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, with many successful applications in natural language processing, computer vision applications, and automatic speech recognition [2], [3], [4], [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], [8], [9], [10], as well as clinical outcome prediction to detect deterioration, such as cardiac arrest, mortality, or intensive care unit (ICU) admission [11], [12], [13], [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:

$$f: X \rightarrow y$$

(1)

where $f(.)$ is a function consisting of parameters $\Theta$, $X$ is the input and $y$ is the output. For example, $X$ 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 an ML application, starting from the input $X$, 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 $L(y, \hat{y}|\Theta)$, and an optimization algorithm. The loss function $L(y, \hat{y}|\Theta)$, 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 $L(y, \hat{y}|\Theta)$ 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 [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].

II. 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 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 $\Theta$ of the model. The trained model is then evaluated on the held-out test set using various performance metrics.

$^1$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. 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. 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.
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 workflow 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 [51]. 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 [45], [33], shock index [32], [41], [45], mean arterial pressure [34], [45], oxygen delivery index [41], absolute successive difference of heart rate, estimated cardiac output, slope of fitted regression lines, or slope projections [32]. 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 [32], [52], [39].

Previous research also computed time series features from waveform data [35], [53], [33], [12]. Those features can be categorized into four types: data adaptive, non-data adaptive, model-based and data-dictated approaches [54]. 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], [33].

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 [36], [47]. 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.
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.

Linear dynamic systems have been previously used to model physiological variables for ICU monitoring [55] and detection of sepsis [56]. Hidden Markov Models (HMMs) were also used to model health trajectories of patients [38], [57]. 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 [58]. Another approach for modeling similar data is the kernel-based support vector regression [36].

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 [59]. Previous studies illustrate the robustness of the single-task GPR [36], [60], [61] in modeling a single physiological time-series variable. Others focus on multi-task GPR [62], [47], [42], which learns similarities across several time-series 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 [63], which generalizes GPR through the use of generative models from deep learning.

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.

C. Representation Learning

Learning a suitable lower-dimensional embedding or representation of a high-dimensional input space is a fundamental component of ML research [51]. The embedding can represent a medical concept [68] or summarize a patient’s hospital visit [69]. It often performs better than the raw input for learning subsequent tasks [70], [71], [72]. 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) [7]. 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 [73], [53], [74], such as for the detection of hypotensive episodes [33], mortality prediction across stroke patients [75], or prediction of hospital readmission [76]. The main limitation of PCA is that it extracts linear features that may not well represent non-linear relationships present in complex clinical data [51]. Another popular technique is independent component analysis (ICA) [77], [44], 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 [48], [46], [78] 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 [79]. The skip-gram algorithm learns the co-occurrence of information inside a context window of a fixed size [7]. It has been used to convert medical codes to dense representations in [40], [68], [48]. Similar to skip-gram, the Global Vectors (GloVe) algorithm was also used to learn the global co-occurrence matrix of medical codes [80].

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 [81], [82] or as a fully-connected (deep) network [11], [81], [78]. One study projected the input into a higher-dimensional space using a convolutional layer [46].

4) Autoencoders and their variants: An autoencoder is a neural network architecture that is often used for dimensionality reduction or feature extraction [83]. 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 [37], [69]. 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 [84], [44], [65], [85]. Another popular variant of autoencoders is the variational autoencoder [86], 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 looked into pre-training well known natural language processing architectures with clinical text for related tasks, such as BERT [87]. 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 onwards, 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 standalone 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.

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 [49], [82], 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
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).

| Ref Year | VS | LT | DI | DC | INT | TEX | Hand-crafted Features | Time-series Modelling | Representation Learning |
|----------|----|----|----|----|-----|-----|-----------------------|----------------------|-----------------------|
| [32] 2008 | ✓ | ✓ | | | | ✓ | ✓ | ✓ | ✓ |
| [33] 2010 | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| [34] 2012 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [35] 2012 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [36] 2012 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [37] 2013 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [38] 2014 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [39] 2015 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [40] 2015 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [41] 2016 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [42] 2016 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [43] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [44] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [45] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [46] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [47] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [48] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [49] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [50] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [51] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [52] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [53] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [54] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [55] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [56] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [57] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [58] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [59] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [60] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [61] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [62] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [63] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [64] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [65] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [66] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [67] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [68] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [69] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [70] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [71] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [72] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [73] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [74] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [75] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [76] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [77] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [78] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [79] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [80] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [81] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [82] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [83] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [84] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [85] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [86] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [87] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [88] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [89] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [90] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [91] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [92] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [93] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [94] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [95] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [96] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [97] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [98] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [99] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| [100] 2017 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

1) Regression Models: Logistic regression is one of the simplest linear classifiers [89] and is often considered as a standard benchmark for sophisticated clinical models [90]. Previous studies used logistic regression to predict hemodynamic instability [32], imminent mortality [91], 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 [92], while another constructed an ensemble model with gradient tree boosting and adaptive boosting to predict the likelihood of transfer to pediatric ICU [45]. 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 [93]. 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 [94]. 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 [35], [95], [96], [41]. 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 [97], [14], [12], [34], [47]. 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 [98], [99], [47] and Gated Recurrent Units (GRU) [81], [48] 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 [82], [88], [100]. 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.
### TABLE II

**Examples of techniques used for outcome prediction in related works.**

| Model                          | Outcome                              | References          | Year |
|--------------------------------|--------------------------------------|---------------------|------|
| Gaussian process classifier    | Cardiac arrhythmia                    | [53]                | 2008 |
|                                | Hemodynamic instability 2 hours in advance | [32]                | 2008 |
| Logistic regression            | 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 |
|                                | ICU transfer and cardiac arrest        | [14]                | 2016 |
|                                | Hypotensive episodes                   | [33], [34]          | 2010, 2012 |
|                                | Ventricular tachycardia                | [12]                | 2016 |
|                                | In-hospital mortality                  | [50]                | 2018 |
| Multi-layer perceptron         | Heart failure                          | [48]                | 2017 |
|                                | Multi-label diagnoses                  | [40]                | 2015 |
| Gated recurrent units          | Congestive heart failure after 6 months | [43]                | 2016 |
| Convolutional neural network   | COPD after 6 months                    | [43]                | 2016 |
|                                | Hospital readmission after discharge   | [88]                | 2017 |
|                                | Mortality                              | [13]                | 2017 |
| Recurrent neural networks      | 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 |

With the increased availability of EHR databases, such techniques have been useful for patient phenotyping and disease subtyping [37], [101]. 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 [102].

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 [103], [104]. 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 [105]. 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 [60] or the Kullback-Leibler (KL) divergence [106], [107]. There are other statistical similarity metrics that can be used to compare distributions, such as the Bhattacharyya distance [108]. Most of the aforementioned related clustering works are based on vital-sign data only and involve small-scale datasets.

### 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 [109].

#### 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.

$$TPR = \frac{TP}{TP + FN}$$ (2)

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

$$TNR = \frac{TN}{TN + FP}$$ (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) [110]. 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 [111].

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

$$PPV = \frac{TP}{TP + FP}$$ (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 [112]. Due to low PPV values, such systems should be considered as risk stratifiers rather than predictors [33].

There are other commonly assessed metrics, such as the F1-score [113], [97] and the likelihood ratio [114]. 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 [115], [49]. The efficiency curve [116], [92] 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 [109]. There are various definitions of interpretability in existing literature and they refer to several distinct ideas [117], [118]. 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 [85], [119]. The local interpretable model-agnostic explanation (LIME) [120] 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 [121]. However, it has also been argued that linear models, rule-based models, and decision trees are not intrinsically interpretable [117]. 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 [122]. Thus, there have been recent breakthroughs in developing inherently interpretable deep learning models instead of performing post-hoc interpretation [123]. 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 [100], [46], [80] or high risk vascular diseases [113]. Other works impose non-negativity [69] or sparsity [37] 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 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 [124].

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 [?]. On an individual-level, factors may include demographics, lifestyle, co-existing 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 [125], [126], [127], 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 [124]. 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] 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, Iltihak J Kullo, and Adelaida M Arruda-olson. Natural language processing of clinical notes for identification of critical limb ischemia. International Journal of Medical Informatics, 111:83–89, 2018.
[3] Varun Gulshan, Lilly Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Vemugopalan, Kasumi Widmer, 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. Journal of the American Medical Association, 316(22):2402–2410, 2016.
[4] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning, 2015.
[5] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, and Brian Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, 2012.
[6] Min Chen, Yixue Hao, Kai Huang, Lin Wang, and Lu Wang. Disease Prediction by Machine Learning over Big Data from Healthcare Communities. IEEE Access, 2017.
[7] Athanasios Tsalas, Max A. Little, Patrick E. McSharry, Jennifer Spielman, and Lorraine O. Ramig. Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease. IEEE Transactions on Biomedical Engineering, 2012.
[8] Kwang Hoon An, Myungjeong Kim, Kristin Teplansky, Jordan R. Green, Thomas F. Campbell, Yana Yvanouva, Daragh Heitzman, and Jun Wang. Automatic early detection of amyotrophic lateral sclerosis from intelligible speech using convolutional neural networks. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018.
[9] Hidehisa Nishi, Naoyuki Ito, Akira Ishii, Isao Ono, Takenori Ogura, Tadashi Sunohara, Hideo Chihara, Ryu Fukumitsu, Masakazu Okawa, Norikazu Yamana, Hiroshi Imamura, Nobutake Sadamasa, Taketo Hatano, Ichiro Nakahara, Nobuyuki Sakai, and Susumu Miyamoto. Deep Learning-Derived High-Level Neuroimaging Features Predict Clinical Outcomes for Large Vessel Occlusion. Stroke, 2020.
[10] June Goo Lee, Sanghoon Jun, Young Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: General overview, 2017.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3007816, IEEE Reviews in Biomedical Engineering.
the Future of Patients from the Electronic Health Records. Scientific reports, 6:26904, 2016.

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, 112:15–20, 2018.

Huih Song, Deepweta Raja, Jayaraman J. Thigagaran, and Andreas Spanias. Attend and Diagnose: Clinical Time Series Analysis using Attention Models. arXiv, 2017.

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.

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.

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

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.

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.

Jimeng Sun, Fei Wang, Jianying Hu, and Shahram Edabollahi. Clustering Overly-Specific Features in Electronic Medical Records. ACM SIGKDD Explorations Newsletter, 14(1):16–22, 2012.

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

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

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.

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.

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.

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

Rasmussen and Williams. Gaussian Processes for Machine Learning. The MIT Press, 2006.

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.

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.

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

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh. Neural Processes. In ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models, 2018.

T. Jayakishnum and A. Santhakumar. Statistical Normalization and Back Propagation for Classification. International Journal of Computer Theory and Engineering, 3(1):89–93, 2011.

Patrick Schwab, Gaetano Sc ebba, Jia Zhang, Marco Delai, and Walter Karlen. Beat by Beat: Classifying Cardiac Arrhythmias with Recurrent Neural Networks. arXiv, 2017.

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.

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.

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

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.

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

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.

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

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

Hayden Wimmer and Loreen Powell. Principle Component Analysis for Feature Extraction and Data Preprocessing in Data Science. In Proceedings of the Conference on Information Systems Applied Research, pages 1–6, 2016.

Songhee Cheon, Junyoung 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.

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.

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

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.

Magus Sahlgren. The distributional hypothesis. Italian Journal of Linguistics, 20(1):33–53, 2008.

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.

Cristóbal Esteban, Danilo Schmidt, Denis Krompaß, and Volker Tresp. Predicting sequences of clinical events by using a personalized latent temporal embedding model. Proceedings - 2015 IEEE International Conference on Healthcare Informatics, ICHI 2015, pages 130–139, 2015.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/RBME.2020.3007816, IEEE Reviews in Biomedical Engineering

[82] 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.

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

[84] 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.

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

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

[87] Emily Alsenter, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Nauman, and Matthew B. A. McDermott. Publicly Available Clinical BERT Embeddings. In Clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL 2019, 2019.

[88] Phuoc Nguyen, Tuyen Tran, Nilmini Wickramasinghe, and Svetla Venkatesh. Deep: A Convolutional Net for Medical Records. IEEE Journal of Biomedical and Health Informatics, 21(1):22–30, 2017.

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

[90] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Elsa Loekito, James Bailey, Rinaldo Bellomo, Graeme K. Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung 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.

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

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

[93] Emily Alsenter, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Nauman, and Matthew B. A. McDermott. Publicly Available Clinical BERT Embeddings. In Clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL 2019, 2019.

[94] Phuoc Nguyen, Tuyen Tran, Nilmini Wickramasinghe, and Svetla Venkatesh. Deep: A Convolutional Net for Medical Records. IEEE Journal of Biomedical and Health Informatics, 21(1):22–30, 2017.

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

[96] Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Elsa Loekito, James Bailey, Rinaldo Bellomo, Graeme K. Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung 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.

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

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

[99] Emily Alsenter, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Nauman, and Matthew B. A. McDermott. Publicly Available Clinical BERT Embeddings. In Clinical Natural Language Processing (ClinicalNLP) Workshop at NAACL 2019, 2019.

[100] Phuoc Nguyen, Tuyen Tran, Nilmini Wickramasinghe, and Svetla Venkatesh. Deep: A Convolutional Net for Medical Records. IEEE Journal of Biomedical and Health Informatics, 21(1):22–30, 2017.
classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[121] 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.

[122] 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.

[123] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[124] Ahmed M. Alaa and Mihaela Van Der Schaar. Autoprognosis: Automated clinical prognostic modeling via Bayesian optimization with structured kernel learning. In *35th International Conference on Machine Learning, ICML 2018*, 2018.

[125] Dong Wang, Xiaodong Wang, and Shaohe Lv. An overview of end-to-end automatic speech recognition, 2019.

[126] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 2011.

[127] Tobias Glasmaechers. Limits of end-to-end learning. In *Journal of Machine Learning Research*, 2017.