Machine Learning for Clinical Outcome Prediction

Farah Shamout, Tingting Zhu, and David A. Clifton

***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.**

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, ranging from natural language processing to computer vision applications [2], [3]. Several ML frameworks have been proposed to predict clinical outcomes within a certain time period in the future, such as cardiac arrest, mortality, or intensive care unit (ICU) admission [4], [5], [6], [7].

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 [7].

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 III) while the second task learns from patterns in the data to produce the predicted label (Section IV). Such models are usually assessed based on clinical utility and interpretability (Section V).

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 [8], minimizes $\mathcal{L}(y, \hat{y}|\Theta)$ in an iterative manner based on the examples present in the dataset.

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

Patient monitoring tools, such as early warning systems [9], 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 [10]. Subsequent studies focused on more discrete clinical events, such as severe sepsis, unexpected cardiac arrest, ICU admission or mortality [11], [12], 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 [13]. 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 [14]. 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

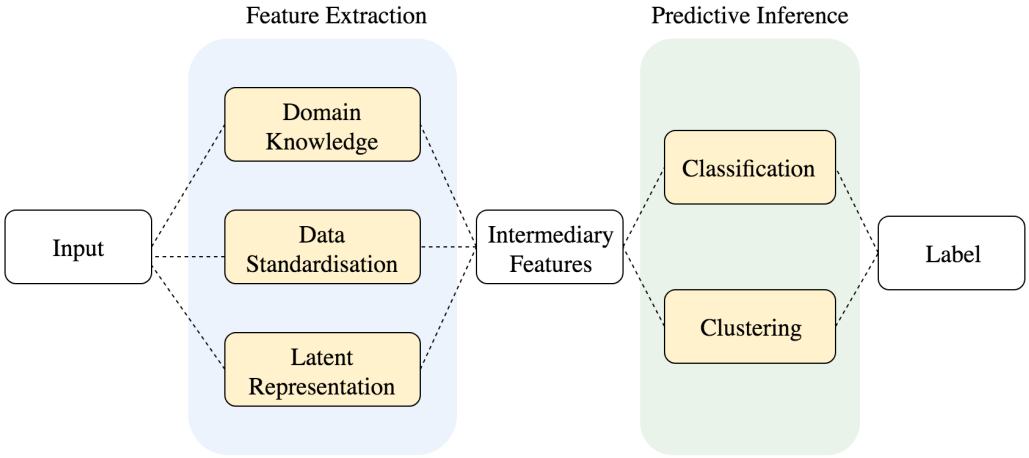

Fig. 1. General ML pipeline that maps an input to a label. The two main steps of the pipeline are (i) extraction of an intermediary feature space and (ii) label prediction using a classification or clustering algorithm.

observations is collected [15], as in clinical early warning systems [16].

II. Electronic Health Records

Various types of data can be used to develop outcome prediction models, such as imaging, speech, or claims data [17]. 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 *training set* and a *testing set*[1], either by a random or a nonrandom split based on location or time. According to the *Transparent Reporting of a multivariable 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 [18]. 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 [19]. 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 [20]. 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 [21]. 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" [22]. 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 [20], [23]. 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" [22]. 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 [20]. 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 [24].

Finally, it is important to verify whether the data was recorded within a reasonable period of time [24]. For example, the recorded collection time of vital signs may precede the time of admission. Although this aspect of

[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.

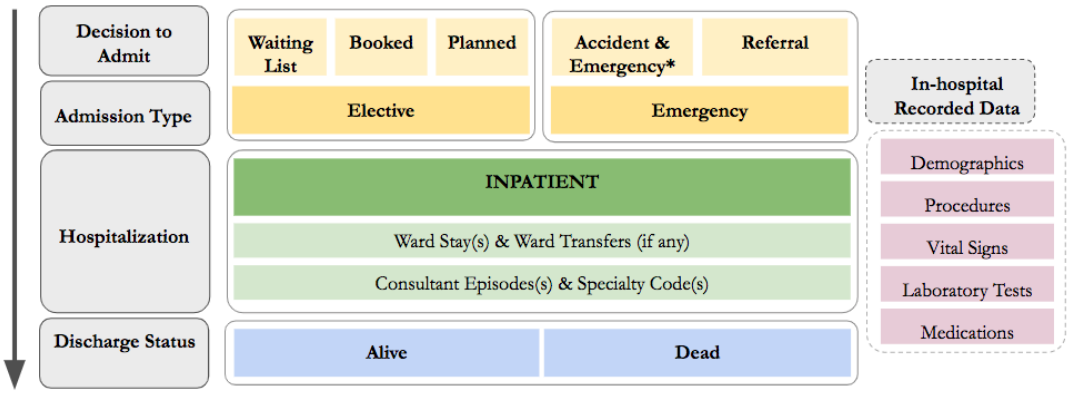

Fig. 2. Visualization of the patient flow in a hospital: Patient is either admitted as an elective or emergency admission, monitored in ward stay(s) during consultant episode(s). Patient may transfer from one ward to another, or may change the consultant during the in-hospital stay. ** Accident & Emergency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

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.

III. Feature Extraction

The performance of clinical predictive models relies on the feature representation of the data, as in other domains [44]. As reported in related works, feature extraction generally involves at least one of domain-expertise for hand-crafted features (Section III-A), data standardization (Section III-B), or representation learning (Section III-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 [38], [26], shock index [25], [34], [38], mean arterial pressure [27], [38], oxygen delivery index [34], absolute successive difference of heart rate, estimated cardiac output, slope of fitted regression lines, or slope projections [25]. 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 [25], [45], [32].

Previous research also computed time series features from waveform data [28], [46], [26], [5]. Those features can be categorized into four types: data adaptive, non-data adaptive, model-based and data-dictated approaches [47]. 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 [5], [26].

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 [29], [40]. 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 [48] and detection of sepsis [49]. Hidden Markov Models (HMMs) were also used to model health trajectories of patients [31], [50]. 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 [51]. Another approach for modeling similar data is the kernel-based support vector regression [29].

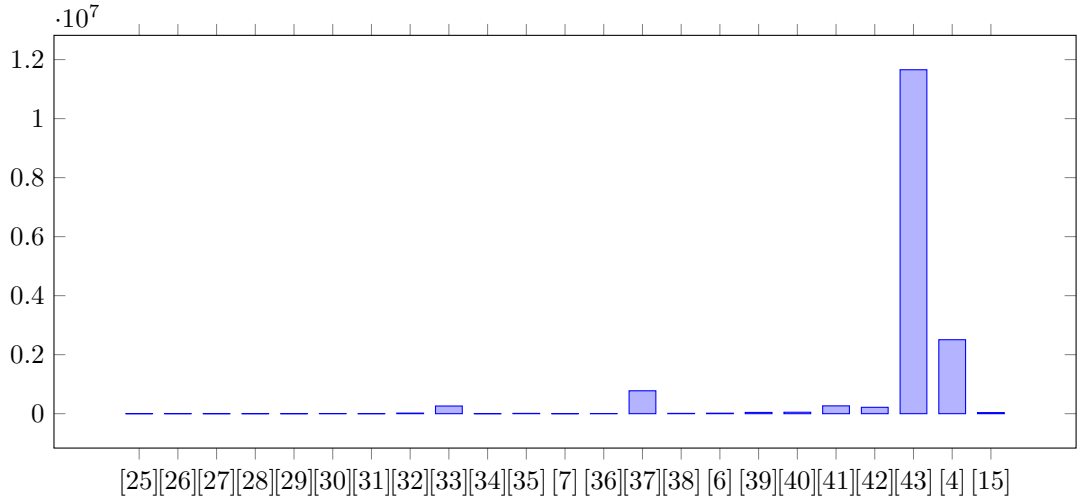

Fig. 3. Dataset sizes reported in the literature in ascending order from left (2008) to right (2019). The vertical axis represents the dataset size, in terms of the number of patient admissions, and the horizontal axis represents the reference number.

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 different characteristics, and then prepare the data for further feature extraction.

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 [52]. Previous studies illustrate the robustness of the single-task GPR [29], [53], [54] in modeling a single physiological time-series variable. Others focus on multi-task GPR [55], [40], [35], 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 [56], 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 [40] or clustering applications [31], [53].

2) Feature Scaling: Empirical studies show that the performance of predictive models relies on the statistical normalisation of the input space [57]. Z-score normalisation with zero mean and unit standard deviation is a widely used tool in feature scaling of numeric clinical variables [58], [59], [42], [6], [60]. Min-max normalisation performs a scaling of the feature values to lie within a range, such as [0,1] in [4]. 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 [44]. The embedding can represent a medical concept [61] or summarize a patient's hospital visit [62]. It often performs better than the raw input for learning subsequent tasks [63], [64], [65]. We now provide an overview of the techniques for obtaining embeddings in related medical applications: (i) standard dimensionality reduction techniques, (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 IV).

1) Standard Dimensionality Reduction Techniques: One of the most popular statistical dimensionality reduction techniques is principal component analysis (PCA) [66]. 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 [67], [46], [68], such as for the detection of hypotensive episodes [26], mortality prediction across stroke patients [69], or prediction of hospital readmission [70]. The main limitation of PCA is that it extracts linear features that may not well represent non-linear relationships present in complex clinical data [44]. Another popular technique is independent component analysis (ICA) [71], [37], 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 [41], [39], [72] 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 [73]. The skip-gram algorithm learns the co-occurrence of information inside a context window of a fixed size [74]. It has been used to convert medical codes to dense representations in [33], [61], [41]. Similar to skip-gram, the Global Vectors (GloVe) algorithm was also used to learn the global co-occurrence matrix of medical codes [75].

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 [76], [77] or as a fully-connected (deep) network [4], [76], [72]. One study projected the input into a higher-dimensional space using a convolutional layer [39].

4) Autoencoders and their variants: An autoencoder is a neural network architecture that is often used for dimensionality reduction or feature extraction [78]. 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 [30], [62]. 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 [79], [37], [58], [80]. Another popular variant of autoencoders is the variational autoencoder [81], 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 input. That may be due to the limited availability of datasets. We also note that representation learning has gained popularity from approximately 2013 and on wards, and we expect it to continue to be an active area of research in the near future. The consistent use of hand-crafted features over the years indicates its effectiveness in training ML models. Additionally, time-series modeling may not be widely used as it requires hyperparameter tuning and high computational resources. It also limits end-to-end training of the pipeline, since some operations cannot be differentiated for gradient descent.

IV. Predictive Inference

The extracted features can then used to train an outcome prediction model. The task can be posed either as a classification (Section IV-A) or clustering (Section IV-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 [42], [77], which have been useful statistical techniques long since before the rise of ML. We also observe that predictions are often defined within a particular future time-frame, ranging from short-term 48 hours prediction windows [4] to 6 months. 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 [38].

1) Regression Models: Logistic regression is one of the simplest linear classifiers [83] and is often considered as a standard benchmark for sophisticated clinical models [84]. Previous studies used logistic regression to predict hemodynamic instability [25], imminent mortality [85], or the composite outcome of cardiac arrest, unplanned ICU admission, and mortality [12]. 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 [86], while another constructed an ensemble model with gradient tree boosting and adaptive boosting to predict the likelihood of transfer to pediatric ICU [38]. 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 [87]. 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 [88]. The

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
[25]	2008	✓						✓		
[26]	2010	✓		✓		✓		✓		
[27]	2012	✓		✓				✓		
[28]	2012	✓		✓				✓		
[50]	2012	✓							✓	
[29]	2013	✓							✓	
[30]	2013		✓						✓	✓
[32]	2014			✓			✓	✓		
[34]	2015	✓	✓					✓		
[35]	2015	✓					✓	✓	✓	
[33]	2015				✓	✓				✓
[7]	2016	✓	✓	✓				✓		
[37]	2016		✓	✓	✓	✓	✓			✓
[62]	2016			✓	✓	✓				✓
[61]	2016		✓		✓	✓				✓
[6]	2017	✓	✓	✓	✓	✓		✓		
[40]	2017	✓	✓	✓	✓	✓			✓	
[41]	2017			✓	✓	✓				✓
[82]	2017				✓	✓		✓		
[38]	2018	✓		✓				✓		
[42]	2018	✓	✓	✓	✓	✓	✓			✓
[4]	2019	✓	✓	✓	✓	✓	✓	✓		✓

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 [28], [89], [90], [34]. 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 [91], [7], [5], [27], [40]. 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 [92], [93], [40] and *Gated Recurrent Units* (GRU) [76], [41] 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 [77], [82], [94]. 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. Within the context of predicting adverse clinical outcomes, this can involve creating a 'dictionary' or cluster of healthy patients and computing a similarity metric for a new patient [45], [53], [95]. Popular similarity metrics are the Kullback-Leibler (KL) divergence [96] and the Mahalanobis distance [45]. Clustering analysis has also been useful for patient phenotyping [30]. The concept of creating patient dictionaries is a subset of novelty detection. An example of such approaches is 'one-class classification' [97], [48].

V. 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 [98].

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

TABLE II

Overview of classifiers used for outcome prediction in related works.

Model	Outcome	References
Novelty detection	ICU readmission	[29]
Logistic regression	Hemodynamic instability 2 hours in advance	[25]
	Gout vs. acute leukaemia	[30]
	Mortality on the same or next day	[85]
Support vector machine	Cardiac arrest within 72 hours	[28], [34]
	Mortality within 72 hours	[28], [35], [31]
Random forest classifier	Diseases within one-year interval	[37]
Support ensemble boosting	Paediatric transfer to ICU	[38]
Gaussian process classier	Cardiac arrhythmia	[46]
Multi-layer perceptron	ICU transfer and cardiac arrest	[7]
	Hypotensive episodes	[26], [27]
	Ventricular tachycardia	[5]
	In-hospital mortality	[43]
Convolutional neural network	Congestive heart failure after 6 months	[36]
	COPD after 6 months	[36]
	Hospital readmission after discharge	[82]
Recurrent neural networks	Mortality	[6]
	Acute kidney injury	[4]
	Diagnosis & medication codes for next visit	[33]
Gated recurrent units	Heart failure	[41]
	Multi-label diagnoses	[33]
Long short term memory networks	Sepsis at least 4 hours in advance	[40]

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) [99].[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. An AUROC higher than 0.8 implies that the model has good diagnostic ability. An AUROC higher than 0.9 means that the model has excellent diagnostic ability [100].

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 [101]. Due to low PPV values, such systems should be considered as risk stratifiers rather than predictors [26].

There are other commonly assessed metrics, such as the F1-score [102], [91] and the likelihood ratio [103]. Some studies also report the false positives to true positives ratio [4] and the inverse of the PPV known as the work-up-to-detection ratio [104], [42]. The efficiency curve [105], [86] 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.

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

B. Interpretability

Despite the good performance of recently introduced ML models, interpretability remains to be a challenge for their clinical utility [98]. There are various definitions of interpretability in existing literature and they refer to several distinct ideas [106], [107]. 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 [80], [108]. The local interpretable model-agnostic explanation (LIME) [109] 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 [110]. However, it has also been argued that linear models, rule-based models, and decision trees are not intrinsically interpretable [106]. 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 [111]. Thus, there have been recent breakthroughs in developing inherently interpretable deep learning models instead of performing post-hoc interpretation [112]. 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 [94], [39], [75] or high risk vascular diseases [102]. Other works impose non-negativity [62] or sparsity [30] constraints on the learned embedding space of medical data.

VI. 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 time-to-event analysis, which focuses on predicting the time until the occurrence of an outcome, rather than predicting a binary label.

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 [113]. 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 in order to perform a variety of predictive tasks, more work needs to be done into developing generalized models that can automatically learn from heterogeneous EHR data to perform diverse tasks.

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] Kun Hsing Yu, Andrew L. Beam, and Isaac S. Kohane. Artificial intelligence in healthcare, 2018.

[2] Naveed Afzal, Vishnu Priya, Sunghwan Sohn, Hongfang Liu, Rajeev Chaudhry, Christopher G Scott, Iftikhar J Kullo, and Adelaide M Arruda-olson. International Journal of Medical Informatics Natural language processing of clinical notes for identi fi cation of critical limb ischemia. *International Journal of Medical Informatics*, 111(September 2017):83–89, 2018.

[3] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA : the journal of the American Medical Association*, 316(22):2402–2410, 2019.

[4] Nenad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O. Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, Geraint Rees, Chris Laing, Clifton R. Baker, Kelly Peterson, Ruth Reeves, Demis Hassabis, Dominic King, Mustafa Suleyman, Trevor Back, Christopher Nielson, Joseph R. Ledsam, and Shakir Mohamed. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572(7767):116–119, 2019.

[5] Hyojeong Lee, Soo-Yong Shin, Myeongsook Seo, Gi-Byoung Nam, and Segyeong Joo. Prediction of Ventricular Tachycardia One Hour before Occurrence Using Artificial Neural Networks. *Scientific Reports*, 6(August):32390, 2016.

[6] M Aczon, D Ledbetter, L Ho, A Gunny, A Flynn, J Williams, and R Wetzel. Dynamic Mortality Risk Predictions in Pediatric Critical Care Using Recurrent Neural Networks. *arXiv*, pages 1–18, 2017.

[7] Scott B. Hu, Deborah J L Wong, Aditi Correa, Ning Li, and Jane C. Deng. Prediction of clinical deterioration in hospitalized adult patients with hematologic malignancies using a neural network model. *PLoS ONE*, 11(8):1–12, 2016.

[8] Sebastian Ruder. An overview of gradient descent optimization algorithms. 2016.

[9] M. E.Beth Smith, Joseph C. Chiovaro, Maya O'Neil, Devan Kansagara, Ana R. Quiñones, Michele Freeman, Makalapua L. Motu'apuaka, and Christopher G. Slatore. Early warning system scores for clinical deterioration in hospitalized patients: A systematic review. *Annals of the American Thoracic Society*, 11(9):1454–1465, 2014.

[10] Lucian L Leape, Troyen A Brennan, Nan Laird, Ann G Lawthers, Russel Localio, Benjamin A Barnes, Leisi Herbert, Joseph P Newhouse, Paul C Weiler, and Howard Hiatt. The Nature of Adverse Events in Hospitalized Patients: Results of the Harvard MEdical Practice Study II. *The New England Journal of Medicine*, 324(6):377–384, 1991.

[11] Daryl Jones, Imogen Mitchell, Ken Hillman, and David Story. Defining clinical deterioration. *Resuscitation*, 84(8):1029–1034, 2013.

[12] Matthew M. Churpek, Trevor C. Yuen, and Dana P. Edelson. Predicting clinical deterioration in the hospital: The impact of outcome selection. *Resuscitation*, 84(5):564–568, 2013.

[13] G Neale, M Woloshynowych, and C Vincent. Exploring the causes of adverse events in NHS hospital practice. *Journal of the Royal Society of Medicine*, 94(7):322–30, 2001.

[14] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of Deep Learning Models on Large Healthcare MIMIC Datasets. 2017.

[15] Farah E. Shamout, Tingting Zhu, Pulkit Sharma, Peter J. Watkinson, and David A. Clifton. Deep Interpretable Early Warning System for the Detection of Clinical Deterioration. *IEEE Journal of Biomedical and Health Informatics*, 2019.

[16] Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Technical report, 2017.

[17] Maggie Makar, Marzyeh Ghassemi, David M. Cutler, and Ziad Obermeyer. Short-Term Mortality Prediction for Elderly Patients Using Medicare Claims Data. *International Journal of Machine Learning and Computing*, 2015.

[18] Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyenberg, Andrew J. Vickers, David Ransohoff, and Gary S. Collins. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Disagnosis (TRIPOD): Explanantion and Elaboration. *Annals of Internal Medicine*, 162(1):W1–W74, 2015.

[19] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 2016.

[20] George Hripcsak and David J Albers. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 20(1):117–21, 2013.

[21] Jon D Patrick, Dung H M Nguyen, Yefeng Wang, and Min Li. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):574–579, 2011.

[22] William R Hogan and Michael M Wagner. Accuracy of data in computer-based patient records. *Journal of the American Medical Informatics Association*, 4(5):342–355, 1997.

[23] 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. *Computers in Biology and Medicine*, 75:203–216, 2016.

[24] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association : JAMIA*, 20:144–151, 2012.

[25] Hanqing Cao, Larry Eshelman, Nicolas Chbat, Larry Nielsen, Brian Gross, and Mohammed Saeed. Predicting ICU hemodynamic instability using continuous multiparameter trends. In *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, volume 2008, pages 3803–6, 2008.

[26] Joon Lee and Roger G Mark. An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *BioMedical Engineering OnLine*, 9(1):62, 2010.

[27] Rob Donald, Tim Howells, Ian Piper, I. Chambers, G. Citerio, P. Enblad, B. Gregson, K. Kiening, J. Mattern, P. Nilsson, A. Ragauskas, Juan Sahuquillo, R. Sinnot, and A. Stell. Early Warning of EUSIG-Defined Hypotensive Events Using a Bayesian Artificial Neural Network Article. *Acta Neurochirurgica Supplementum*, 114(January 2012):87–91, 2012.

[28] Marcus Eng Hock Ong, Christina Hui Lee Ng, Ken Goh, Nan Liu, Zhi Xiong Koh, Nur Shahidah, Tong Tong Zhang, Stephanie Fook-Chong, and Zhiping Lin. 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. *Critical care (London, England)*, 16(3):R108, 2012.

[29] David A Clifton and Marco Pimentel. Gaussian Processes for Personalized e-Health Monitoring With Wearable Sensors Gaussian Processes for Personalized e-Health Monitoring With Wearable Sensors. *IEEE Transactions on Biomedical Engineering*, 60(March 2013):193–197, 2013.

[30] Thomas A. Lasko, Joshua C. Denny, and Mia A. Levy. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE*, 8(6), 2013.

[31] Shima Ghassempour, Federico Girosi, and Anthony Maeder. Clustering multivariate time series using Hidden Markov

Models. *International Journal of Environmental Research and Public Health*, 11(3):2741–2763, 2014.

[32] Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. *Bone*, 23(1):1–7, 2014.

[33] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2015.

[34] Curtis E Kennedy, Noriaki Aoki, Michele Mariscalco, and James P Turley. Using Time Series Analysis to Predict Cardiac Arrest in a PICU. *Pediatric critical care medicine : a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies*, 16(9):332–9, 2015.

[35] Marzyeh Ghassemi, Tristan Naumann, Thomas Brennan, David a Clifton, and Peter Szolovits. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse, Heterogeneous Clinical Data. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 446–453, 2015.

[36] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2016.*, pages 432–440, 2016.

[37] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific reports*, 6(April):26094, 2016.

[38] Jonathan Rubin, Cristhian Potes, Minnan Xu-Wilson, Junzi Dong, Asif Rahman, Hiep Nguyen, and David Moromisato. An ensemble boosting model for predicting transfer to the pediatric intensive care unit. *International Journal of Medical Informatics*, 112(January):15–20, 2018.

[39] Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. Attend and Diagnose: Clinical Time Series Analysis using Attention Models. *arXiv*, 2017.

[40] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[41] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[42] Alvin Rajkomar and Others. Scalable and accurate deep learning for electronic health records. *Nature Digital Medicine*, 1(1):1–10, 2018.

[43] Joon myoung Kwon, Youngnam Lee, Yeha Lee, Seungwoo Lee, Hyunho Park, and Jinsik Park. Validation of deep-learning-based triage and acuity score using a large national dataset. *PLoS ONE*, 2018.

[44] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[45] Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi. Clustering Overly-Specific Features in Electronic Medical Records. *ACM SIGKDD Explorations Newsletter*, 14(1):16, 2012.

[46] G. Skolidis, R. H. Clayton, and G. Sanguinetti. Automatic Classification of Arrhythmic Beats Using Gaussian Processes. *Computers in Cardiology*, 35:921–924, 2008.

[47] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering - A decade review. *Information Systems*, 53(October 2016):16–38, 2015.

[48] John A. Quinn, Christopher K.I. Williams, and Neil McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1537–1551, 2009.

[49] Ioan Stanculescu, Christopher K I Williams, and Yvonne Freer. A Hierarchical Switching Linear Dynamical System Applied to the Detection of Sepsis in Neonatal Condition Monitoring. *UAI'14 Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 752–761, 2014.

[50] Li Wei H. Lehman, Shamim Nemati, Ryan P. Adams, and Roger G. Mark. Discovering shared dynamics in physiological signals: Application to patient monitoring in ICU. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 5939–5942, 2012.

[51] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A Critical Review of Recurrent Neural Networks for Sequence Learning. 2015.

[52] Rasmussen and Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[53] Marco A.F. Pimentel, David A. Clifton, and Lionel Tarassenko. Gaussian process clustering for the functional characterisation of vital-sign trajectories. In *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2013.

[54] Glen Wright Colopy, Stephen J. Roberts, and David A. Clifton. Gaussian Processes for Personalized Interpretable Volatility Metrics in the Step-Down Ward. *IEEE Journal of Biomedical and Health Informatics*, 2019.

[55] Robert Dürichen, Marco A F Pimentel, Lei Clifton, Achim Schweikard, and David A. Clifton. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, 62(1):314–322, 2015.

[56] Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural Processes. 2018.

[57] T Jayalakshmi and A. Santhakumaran. Statistical Normalization and Back Propagationfor Classification. *International Journal of Computer Theory and Engineering*, 3(1):89–93, 2011.

[58] Patrick Schwab, Gaetano Scebba, Jia Zhang, Marco Delai, and Walter Karlen. Beat by Beat: Classifying Cardiac Arrhythmias with Recurrent Neural Networks. *arXiv*, 2017.

[59] Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical Intervention Prediction and Understanding using Deep Networks. *arXiv*, pages 1–16, 2017.

[60] Narges Razavian, Jake Marcus, and David Sontag. Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, pages 1–27, 2016.

[61] Youngduck Choi, Chill Yi-i Chiu Ms, and David Sontag. Learning Low-Dimensional Representations of Medical Concepts. *AMIA Joint Summits on Translational Science proceedings*, pages 41–50, 2016.

[62] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer Representation Learning for Medical Concepts. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pages 1495–1504, 2016.

[63] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, and Pascal Lamblin. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, 2009.

[64] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

[65] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys*, 2019.

[66] Lindsay I Smith. A tutorial on Principal Components Analysis Introduction. *Statistics*, 2002.

[67] Paul Sajda. Machine Learning for Detection and Diagnosis of Disease. *Annual Review of Biomedical Engineering*, 8:537–65, 2006.

[68] Hayden Wimmer and Loreen Powell. Principle Component Analysis for Feature Reduction and Data Preprocessing in

Data Science. In *Proceedings of the Conference on Information Systems Applied Research*, pages 1–6, 2016.

[69] Songhee Cheon, Jungyoon Kim, and Jihye Lim. The Use of Deep Learning to Predict Stroke Patient Mortality. *International journal of environmental research and public health*, 16(11), 2019.

[70] Denis Krompaß, Cristóbal Esteban, Volker Tresp, Martin Sedlmayr, and Thomas Ganslandt. Exploiting Latent Embeddings of Nominal Clinical Data for Predicting Hospital Readmission. *KI - Künstliche Intelligenz*, 29(2):153–159, 2015.

[71] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 2000.

[72] Cristóbal Esteban, Oliver Staeck, Yinchong Yang, and Volker Tresp. Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks. In *IEEE International Conference on Healthcare Informatics (ICHI)*, pages 93–101, 2016.

[73] Magnus Sahlgren. The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–53, 2008.

[74] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality arXiv : 1310 . 4546v1 [cs . CL] 16 Oct 2013. *arXiv preprint arXiv:1310.4546*, 2013.

[75] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. GRAM: Graph-based Attention Model for Healthcare Representation Learning. *arXiv*, pages 1–15, 2016.

[76] Cristóbal Esteban, Danilo Schmidt, Denis Krompaß, and Volker Tresp. Predicting sequences of clinical events by using a personalized temporal latent embedding model. *Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015*, pages 130–139, 2015.

[77] Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In *NIPS Proceedings*, 2016.

[78] Aaron Courville Ian Goodfellow, Yoshua Bengio. Deep Learning Book. *Deep Learning*, 2015.

[79] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre Antoine Manzagol. Stacked denoising autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research*, 2010.

[80] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. 2015.

[81] Carl Doersch. Tutorial on Variational Autoencoders. 2016.

[82] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. Deepr: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics*, 21(1):22–30, 2017.

[83] A. J. Scott, D. W. Hosmer, and S. Lemeshow. Applied Logistic Regression. *Biometrics*, 2006.

[84] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019.

[85] Elsa Loekito, James Bailey, Rinaldo Bellomo, Graeme K. Hart, Colin Hegarty, Peter Davey, Christopher Bain, David Pilcher, and Hans Schneider. Common laboratory tests predict imminent death in ward patients. *Resuscitation*, 84(3):280–285, 2013.

[86] Stuart W. Jarvis, Caroline Kovacs, Tessy Badriyah, Jim Briggs, Mohammed A. Mohammed, Paul Meredith, Paul E. Schmidt, Peter I. Featherstone, David R. Prytherch, and Gary B. Smith. 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*, 84(11):1494–1499, 2013.

[87] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[88] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 1998.

[89] Anneleen Daemen, Dirk Timmerman, Thierry Van den Bosch, Cecilia Bottomley, Emma Kirk, Caroline Van Holsbeke, Lil Valentin, Tom Bourne, and Bart De Moor. Improved modeling of clinical data with kernel methods. *Artificial Intelligence in Medicine*, 54(2):103–114, 2012.

[90] Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association : JAMIA*, 20(e2):253–9, 2013.

[91] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, pages 1–16, 2017.

[92] Sepp Hochreiter and J Urgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[93] Zachary C. Lipton, David C. Kale, Charles Elkan, and Randall Wetzel. Learning to Diagnose with LSTM Recurrent Neural Networks. In *Proceedings of ICLR 2016*, pages 1–18, 2015.

[94] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[95] Tingting Zhu, Glen Wright Colopy, Clare MacEwen, Katherine Niehaus, Yang Yang, Chris W. Pugh, and David A. Clifton. Patient-Specific Physiological Monitoring and Prediction Using Structured Gaussian Processes. *IEEE Access*, 7:58094–58103, 2019.

[96] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 2007.

[97] Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[98] Muhammad Aurangzeb Ahmad, Ankur Teredesai, and Carly Eckert. Interpretable machine learning in healthcare. In *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 2018.

[99] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[100] Gary B. Smith, David R. Prytherch, Paul E. Schmidt, and Peter I. Featherstone. Review and performance evaluation of aggregate weighted 'track and trigger' systems. *Resuscitation*, 77(2):170–179, 2008.

[101] Peter J. Watkinson, Marco A.F. Pimentel, David A. Clifton, and Lionel Tarassenko. Manual centile-based early warning scores derived from statistical distributions of observational vital-sign data. *Resuscitation*, 129(June):55–60, 2018.

[102] You Jin Kim, Yun-Geun Lee, Jeong Whun Kim, Jin Joo Park, Borim Ryu, and Jung-Woo Ha. High Risk Prediction from Electronic Medical Records via Deep Attention Networks. In *NIPS Proceedings*, 2017.

[103] Marko Hoikka, Tom Silfvast, and Tero I. Ala-Kokko. Does the prehospital National Early Warning Score predict the short-term mortality of unselected emergency patients? *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, 2018.

[104] Santiago Romero-Brufau, Jeanne M. Huddleston, Gabriel J. Escobar, and Mark Liebow. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. *Critical Care*, 19(1):19–24, 2015.

[105] David R. Prytherch, Gary B. Smith, Paul E. Schmidt, and Peter I. Featherstone. ViEWS-Towards a national early warning score for detecting adult inpatient deterioration. *Resuscitation*, 81(8):932–937, 2010.

[106] Zachary C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):35–43, 2018.

[107] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017.

[108] Zhengping Che, Sanjay Purushotham, Robinder Khemani, and Yan Liu. Interpretable Deep Models for ICU Outcome Prediction. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2016:371–380, 2016.

[109] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[110] Shane Nanayakkara, Sam Fogarty, Michael Tremeer, Kelvin Ross, Brent Richards, Christoph Bergmeir, Sheng Xu, Dion Stub, Karen Smith, Mark Tacey, Danny Liew, David Pilcher, and David M. Kaye. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. *PLoS Medicine*, 15(11):1–16, 2018.

[111] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–158, 2012.

[112] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

[113] Daryl Jones, Imogen Mitchell, Ken Hillman, and David Story. Defining clinical deterioration. *Resuscitation*, 84(8):1029–1034, 2013.