimum length of stay. Finally, meaningful features as input variables to the ML model can be extracted using a variety of techniques.

IV. Feature Extraction

The performance of clinical predictive models relies on the feature representation of the data, as in other domains [32]. As reported in related works, feature extraction generally involves

[1] In clinical studies, the test set is usually termed the validation set, not to be confused with the portion of the training set used for ML-oriented tasks, such as hyperparameter selection.



Step 1: Patient Cohort Extraction

1. Elective vs. Emergency Admissions
2. Surgical vs. Non-surgical patients
3. Adults vs. Pediatrics

Step 2: Data Preparation

1. Physiologically implausible values
2. Sparsity & missingness
3. Varying lengths of stay
4. Heterogeneous data types
5. Effects of interventions
6. Multiple clinical outcomes

Fig. 4. Clinical outcome prediction models first extract a cohort of interest based on a specific patient inclusion and exclusion criteria, and then prepare the data for further downstream tasks by assessing the characteristics of the raw dataset.

at least one of domain-expertise for hand-crafted features (Section IV-A), data standardization (Section IV-B), or representation learning (Section IV-C).

A. Hand-Crafted Features

Domain expertise is commonly used to provide guidance on the design of the data pre-processing pipeline. This involves (i) preliminary feature selection from the input space, (ii) designing hand-crafted features, and (iii) incorporating prior knowledge of the structure of the data in the model design.

Examples of hand-crafted features in related works are pulse pressure [33], [34], shock index [33], [35], [36], mean arterial pressure [33], [37], oxygen delivery index [36], absolute successive difference of heart rate, estimated cardiac output, slope of fitted regression lines, or slope projections [35]. Statistical measures can be obtained from the distributions of the raw data, such as minimum and maximum extremes, moments (mean, standard deviation, and skewness), percentiles or the difference between two percentiles [35], [38].

Previous research also computed time series features from waveform data [12], [34], [39], [40]. Those features can be categorized into four types: data adaptive, non-data adaptive, model-based and data-dictated approaches [41]. Fourier and wavelet transforms, for instance, decompose raw signals into frequency and wavelets respectively. Time domain, Poincaré nonlinear, cross-correlation analysis and geometric measures have also been used to investigate variability of vital signs [12], [34].

Deriving hand-crafted features is a powerful tool in the design of ML models and has been used extensively over the years. However, it is a time-consuming and labor-intensive process, requires expert knowledge, and may not scale well to new problems.

B. Data Standardization

ML algorithms require further data preparation steps to ensure stability of learning. Here, related works reduce the noise, sparsity and irregularity of the clinical data, as well as align the scales of the various predictor variables.

1) Time-Series Modeling: Time-series modeling is widely used in studies pertaining to early warning models [42], [43]. It is often used either (i) to infer a pattern of the physiological trajectory or (ii) as an interpolation technique to overcome the sparsity and irregularity of physiological data.

Linear dynamic systems have been previously used to model physiological variables for ICU monitoring [44] and detection of sepsis [45]. Hidden Markov Models (HMMs) were also used to model health trajectories of patients [46], [47]. However, such models cannot easily adapt to irregularly sampled vital-sign data. Additionally, each hidden state in an HMM only depends on the previous state [48]. Another approach for modeling similar data is the kernel-based support vector regression [42].

One of the most popular techniques for time series modeling within the clinical domain is Gaussian Process Regression (GPR). GPR is based on a non-parametric stochastic process that offers a probabilistic approach for time-series modeling by providing confidence intervals for estimated values at unobserved time instances. A comprehensive introduction to GPR can be found in [49]. Previous studies illustrate the robustness of the single-task GPR [42], [50], [51] in modeling a single physiological time-series variable. Others focus on multi-task GPR [43], [52], [53], which learns similarities across several time-series data data and models them simultaneously. The use of GPR relies heavily on the choice of the kernel that encodes prior knowledge of any nonlinear time-series dynamics that might be hypothesized to exist in the data.

Most recently, neural processes, a class of neural latent variable models, were also introduced as a probabilistic regression approach [54], which generalizes GPR through the use of generative models from deep learning.

Modeling the physiological trajectory of patients has become increasingly popular for further use in classification [43] or clustering applications [46], [50].

2) Feature Scaling: Empirical studies show that the performance of predictive models relies on the statistical normalisation of the input space [55]. Z-score normalisation with zero mean and unit standard deviation is a widely used tool in feature scaling of numeric clinical variables [13], [56]–[59]. Min-max normalisation performs a scaling of the feature values to lie within a range, such as [0,1] in [11]. A rigorous comparison of the different normalisation techniques in the context of clinical deterioration does not exist. The current practice is to choose the normalisation technique based on its effect on the performance of the respective classifier. This presents an opportunity for future research.

C. Representation Learning

Learning a suitable lower-dimensional *embedding* or *representation* of a high-dimensional input space is a fundamental component of ML research [32]. The embedding can represent a medical concept [60] or summarize a patient's hospital visit [61]. It often performs better than the raw input for learning subsequent tasks [62]–[64]. We now provide an overview of the techniques for obtaining embeddings in related medical applications: (i) standard dimensionality reduction techniques,

TABLE I
OVERVIEW OF FEATURE REPRESENTATION TECHNIQUES ADOPTED IN RELATED WORKS USING A VARIETY OF PREDICTOR VARIABLES: VITAL SIGNS (VS), LABORATORY TESTS (LT), DEMOGRAPHIC INFORMATION (DI), DIAGNOSTIC CODES (DC), INTERVENTIONS (INT) SUCH AS PROCEDURES AND MEDICATIONS, AND FREE TEXT (TEX)

		Predictor Variables						Feature Representation		
Ref	Year	VS	LT	DI	DC	INT	TEX	Hand-crafted Features	Time-series Modelling	Representation Learning
[32]	2008	✓						✓		
[33]	2010	✓		✓		✓		✓		
[34]	2012	✓		✓				✓		
[35]	2012	✓		✓				✓		
[57]	2012	✓							✓	
[36]	2013	✓							✓	
[37]	2013		✓						✓	✓
[39]	2014			✓			✓	✓		
[41]	2015	✓	✓					✓		
[42]	2015	✓					✓	✓	✓	
[40]	2015				✓	✓				✓
[14]	2016	✓	✓	✓				✓		
[44]	2016		✓	✓	✓	✓	✓			✓
[69]	2016			✓	✓	✓				✓
[68]	2016		✓		✓	✓				✓
[13]	2017	✓	✓	✓	✓	✓		✓		
[47]	2017	✓	✓	✓	✓	✓			✓	
[48]	2017			✓	✓	✓				✓
[88]	2017				✓	✓		✓		
[45]	2018	✓		✓				✓		
[49]	2018	✓	✓	✓	✓	✓	✓			✓
[11]	2019	✓	✓	✓	✓	✓	✓	✓		✓

(ii) distributed representations used in language modelling, (iii) using embedding layers as part of a larger model, or (iv) through the latent space of autoencoders and their variants. Such compact representations are then further used as inputs for classification or clustering purposes (covered in Section V).

1) Standard Dimensionality Reduction Techniques: One of the most popular statistical dimensionality reduction techniques is principal component analysis (PCA) [127] . PCA transforms a set of possibly correlated variables to a set of linearly uncorrelated components. It has been used to extract features for various clinical applications [40], [65], [66], such as for the detection of hypotensive episodes [34], mortality prediction across stroke patients [67], or prediction of hospital readmission [68]. The main limitation of PCA is that it extracts linear features that may not well represent non-linear relationships present in complex clinical data [32]. Another popular technique is independent component analysis (ICA) [69], [70], which transforms the variables to a set of independent components.

2) Distributed Concept Representations: Patient records may contain discrete categorical codes, such as diagnosis, medication, or treatment codes. Several studies [71]–[73] propose learning from such variables using embedding techniques derived from the distributional hypothesis in semantic modeling. The distributional hypothesis states that words that appear in similar contexts in large samples of language data are semantically similar [74]. The skip-gram algorithm learns the co-occurrence of information inside a context window of a fixed size [128]. It has been used to convert medical codes to dense representations in [60], [71], [75]. Similar to skip-gram, the Global Vectors (GloVe) algorithm was also used to learn the global co-occurrence matrix of medical codes [76].

3) Embedding Layers: Embedding layers can also be integrated as part of a larger model to transform high-dimensional features into a lower-dimensional space. The embedding can consist of a simple linear transformation [77], [78] or as a fully-connected (deep) network [11], [73], [77]. One study projected the input into a higher-dimensional space using a convolutional layer [72].

4) Autoencoders and Their Variants: An autoencoder is a neural network architecture that is often used for dimensionality reduction or feature extraction [79]. It first transforms the input space to a (typically noise-free) lower-dimensional representation using an encoder, and then reconstructs the input from this compact representation. The sparse autoencoder (SAE) enforces a sparsity constraint on the learned representation, and it has been used to learn latent representations of clinical data [61], [80]. The denoising autoencoder (DAE) reconstructs the input from a partially corrupted version of the input. The stacked DAE, which consists of several autoencoders that are initially pre-trained independently then connected in one network, has also been used for clinical applications [56], [70], [81], [82]. Another popular variant of autoencoders is the variational autoencoder [83], which is a generative model that learns a probabilistic latent space, unlike the previously mentioned discriminative autoencoders.

In Table I, we summarize the feature extraction techniques in related outcome prediction studies. In terms of variable selection, we observe that free clinical text is the least-used data type. That may be due to the limited availability of datasets and the complexity of processing free text, such as due to the prevalence of abbreviations. However, recent works have have looked into pre-training well known natural language

TABLE II
EXAMPLES OF TECHNIQUES USED FOR OUTCOME PREDICTION IN RELATED WORKS

Model	Outcome	References	Year
Gaussian process classifier	Cardiac arrhythmia	[53]	2008
Logistic regression	Hemodynamic instability 2 hours in advance	[32]	2008
	Gout vs. acute leukaemia	[37]	2013
	Mortality on the same or next day	[91]	2013
Novelty detection	ICU readmission	[36]	2013
Support vector machine	Cardiac arrest within 72 hours	[35], [41]	2012, 2015
	Mortality within 72 hours	[35], [38], [42]	2012, 2014, 2015
Random forest classifier	Diseases within one-year interval	[44]	2016
Support ensemble boosting	Paediatric transfer to ICU	[45]	2018
Multi-layer perceptron	ICU transfer and cardiac arrest	[14]	2016
	Hypotensive episodes	[33], [34]	2010, 2012
	Ventricular tachycardia	[12]	2016
	In-hospital mortality	[50]	2018
Gated recurrent units	Heart failure	[48]	2017
	Multi-label diagnoses	[40]	2015
Convolutional neural network	Congestive heart failure after 6 months	[43]	2016
	COPD after 6 months	[43]	2016
	Hospital readmission after discharge	[88]	2017
Recurrent neural networks	Mortality	[13]	2017
	Acute kidney injury	[11]	2019
	Diagnosis & medication codes for next visit	[40]	2015
Long short term memory networks	Sepsis at least 4 hours in advance	[47]	2017
	ICU admission, mortality, & cardiac arrest	[22]	2020

processing architectures with clinical text for related tasks, such as BERT [84]. Therefore, we expect textual data to become increasingly popular in future clinical applications as research in natural language processing develops. We also note that representation learning has gained popularity from approximately 2013 on wards, and we expect it to continue to be an active area of research in the near future, since it can also support the development of end-to-end models. The consistent use of hand-crafted features over the years indicates its effectiveness in improving the learning of ML models. Unlike hand-crafted features that are easy to compute, time-series modeling is not as widely used since it requires extensive hyperparameter tuning (e.g., choice of kernel). Both hand-crafted features and time-series modeling limit end-to-end training of the pipeline, since they are usually incorporated as stand-alone intermediate data processing steps. Another interesting trend is that more types of predictor variables are being included in prediction models over time, due to the increased availability of EHR data and computational resources.

V. PREDICTIVE INFERENCE

The extracted features can then used to train an outcome prediction model. The task can be posed either as a classification (Section V-A) or clustering (Section V-B) problem.

A. Outcome Classification Framework

Table II summarizes the different classification models that have been used to predict various clinical outcomes, as presented in recent papers. Most papers compare the performance of their models to those of simple ML techniques, such as regression [58], [78], which have been useful statistical techniques long since before the rise of ML. We also observe a trend in ML model selection over time, where sophisticated deep learning models, such as convolutional neural networks or long short term memory networks, were used most recently. We also note that predictions are often defined within a particular future time-frame, ranging from short-term 48 hours prediction windows [11] to 6 months, in order to frame the problem as a classification task. The varying definitions in the literature of what exactly constitutes an outcome makes it challenging to compare methods directly. Additionally, some studies tend to focus on specific patient subgroups, such as pediatrics [33].

1) Regression Models: Logistic regression is one of the simplest linear classifiers [86] and is often considered as a standard benchmark for sophisticated clinical models [87]. Previous studies used logistic regression to predict hemodynamic instability [35], imminent mortality [88], or the composite outcome of cardiac arrest, unplanned ICU admission, and mortality [19]. However, logistic regression cannot learn non-linear relationships and assumes independence across the input variables.

Decision tree learning involves the stratification of the feature space based on a criterion defined by information theory, such as entropy. One study developed an early warning score based on decision trees, using seven routinely-collected laboratory tests [89], while another constructed an ensemble model with gradient tree boosting and adaptive boosting to predict the likelihood of transfer to pediatric ICU [33]. Despite the high interpretability of the aforementioned studies, they heavily rely on task-specific hand-engineered features and do not learn complex patterns in the data.

2) Kernel Methods: Kernel methods rely on a user-defined kernel function that estimates the 'similarity' between pairs of data [90]. The support vector machine is a popular example of kernel methods. It projects data into a higher-dimensional space and finds the optimal discriminatory hyper-planes between classes [91]. The use of support vector machines heavily relies on the choice of the kernel and regularization, and they have shown strong performance in recent clinical applications [36], [39], [92], [93]. Computing the kernel matrix for all pairs of data may be computationally expensive for large clinical datasets especially when a non-linear kernel is used. Further work must investigate approximation techniques for applications involving large-scale medical data.

3) Deep Learning: Deep learning models are also becoming increasingly popular for outcome prediction tasks [12], [14], [37], [43], [96]. The simplest form of neural networks is the multi-layer perceptron (MLP), which consists of fully-connected perceptrons. The main limitation of the MLP is its inability to account for temporal dependencies. Recurrent neural networks and their variants seek to model temporal behaviour through feedback connections. Both *Long Short Term Memory* (LSTM) networks [43], [97], [98] and *Gated Recurrent Units* (GRU) [71], [77] were constructed to predict (and alert in advance of) clinical outcomes. There is also a growing interest in developing 'end-to-end' architectures that can jointly extract features and perform classification [78], [85], [99]. Although deep learning techniques are typically characterized by strong performance, their decision-making process lacks interpretability.

B. Clustering for Abnormality Detection

Clustering algorithms are unsupervised learning techniques that group data based on similarity measures. With the increased availability of EHR databases, such techniques have been useful for patient phenotyping and disease subtyping [80], [100]. As for detecting deterioration prior to adverse events, most existing works adopt the novelty detection framework using vital signs, also known as 'one-class classification.' A full review of novelty detection methods has been created in [101].

Such approaches involve creating a 'dictionary' or cluster of healthy patients and computing a similarity metric for new patients. Kernel density estimators are non-parametric methods that can estimate the underlying probability distribution from multi-variate data. Early works demonstrated the use of unconditional probability density function, one-class support vector machine, and Gaussian mixture models to assess the patient's status using routine measurements of vital signs with respect to a 'normal' distribution [102], [103]. Another study used a weighted sum of Gaussian kernels to estimate the distribution of the vital signs of 'normal' patients, and the departure from normality was quantified using a novelty score based on likelihood [104]. Later works focused on assessing the patient based on a time-series representation of the vital-sign data. Some considered clustering of GPR-derived latent representations to model vital-sign data trajectories, and compute the similarity of a new test trajectory based on its local likelihood with respect to the training set [50] or the Kullback-Leibler (KL) divergence [105], [106]. There are other statistical similarity metrics that can be used to compare distributions, such as the Bhattacharyya distance [107]. Most of the aforementioned related clustering works are based on vital-sign data only and involve small-scale datasets.

VI. Performance Evaluation

The performance of supervised outcome prediction models on the *testing set* is evaluated using various statistical methods. Those statistical methods mainly assess the performance of the model in terms of accuracy metrics. In recent years, model interpretability has also become an area of interest as it directly reflects how we translate technologies into clinical practice [108].

A. Performance Metrics

Model discrimination refers to the model's ability in separating classes of interest. In the context of outcome prediction models, we will here refer to patients who experience an adverse outcome as the *positive class*, and those who do not as the *negative class*. Many ML models are trained to compute the probability of the positive class, which is then converted to a binary value by fixing a decision threshold. The predictions are then compared to the true labels and can classified into one of four categories: (1) True Positives (TP): model correctly predicts the positive class, (2) True Negatives (TN): model correctly predicts the negative class, (3) False Positives (FP): model incorrectly predicts the positive class, and (4) False Negatives (FN): model incorrectly predicts the negative class.

Accuracy, which summarizes the proportion of correctly classified samples across all samples, is highly biased when using highly imbalanced datasets. Therefore, other metrics are usually considered. Sensitivity, or the True Positive Rate (TPR), assesses the model's ability to correctly predict the positive class.

$$TPR = \frac{TP}{TP + FN} \tag{2}$$

Specificity, also known as the True Negative Rate (TNR), assesses the model's ability to correctly predict the negative class.

$$TNR = \frac{TN}{TN + FP} \tag{3}$$

The receiving operator characteristic (ROC) curve plots the TPR on the vertical axis and (1-TNR), also known as the False Positive Rate (FPR), on the horizontal axis. The integral under the curve is the Area Under the Receiving Operator Characteristic Curve (AUROC) [109].[2] The AUROC assesses the model's overall diagnostic ability as the decision threshold is varied. An AUROC of 0.5 means that the model is making predictions at random in a two-class setting. One related study mentions that an AUROC higher than 0.8 implies that the model has good diagnostic ability and an AUROC higher than 0.9 means that the model has excellent diagnostic ability [110].

[2] Some studies refer to the AUROC as the 'concordance-statistic' (C-statistic).

Precision, also known as the Positive Predictive Value (PPV), assesses the proportion of correctly predicted positive class across all of the true positive class.

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

The Precision-Recall curve, where recall is essentially sensitivity, plots the TPR on the horizontal axis and the Precision on the vertical axis and integrates the area under the curve. The integral under the curve is the Area under Precision-Recall Curve (AUPRC). Unlike the AUROC, the AUPRC and PPV are highly sensitive to class imbalance. Outcome prediction models are generally characterized with low AUPRC and PPV values [111]. Due to low PPV values, such systems should be considered as risk stratifiers rather than predictors [34].

There are other commonly assessed metrics, such as the F1-score [96], [112] and the likelihood ratio [113]. Some studies also report the false positives to true positives ratio [11] and the inverse of the PPV known as the work-up-to-detection ratio [58], [114]. The efficiency curve [89], [115] is a qualitative summary that plots the number of positives generated at different decision thresholds against the sensitivity of the model. This tool is essential to evaluate the trade-off between the total number of positives and the number of false positives.

B. Interpretability

Despite the good performance of recently introduced ML models, interpretability remains to be a challenge for their clinical utility [108]. There are various definitions of interpretability in existing literature and they refer to several distinct ideas [116], [117]. Most of these ideas pertaining to the clinical domain revolve around trustworthiness of the results and transparency of the model. In the context of this review, we summarize the efforts of outcome prediction models that considered interpretability as a key component of model assessment.

Mimic learning assumes that shallow models, such as linear models, are interpretable. It aims to identify the features that are potentially relevant to the prediction. It involves first training a deep learning model for a specific clinical task. It then trains a shallow model, such as gradient boosting trees, to mimic the behaviour of the deep learning model [82], [118]. The local interpretable model-agnostic explanation (LIME) [119] generates a local explanation of the model behaviour using a shallow model. It has been even used to explain ML models for the prediction of in-hospital mortality [120]. However, it has also been argued that linear models, rule-based models, and decision trees are not intrinsically interpretable [116]. Other post-hoc interpretability techniques such as *saliency maps* rely on qualitative visual interpretations commonly used in computer vision applications.

It is often argued that deep learning models compromise interpretability for high accuracy [121]. Thus, there have been recent breakthroughs in developing inherently interpretable deep learning models instead of performing post-hoc interpretation [122]. For instance, attention mechanisms are incorporated within deep learning models and assign normalised weights to a set of features. The weights indicate the feature importance for the prediction of a future diagnosis [72], [76], [99] or high risk vascular diseases [112]. Other works impose non-negativity [61] or sparsity [80] constraints on the learned embedding space of medical data.

VII. Moving Forward

The prediction of clinical outcomes is essential to detect deterioration in a timely manner and to ease burden off clinical staff. The development of the ML pipelines and their subsequent performance can also be improved by accounting for a few considerations.

A. Noisy Outcome Labels

To train outcome prediction models, outcome labels are currently being defined based on the occurrence of discrete clinical events. However, such labels may be noisy or inaccurate since EHRs only reflect parts of the hospital experience. For example, while a patient may experience cardiac arrest, the patient may be on terminal care pathways with ‘do not resuscitate orders,’ and such information may not be present in the available dataset.

Additionally, outcome labels are defined based on a specific time-window, where the features are associated with a positive outcome label only if they are within N hours to an outcome. This creates a strict cut-off where data collected prior to this N-hours window is not associated with a future outcome. Realistically speaking, deterioration is likely to develop gradually over time, yet this is the state-of-the-art approach in developing outcome prediction models within clinical practice. Future work should consider both classification and time-to-event analysis, where the latter focuses on predicting the time until the occurrence of an outcome, rather than just predicting a binary label [123].

B. Personalized Predictive Models

Most of the outcome prediction models are developed and evaluated population-wide and recent improvements show marginal improvements. As more data is collected per patient, we hypothesize that the predictive power of such models could improve by developing patient-specific models, that account for individual-, disease-, and organizational-based factors [129]. On an individual-level, factors may include demographics, lifestyle, coexisting medical conditions, or genetic information. Disease-related factors may include degree of severity, medications and therapy, rate of progression, interventions, surgeries, and procedures. Organizational-factors may include type of hospital, time of the day, staff ratio, or staff training. This also motivates the advancement of internet of things in healthcare to enhance the collection of integrated data, and would certainly allow us to move forward towards ‘precision medicine.’

Additionally, in the development of machine learning and deep learning models, it is assumed that the data samples are independent and identically distributed (i.i.d.) random sets. However, this may not be the case in practice, since some data samples may belong to the same patient and spatio-temporal patterns may be indicative of deterioration prior to an outcome.

C. General Learning Models

Deep neural networks are powerful processing techniques. However, most of the state-of-the-art models seek to learn how to predict a specific outcome or a particular task, which can generally be referred to as 'narrow AI.' While some of the motivation behind using representation learning has been to learn general patient representations as inputs for downstream predictive tasks, more work needs to be done into developing generalized models that can automatically learn from heterogeneous EHR data to perform diverse tasks simultaneously, such as disease diagnosis and patient prognosis.

Additionally, end-to-end models have shown recent success in applications such as speech recognition and natural language processing [124]–[126], since they can bypass intermediate data processing steps that are typically present in traditional ML pipelines. In the context of clinical outcome prediction models, this requires major improvements in the collection and curation of EHR data across several dimensions, especially completeness, complexity, and accuracy. To overcome the challenge of manually designing ML pipelines, some works have suggested frameworks to automatically optimize the configuration of the pipeline, such as `AutoPrognosis` [123]. Future works should further investigate the extension of end-to-end training for EHR data to improve efficiency and minimize biases.

While recently developed ML models perform well within retrospective studies, validating their success in practice requires prospective analysis. The progress of the field relies on increased multidisciplinary collaborations between ML research scientists and clinicians. While it will take time for both parties to speak the same language, we hope that this review would demystify the overall ML pipeline and summarize the assumptions and techniques of the state-of-the-art.

References

[1] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nat. Biomed. Eng.*, vol. 2, no. 10, pp. 719–731, Oct. 2018.

[2] N. Afzal *et al.*, "Natural language processing of clinical notes for identification of critical limb ischemia," *Int. J. Med. Informat.*, vol. 111, pp. 83–89, 2018.

[3] V. Gulshan *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Amer. Med. Assoc.*, vol. 316, no. 22, pp. 2402–2410, 2016.

[4] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[5] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[6] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *IEEE Access*, vol. 5, pp. 8869–8879, 2017.

[7] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[8] K. H. An *et al.*, "Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc., INTERSPEECH*, 2018, pp. 1913–1917.

[9] H. Nishi *et al.*, "Deep learning-derived high-level neuroimaging features predict clinical outcomes for large vessel occlusion," *Stroke*, vol. 51, pp. 1484–1492, 2020.

[10] J. G. Lee *et al.*, "Deep learning in medical imaging: General overview," *Korean J. Radiol.*, vol. 18, no. 4, pp. 570–584, 2017.

[11] N. Tomašev *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, pp. 116–119, 2019.

[12] H. Lee, S.-Y. Shin, M. Seo, G.-B. Nam, and S. Joo, "Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks," *Sci. Rep.*, vol. 6, 2016, Art. no. 32390.

[13] M. Aczon *et al.*, "Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks," 2017, *arXiv:1701.06675*.

[14] S. B. Hu, D. J. L. Wong, A. Correa, N. Li, and J. C. Deng, "Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model," *PLoS ONE*, vol. 11, no. 8, pp. 1–12, 2016.

[15] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*.

[16] M. E. Smith *et al.*, "Early warning system scores for clinical deterioration in hospitalized patients: A systematic review," *Ann. Amer. Thoracic Soc.*, vol. 11, no. 9, pp. 1454–1465, 2014.

[17] L. L. Leape *et al.*, "The nature of adverse events in hospitalized patients," *New England J. Med.*, vol. 324, no. 6, pp. 377–384, 1991.

[18] D. Jones, I. Mitchell, K. Hillman, and D. Story, "Defining clinical deterioration," *Resuscitation*, vol. 84, no. 8, pp. 1029–1034, 2013.

[19] M. M. Churpek, T. C. Yuen, and D. P. Edelson, "Predicting clinical deterioration in the hospital: The impact of outcome selection," *Resuscitation*, vol. 84, no. 5, pp. 564–568, 2013.

[20] G. Neale, M. Woloshynowych, and C. Vincent, "Exploring the causes of adverse events in NHS hospital practice," *J. Roy. Soc. Med.*, vol. 94, no. 7, pp. 322–330, 2001.

[21] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmark of deep learning models on large healthcare MIMIC datasets," 2017, *arXiv:1710.08531*.

[22] F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton, "Deep interpretable early warning system for the detection of clinical deterioration," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 437–446, Feb. 2019.

[23] "National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS," Royal College of Physicians, London, U.K., Tech. Rep., 2017.

[24] M. Makar, M. Ghassemi, D. M. Cutler, and Z. Obermeyer, "Short-term mortality prediction for elderly patients using medicare claims data," *Int. J. Mach. Learn. Comput.*, vol. 5, no. 3, pp. 192–197, 2015.

[25] K. G. Moons *et al.*, "Transparent Reporting of a multivariable prediction model for individual prognosis or disagnosis (TRIPOD): Explanantion and elaboration," *Ann. Internal Med.*, vol. 162, no. 1, pp. W1–W74, 2015.

[26] A. E. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Sci. Data*, vol. 3, 2016, Art. no. 160035.

[27] G. Hripcsak and D. J. Albers, "Next-generation phenotyping of electronic health records," *J. Amer. Med. Informat. Assoc.*, vol. 20, no. 1, pp. 117–21, 2013.

[28] J. D. Patrick, D. H. M. Nguyen, Y. Wang, and M. Li, "A knowledge discovery and reuse pipeline for information extraction in clinical notes," *J. Amer. Med. Informat. Assoc.*, vol. 18, no. 5, pp. 574–579, 2011.

[29] W. R. Hogan and M. M. Wagner, "Accuracy of data in computer-based patient records," *J. Amer. Med. Informat. Assoc.*, vol. 4, no. 5, pp. 342–355, 1997.

[30] E. M. Mirkes, T. J. Coats, J. Levesley, and A. N. Gorban, "Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes," *Comput. Biol. Med.*, vol. 75, pp. 203–216, 2016.

[31] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *J. Amer. Med. Informat. Assoc.*, vol. 20, pp. 144–151, 2012.

[32] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[33] J. Rubin *et al.*, "An ensemble boosting model for predicting transfer to the pediatric intensive care unit," *Int. J. Med. Informat.*, vol. 112, pp. 15–20, 2018.

[34] J. Lee and R. G. Mark, "An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care," *BioMed. Eng. OnLine*, vol. 9, no. 1, p. 62, 2010, Art. no. 62.

[35] H. Cao, L. Eshelman, N. Chbat, L. Nielsen, B. Gross, and M. Saeed, "Predicting ICU hemodynamic instability using continuous multiparameter trends," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2008, pp. 3803–3806.

[36] C. E. Kennedy, N. Aoki, M. Mariscalco, and J. P. Turley, "Using time series analysis to predict cardiac arrest in a PICU," *Pediatric Crit. Care Med.*, vol. 16, no. 9, pp. 332–329, 2015.

[37] R. Donald *et al.*, "Early warning of EUSIG-defined hypotensive events using a Bayesian artificial neural network article," *Acta Neurochirurgica Supplementum*, vol. 114, pp. 87–91, 2012.

[38] M. Ghassemi *et al.*, "Unfolding physiological state: Mortality modelling in intensive care units," *Bone*, vol. 23, no. 1, pp. 1–7, 2014.

[39] M. E. H. Ong *et al.*, "Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score," *Crit. Care*, vol. 16, no. 3, 2012, Art. no. R108.

[40] G. Skolidis, R. H. Clayton, and G. Sanguinetti, "Automatic classification of arrhythmic beats using Gaussian processes," *Comput. Cardiol.*, vol. 35, pp. 921–924, 2008.

[41] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering—A decade review," *Inf. Syst.*, vol. 53, pp. 16–38, 2015.

[42] L. Clifton, D. A. Clifton, M. A. Pimentel, P. J. Watkinson, and L. Tarassenko, "Gaussian processes for personalized e-health monitoring with wearable sensors," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 193–197, Jan. 2013.

[43] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask Gaussian process RNN classifier," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1174–1182.

[44] J. A. Quinn, C. K. Williams, and N. McIntosh, "Factorial switching linear dynamical systems applied to physiological condition monitoring," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1537–1551, Sep. 2009.

[45] I. Stanculescu, C. K. I. Williams, and Y. Freer, "A hierarchical switching linear dynamical system applied to the detection of sepsis in neonatal condition monitoring," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, 2014, pp. 752–761.

[46] S. Ghassempour, F. Girosi, and A. Maeder, "Clustering multivariate time series using hidden Markov models," *Int. J. Environ. Res. Public Health*, vol. 11, no. 3, pp. 2741–2763, 2014.

[47] L. W. H. Lehman, S. Nemati, R. P. Adams, and R. G. Mark, "Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 5939–5942.

[48] Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," 2015, *arXiv:1506.00019*.

[49] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

[50] M. A. Pimentel, D. A. Clifton, and L. Tarassenko, "Gaussian process clustering for the functional characterisation of vital-sign trajectories," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, 2013, pp. 1–6.

[51] G. W. Colopy, S. J. Roberts, and D. A. Clifton, "Gaussian processes for personalized interpretable volatility metrics in the step-down ward," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 3, pp. 949–959, May 2019.

[52] R. Dürichen, M. A. F. Pimentel, L. Clifton, A. Schweikard, and D. A. Clifton, "Multitask Gaussian processes for multivariate physiological time-series analysis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 314–322, Jan. 2015.

[53] M. Ghassemi, T. Naumann, T. Brennan, D. A. Clifton, and P. Szolovits, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 446–453.

[54] M. Garnelo *et al.*, "Neural processes," 2018, *arXiv:1807.01622*.

[55] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagationfor classification," *Int. J. Comput. Theory Eng.*, vol. 3, no. 1, pp. 89–93, 2011.

[56] P. Schwab, G. C. Scebba, J. Zhang, M. Delai, and W. Karlen, "Beat by beat: Classifying cardiac arrhythmias with recurrent neural networks," *Comput. Cardiology (CinC)*, 2017, pp. 1–4.

[57] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical intervention prediction and understanding using deep networks," 2017, *arXiv:1705.08498*.

[58] A. Rajkomar *et al.*, "Scalable and accurate deep learning for electronic health records," *Nat. Digi. Med.*, vol. 1, no. 1, pp. 1–10, 2018.

[59] N. Razavian, J. Marcus, and D. Sontag, "Multi-task prediction of disease onsets from longitudinal lab tests," in *Proc. 1st Mach. Learn. Healthcare Conf.*, 2016, pp. 1–27.

[60] Y. Choi, C. Y.-i. C. Ms, and D. Sontag, "Learning low-dimensional representations of medical concepts," *AMIA Joint Summits Transl. Sci. Proc.*, vol. 2016, pp. 41–50, 2016.

[61] E. Choi *et al.*, "Multi-layer representation learning for medical concepts," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1495–1504.

[62] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.*, vol. 10, no. 1, pp. 1–40, 2009.

[63] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[64] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, pp. 1–36, 2019.

[65] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–65, 2006.

[66] H. Wimmer and L. Powell, "Principle component analysis for feature reduction and data preprocessing in data science," in *Proc. Conf. Inf. Syst. Appl. Res.*, 2016, pp. 1–6.

[67] S. Cheon, J. Kim, and J. Lim, "The use of deep learning to predict stroke patient mortality," *Int. J. Environmental Res. Public Health*, vol. 16, no. 11, 2019, Art. no. 1876.

[68] D. Krompaß, C. Esteban, V. Tresp, M. Sedlmayr, and T. Ganslandt, "Exploiting latent embeddings of nominal clinical data for predicting hospital readmission," *KI - Künstliche Intelligenz*, vol. 29, no. 2, pp. 153–159, 2015.

[69] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–30, 2000.

[70] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, 2016, Art. no. 26094.

[71] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *J. Amer. Med. Informat. Assoc.*, vol. 24, no. 2, pp. 361–370, 2017.

[72] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4091–4098.

[73] C. Esteban, O. Staeck, Y. Yang, and V. Tresp, "Predicting clinical events by combining static and dynamic information using recurrent neural networks," in *Proc. IEEE Int. Conf. Healthcare Informat.*, 2016, pp. 93–101.

[74] M. Sahlgren, "The distributional hypothesis," *Italian J. Linguistics*, vol. 20, no. 1, pp. 33–53, 2008.

[75] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Healthcare Conf.*, 2015, pp. 301–318.

[76] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 787–795.

[77] C. Esteban, D. Schmidt, D. Krompaß, and V. Tresp, "Predicting sequences of clinical events by using a personalized temporal latent embedding model," in *Proc. IEEE Int. Conf. Healthcare Informat.*, 2015, pp. 130–139.

[78] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.

[79] A. C. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[80] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PLoS ONE*, vol. 8, no. 6, 2013, Art. no. e66341.

[81] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3480, 2010.

[82] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Distilling knowledge from deep networks with applications to healthcare domain," 2015, *arXiv:1512.03542*.

[83] C. Doersch, "Tutorial on variational autoencoders," 2016, *arXiv:1606.05908*.

[84] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," in *Proc. Clin. Natural Lang. Process. Workshop*, 2019, *arXiv:1904.03323*.

[85] P. Nguyen, T. Tran, N. Wickramasinghe, and S. Venkatesh, "Deepr: A convolutional net for medical records," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 22–30, Jan. 2017.

[86] D. W. Hosme, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[87] E. Christodoulou *et al.*, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clinical Epidemiol.*, vol. 110, pp. 12–22, 2019.

[88] E. Loekito *et al.*, "Common laboratory tests predict imminent death in ward patients," *Resuscitation*, vol. 84, no. 3, pp. 280–285, 2013.

[89] S. W. Jarvis *et al.*, "Development and validation of a decision tree early warning score based on routine laboratory test results for the discrimination of hospital mortality in emergency medical admissions," *Resuscitation*, vol. 84, no. 11, pp. 1494–1499, 2013.

[90] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Ann. Statist.*, vol. 36, no. 3, pp. 1171–1220, 2008.

[91] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discovery*, vol. 2, pp. 121–167, 1998.

[92] A. Daemen *et al.*, "Improved modeling of clinical data with kernel methods," *Artif. Intell. Med.*, vol. 54, no. 2, pp. 103–114, 2012.

[93] Y. Chen *et al.*, "Applying active learning to high-throughput phenotyping algorithms for electronic health records data," *J. Amer. Med. Informat. Assoc.*, vol. 20, pp. e253–e259, 2013.

[94] J. M. Kwon, Y. Lee, Y. Lee, S. Lee, H. Park, and J. Park, "Validation of deep-learning-based triage and acuity score using a large national dataset," *PLoS ONE*, vol. 13, 2018, Art. no. e0205836.

[95] Y. Cheng, F. Wang, P. Zhang, and J. Hu, "Risk prediction with electronic health records: A deep learning approach," in *Proc. Proc. SIAM Int. Conf. Data Mining. Soc. Ind. Appl. Math.*, 2016, pp. 432–440.

[96] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018.

[97] S. Hochreiter and J. U. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[98] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–18.

[99] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 1903–1911.

[100] P. Schulam, F. Wigley, and S. Saria, "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery," in *Proc. Nat. Conf. Artif. Intell.*, 2015, pp. 2956–2964.

[101] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Process.*, vol. 99, pp. 215–249, 2014.

[102] L. Tarassenko, A. Hann, and D. Young, "Integrated monitoring and analysis for early warning of patient deterioration," *Brit. J. Anaesthesia*, vol. 97, no. 1, pp. 64–68, 2006.

[103] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko, "Identification of patient deterioration in vital-sign data using one-class support vector machines," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, 2011, pp. 125–131.

[104] M. A. Pimentel, D. A. Clifton, L. Clifton, P. J. Watkinson, and L. Tarassenko, "Modelling physiological deterioration in post-operative patient vital-sign data," *Med. Biol. Eng. Comput.*, vol. 51, pp. 869–77, 2013.

[105] T. Zhu *et al.*, "Patient-specific physiological monitoring and prediction using structured Gaussian processes," *IEEE Access*, vol. 7, pp. 58 094–58 103, 2019.

[106] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, 2007.

[107] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Meth. Soc.*, vol. 7, no. 4, pp. 401–406, 1946.

[108] M. A. Ahmad, A. Teredesai, and C. Eckert, "Interpretable machine learning in healthcare," in *Proc. IEEE Int. Conf. Healthcare Informat.*, 2018, pp. 559–560.

[109] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[110] G. B. Smith, D. R. Prytherch, P. E. Schmidt, and P. I. Featherstone, "Review and performance evaluation of aggregate weighted 'track and trigger' systems," *Resuscitation*, vol. 77, no. 2, pp. 170–179, 2008.

[111] P. J. Watkinson, M. A. Pimentel, D. A. Clifton, and L. Tarassenko, "Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data," *Resuscitation*, vol. 129, pp. 55–60, 2018.

[112] Y. J. Kim, Y.-G. Lee, J. W. Kim, J. J. Park, B. Ryu, and J.-W. Ha, "High risk prediction from electronic medical records via deep attention networks," 2017, *arXiv:1712.00010*.

[113] M. Hoikka, T. Silfvast, and T. I. Ala-Kokko, "Does the prehospital national early warning score predict the short-term mortality of unselected emergency patients?" *Scand. J. Trauma, Resuscitation Emergency Medicine*, vol. 26, no. 1, 2018, Art. no. 48.

[114] S. Romero-Brufau, J. M. Huddleston, G. J. Escobar, and M. Liebow, "Why the C-statistic is not informative to evaluate early warning scores and what metrics to use," *Crit. Care*, vol. 19, no. 1, pp. 19–24, 2015.

[115] D. R. Prytherch, G. B. Smith, P. E. Schmidt, and P. I. Featherstone, "ViEWS-towards a national early warning score for detecting adult inpatient deterioration," *Resuscitation*, vol. 81, no. 8, pp. 932–937, 2010.

[116] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 35–43, 2018.

[117] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[118] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for ICU outcome prediction," in *Proc. Annu. Symp. AMIA Symp.*, 2016, vol. 2016, pp. 371–380.

[119] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 97–101.

[120] S. Nanayakkara *et al.*, "Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study," *PLoS Med.*, vol. 15, no. 11, pp. 1–16, 2018.

[121] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 150–158.

[122] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[123] A. M. Alaa and M. Van Der Schaar, "Autoprognosis: Automated clinical prognostic modeling via Bayesian optimization with structured kernel learning," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 139–148.

[124] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," 2019, Art. no. 1018.

[125] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12, pp. 2493–2537, 2011.

[126] T. Glasmachers, "Limits of end-to-end learning," 2017, *arXiv:1704.08305*.

[127] L. I. Smith, "A tutorial on principal components analysis," Tech. Rep., 2002.

[128] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546v1*.

[129] D. Jones, I. Mitchell, K. Hillman, and D. Story, "Defining clinical deterioration," *Resuscitation*, vol. 84, no. 8, pp. 1029–1034, 2013.