Deep learning and the electrocardiogram: review of the current state-of-the-art

Sulaiman Somani 1, Adam J. Russak 1,2, Felix Richter 1, Shan Zhao 1,3, Akhil Vaid 1, Fayzan Chaudhry 1,4, Jessica K. De Freitas 1,4, Nidhi Naik 1, Riccardo Miotto 1,4, Girish N. Nadkarni 1,2,5, Jagat Narula 6,7, Edgar Argulian 6,7, and Benjamin S. Glicksberg 1,4*

1The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Pl., New York, NY 10029, USA; 2Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 3Department of Anesthesiology, Perioperative and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 4Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 5The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA; 6Mount Sinai Heart, Icahn School of Medicine at Mount Sinai, New York, NY, USA; and 7Department of Cardiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Received 21 July 2020; editorial decision 20 November 2020; accepted after revision 25 November 2020; online publish-ahead-of-print 10 February 2021

Abstract
In the recent decade, deep learning, a subset of artificial intelligence and machine learning, has been used to identify patterns in big healthcare datasets for disease phenotyping, event predictions, and complex decision making. Public datasets for electrocardiograms (ECGs) have existed since the 1980s and have been used for very specific tasks in cardiology, such as arrhythmia, ischemia, and cardiomyopathy detection. Recently, private institutions have begun curating large ECG databases that are orders of magnitude larger than the public databases for ingestion by deep learning models. These efforts have demonstrated not only improved performance and generalizability in these aforementioned tasks but also application to novel clinical scenarios. This review focuses on orienting the clinician towards fundamental tenets of deep learning, state-of-the-art prior to its use for ECG analysis, and current applications of deep learning on ECGs, as well as their limitations and future areas of improvement.

Keywords
Deep learning • Big data • Artificial intelligence • Electrocardiogram • Cardiovascular medicine

Introduction
The field of deep learning (DL), which has seen a dramatic rise in the past decade, is a form of data-driven modelling that serves to identify patterns in data and/or make predictions. It has made substantial impacts in multiple aspects of modern life, from allowing the human voice to execute commands on smartphones to hyperpersonalizing advertisements.1 In the healthcare space, DL has been leveraged to predict diabetic retinopathy from fundoscopic images,2 diagnose melanoma from pictures of skin lesions,3 and segment the ventricle from a cardiac MRI,4 the latter most of which was recently approved by the FDA, among countless other examples.5–7

Given the vast array of imaging modalities (e.g., CT, MRI, echocardiogram) present in cardiology, DL has also been utilized extensively on cardiovascular data to address key clinical issues.8–10 Though not formally an imaging modality, electrocardiograms (ECG) may be considered different channels (i.e. leads) of one-dimensional images (i.e. signal intensity in volts over time). While other reviews11–16,84 have extensively reported the technical details of various applications of DL or focused on machine learning (ML) applications for ECG analysis, a focus on developing an intuitive understanding for the clinician as well as a clinical perspective on the impact of these advances remains lacking. Additionally, the original research articles showcased in these publications are generally over-representative of small open-source datasets, which are marred with concerns of...
external validity. In addition, there have been many publications recently using DL on ECGs in large, privately curated datasets to solve novel problems, which remain unaddressed by a review.

This review will first aim to establish a foundation of knowledge for DL, with an emphasis on explaining why it is best suited for many ECG-related analyses. Subsequently, we will provide an overview of how ECGs can be represented as a data form for DL, with brief coverage on openly available and private datasets. The Application section will build on this knowledge base and explore original DL research on ECGs that focuses on tasks in five domains: arrhythmias, cardiomyopathies, myocardial ischaemia, valvulopathy, and non-cardiac areas of use. This review will conclude with a recapitulation of the current state, limitations, promising endeavours, and recommendations for future clinical and research practice.

**On artificial intelligence, machine learning, and deep learning**

While a thorough discussion on the details of artificial intelligence (AI) is beyond the scope of this paper, the field and its recent advances will be refreshed for the reader’s benefit. More interested readers are recommended to explore other seminal articles of literature that more exhaustively cover essential knowledge for original research appraisal and endeavours.

Simplistically, AI refers to the idea of a computer model that makes decisions using *a priori* information and improves its performance with experience (i.e. more data). Such clinically related tasks may involve detecting cancerous nodules from CT scans, identifying clusters of disease phenotypes, or optimizing treatment regimens in patients over time.10,18 Given its broad definition, AI is necessarily classified into multiple subsets, notably ML and, more recently, DL, which is a subset of ML. Briefly, both ML and DL seek to use data, rather than a fully empirical set of human-generated rules, to solve a problem. Take, for example, the simple task of converting a temperature from Celsius to Fahrenheit. The empirical approach to solving this problem is to explicitly write a program that takes, as an input, a temperature in °C and converts it into an output, its equivalent temperature in °F, by multiplying the input temperature by 1.8 and adding 32. If we suppose that this conversion equation was not known, one can use linear regression, which is common to both statistics and ML as a simple linear model, to offer the computer an initial guess of a representative equation Temp (F) = m × Temp (C) + b. A starting guess is offered for the unknown parameters (in this case m and b) to represent this information (also called a ‘model’), supply it a table of temperatures in °C (called ‘features’) and corresponding °F (referred to as ‘labels’), provide another set of instructions to fit this data to the underlying equation (i.e. ‘optimization’) by minimizing its prediction error (i.e. ‘loss’ or ‘cost function’), and finally execute this instruction set to continually update the parameters with some logic to ultimately fit this data to the underlying equation (i.e. ‘training’). Though simplistically represented, each parenthetical reference above recognizes a key aspect to some of the most integral and defining components for an AI algorithm that, when tuned appropriately, create novel techniques and entire subspecialties in data-driven AI.

Additionally, while much of probability and statistics is used to mathematically derive and establish the basis for many machine and DL models,10 the priority of statistical models tends to lie in inference and understanding of the dataset’s features and their impact on the outcome of interest with generally parametric models. These models tend to be simpler and not capture non-linearity as well as that of ML or DL models. However, in equivalent and supervised tasks, the simplest AI models prioritize optimizing on outcome prediction instead by engendering more complex model representations.20 The main drawback, however, is that interpretation of the model’s learned parameters becomes significantly harder than that of its counterparts from more statistical frameworks.

Nonetheless, there are nuances between ML and DL that set them apart and are worth discussing. Predominantly, DL separates itself from its parent and predecessor, ML, by the difference in its underlying architecture (which certainly also impacts other facets of the pipeline). Deep learning models are composed of many simple linear models (‘nodes’) arranged in series (each series termed ‘layers’, the number and depth of which contribute eponymously to these models being referred to as ‘deep’) with intervening non-linearities to encourage more complex information representation (Figure 1). This sort of hierarchical structure encourages learning simple representations at each layer that build up to learning complex concepts. In the most intuitive example in image recognition tasks, as work by Olshausen et al. and others has shown,10,18,21 this amounts to each layer (e.g. convolutional, discussed below) in the series learning simple entities (e.g. lines, circles) that build up into more sophisticated representations (e.g. beaks, feathers, eyes).19

By designing models with increased capacity, DL by virtue reduces the need for extensive, manual feature engineering on certain data-sets that are not as natively compatible (e.g. raw ECG waveforms, variable-length sequences) with typical ML models. For example, Narula et al. demonstrate the use of an ML algorithm to distinguish physiologic hypertrophy from hypertrophic cardiomyopathy (HCM) using information such as LV volume and wall strain derived from speckle-tracking echocardiogram data. Simplistically speaking, however, DL, by virtue of its greater capacity to perform cohesive tasks like vision and computer knowledge representation, may obviate the need for such manual labelling by its ability to process raw echocardiogram video data and automatically learn important features (which may or may not include or be derived from the aforementioned features) in order to perform the classification step. It is worth noting that these engineered features may also be used for training DL models, but that DL models operating on such and other structured, tabular data (e.g. patient demographics, lab values) have largely been unable to demonstrate an improvement over comparable statistical or ML frameworks, where data complexity is not high enough to provide deep models with an advantage over well-performing shallow models.23-25

Of critical importance, the need to relinquish *a priori* feature establishment may not be apparent to the reader. For example, with respect to the ECG, frameworks for its interpretation (e.g. rate, rhythm, axis, intervals, ventricles) already exist to classify and localize various cardiac diseases. However, despite the relative robustness of these systems, it would be naive to discount the possible existence of other morphologies indiscernible to the human eye, either locally or as relationships between beats, given the complexity of the cardiac conduction system. In signal processing and imaging, there are many underived features in the raw waveforms and pixels, respectively, which the high-fidelity automatic feature engineering DL offers may
take advantage of. Certainly, such indescribable patterns must exist, and though not fully proven, must explain the encouraging results of Attia et al.\textsuperscript{26} in predicting paroxysmal atrial fibrillation (AF) in patients from a benign, normal sinus rhythm ECG.

However, often the cost of this luxury in capturing complex data representations and improved prediction performance is the aforementioned loss of model interpretability, blanching the technique’s reputation as ‘black-box’. Though methods have been developed to gain more insight into the parameters learned by these models, a notable side effect is overfitting, which is typically caused by having a model with more capacity than relevant information present in the data and required to perform well on the task. This facet permits the model to learn inappropriate aspects about the data, giving the false impression of performing well and causing poor generalizability to other datasets.\textsuperscript{27} Typically, this issue arises when large density models are used to perform prediction on small datasets, which is a slippery slope that can easily occur when trying to improve a model’s performance. Overfitting may also occur in response to biases present in the dataset, notably when limiting data acquisition from a single site or manufacturer or when restricting to a subset of the general population.\textsuperscript{19}

To avoid such pitfalls, it is essential to consider the quality of the dataset, which, if poor enough, may never be overcompensated by any degree of model adjustments.\textsuperscript{28} Best practices dictate use of a training set (usually 60–80\% of a given dataset but will vary based on data availability and outcome prevalence) for the model to learn the parameters for a given network configuration, a validation set (anywhere from 10\% to 20\% of the dataset) to learn the best configuration for the model (i.e. the size and number of layers, type of non-linear activations in the models, etc.), and a test set (usually 10–20\% of the dataset) to report the final model’s performance. Commonly reported metrics to assess model performance include precision or positive predictive value (PPV), recall (sensitivity), specificity, area

**Figure 1** Understanding important layer types. Two common layer types used in deep learning pipelines for image processing are fully connected layers (top), which function simply as many linear regression models with a non-linear activation function that increases the informational capacity of the model. Convolutional layers (bottom) are composed of many ‘kernels’ that learn particular patterns to pick up (small gradient boxes) and scan across an input signal where that pattern may be present. In this example, the kernels from the top to below represent the shape of a R-S wave, a P-wave, and T-P wave segment, and their relative strengths of detection (high: yellow, low: blue) are shown for the input ECG signal (magenta). The resulting signals demonstrate localization of these key kernel patterns that helps the deep learning model learn both the presence and relationship of such features in the input signal. ECGs, electrocardiograms.
under the receiver operator characteristic curve, i.e. AUC-ROC (which reflects the model’s ability to distinguish between different task outcomes), and the F1-statistic (which measures model performance especially in the setting of class imbalance, when one outcome or characteristic is significantly overrepresented in the dataset). While the AUC-ROC, also known as the c-statistic, tends to be the most heavily reported and investigated value, it is important to consider all metrics during appraisal since these metrics are sensitive to the system’s inherent limitations (i.e. class imbalance).29

Finally, we conclude with an overview and intuitive description of the most common DL architectures encountered during the literature retrieval process. By far, convolutional neural networks (CNN) are the most common architecture used for analysing ECGs. At the heart of these networks is the use of the convolution operation, which is a classical technique in signal processing for localizing key features and reducing noise. Convolution refers to the act of taking a small pattern (so-called ‘kernel’) and identifying where in the input that pattern arises (Figure 1), akin to a sliding window. The resulting ‘heat map’ of activity helps to identify where such patterns exist in the image, which can then be used to localize important features, retain global information through successive layers, and remove artefacts deemed unnecessary by the neural network during training. For example, one of the simplest convolutional kernels functions as an edge detector by detecting horizontal or vertical changes in a signal. Serial combinations in parallel and series of these simple edge detectors can allow the CNN to learn how edges combine to form more complex shapes, like the number 9. This generic operation allows sophisticated architectures to be built (i.e. AlexNet,30 GoogleLeNet,31 DenseNet,32 ResNet33) that achieve state-of-the-art performance on standard image competition datasets (e.g. ImageNet34) and serve as inspiration for the development of other models.

While CNNs are well-suited for fixed-length spatial data, recurrent neural networks (RNNs), on the other hand, are designed to address problems that are represented as fixed- or variable-length sequences (i.e. word sentences, signals) and characterize the temporal and spatial relationship of data. The core node in this architecture operates in a loop: for each element in the sequence, it transforms that sequence into an output and hidden representation, the latter of which serves as an additional input for the next element in the sequence. This way, this architecture maintains a memory of the important parts of the sequence and updates the output with that information. Further improvements on this basic design include bi-directional RNNs, gated recurrent units (GRUs), long-short-term memory (LSTM), and attention-transformer networks, which help address the shortcomings of a naïve RNNs and achieve state-of-the-art performance in speech recognition, neural (language) translation, and music generation.19

As is evident, the classical tasks to which these networks are derived do not readily seem amenable to ECG analysis, given the cyclic format (i.e. heartbeats) and its spatial and temporal duality. Therefore, it is worthwhile to discuss the ECG from a data perspective and how it maintains a high level of compatibility with DL to be served to different types of architectures.

**Electrocardiograms as data**

Historically, the heartbeat classification and segment identification of the P-QRS-T were the first data analysis tasks to be performed, and they were achieved from a signal processing approach. These ECGs, originally a time series with a signal intensity, were decomposed into wavelike components with Fourier transformation, Hermite techniques, and wavelet transformations. This may be considered a form of feature extraction since these transformations make important features, such as irregularity in rhythm or rhythm frequency, more discernible for downstream models. Such wavelet-based convolutional techniques have achieved a 93% accuracy on the MIT-BIH arrhythmia database.35 However, ML and DL models have generally achieved better performance with a promise of better generalization and have been favoured since.36,37

In that light, for data-driven model development, it becomes important to identify the best way to represent this signal for the task being solved (Figure 2). The ECG signal may actually be represented in a variety of fashions, each of which may be amenable to a DL pipeline. First, the ECG itself may be subsampled into individual heartbeats of fixed length, which can generate hundreds to thousands of samples per ECG from which features may be derived and used in a more traditional DL network, such as a fully connected neural network. Additionally, it can be sent as a 2D boolean (zeros or ones) image instead of a 1D signal, which is amenable for diagnosing conditions from a fixed-length ECG strip and is highly compatible for use in more traditional image-based CNN architectures. This signal may be one-dimensional or multi-dimensional, depending on the number of leads used, allowing more information to be captured. Finally, the ECG may be represented as a sequence of beats, each linked to the other in time, and treated as a time series that may be analysed by an RNN-type framework.

The type of representation chosen for ECG analysis will ultimately depend on the dataset available. A list of the most commonly available datasets encountered in the literature search is shown in Table 1. The MIT-BIH AF database was the earliest to be released, containing 25 two-lead ECGs, each of which was ~10h long. As other databases followed from the same institution (MIT-BIH), the low number of unique patient ECGs was compensated for by their length, which was subsampled to generate thousands of smaller length ECGs centred around each beat and motivated the research endeavours attempting to perfect beat classification in the early days.38 The Computing in Cardiology Challenge datasets, by introducing much larger datasets, set the stage for novel task definitions (ranging from AF classification, ECG abnormalities, ECG quality, and sleep arousal classification).39 Additionally, though less clean and without extensive annotations for extensive ML or DL tasks, the Mimic database40 gained popularity as well, offering >67,000 ECGs for ICU patients. The past half-decade, however, has also seen a growth in institutional datasets (Table 2), which have surpassed the number of annotated ECGs in these open databases by orders of magnitude. While the number of institutions with published evidence of such databases is few, the retrospective collection of ECG data has allowed more cohort-based questions to be asked, many of which are discussed in the sections below.

**Applications**

This review filtered 31 original research papers to address the applications of DL on ECG identification, starting from a PubMed query
for [(‘deep learning’ OR ‘machine learning’ OR ‘artificial intelligence’) AND (‘electrocardiogram’ OR ‘ECG’ OR ‘ecg’ OR ‘electrocardiograph’) between 1 April 2015 and 15 May 2020 (Figure 3). Since many of the original research articles performed beat classification using the open source datasets and were exhaustively addressed in prior reviews, only papers utilizing >1000 unique ECGs (including both training and test data) were included.

Arrhythmias

Conduction system abnormalities are the most natural cardiac disorders to tackle with ECGs. Motivated by a relatively high adult population prevalence of around 3%, significant work has been devoted to diagnosing AF, the most common arrhythmia, with few ML works on diagnosing other aberrant waveforms (e.g. ventricular tachyarrhythmias). The problem of its identification by ECG has been subject to many research endeavours encompassing all strokes of AI, such as signal processing, ML, and DL, the lattermost of which is detailed in Table 1.

For what may be the most unique but clinically relevant application, Attia et al. used DL to predict paroxysmal AF from a patient’s first clinically benign (i.e. normal sinus rhythm) ECG with the knowledge that they were ultimately diagnosed at least 30 days after this benign ECG with AF. Using a CNN architecture with residual blocks, which allow deeper models to be trained more efficiently, the authors used 454,789 ECGs from 126,526 patients for training and achieved promising performance. While the study design may suffer from heavy selection bias in failing to address patients with ultimately undiagnosed AF and offers no values for a negative predictive value (NPV) despite suggesting the utility of this model as a screening test, the true utility of this work remains in the innovative approach to using ECG data in a novel way and entertaining the possible adjuvant role of DL in conjunction with CHADS2-VASC for recommending anticoagulation in patients with etiologically cryptogenic stroke and, more generally, the risk of stroke secondary to underlying AF.

DL models on ECGs have also been shown to perform at the level of medical professionals. Using only a single ECG lead, Hannun et
al.\textsuperscript{42} curated a dataset composed of 91,232 ECGs from 53,549 patients in an ambulatory setting. At the cost of having a small testing set, the authors benchmarked the model’s encouraging performance by having expert cardiologists manually annotate all 328 test set ECGs. In this case, these experts performed worse compared with the model in detecting all arrhythmias except junctional rhythm and ventricular tachycardia. At a larger scale, Ribeiro\textsuperscript{et al.}\textsuperscript{42,43} demonstrate end-to-end training on the largest ECG database found in this review, comprising of 1,558,415 ECGs from a tele-ECG service in southeast Brazil, to train a CNN with residual connections to diagnose various arrhythmias, such as AVB Type I, RBBB, LBBB, sinus tachycardia and bradycardia, and AF. Somewhat similar to the case with Hannun\textsuperscript{et al.}, the performance of this model, as judged by its PPV, sensitivity, specificity, and AUC, was marginally better when compared with a cohort of medical trainees (residents and medical students).

Extending this multi-classification further, Smith\textsuperscript{et al.}\textsuperscript{44} additionally refined the ECG classification problem in the scope of triaging ECGs in the ED as normal, abnormal, or emergent, subtype by the etiology (e.g. ventricular rhythm emergency vs. significant AV conduction) at a single centre study in MN, USA. They investigated the performance of a pre-trained DL model from an industrial partner (Cardiologs Technologies) against conventional, on-board algorithms that detect these abnormalities on the ECG machines themselves (Mortara/Veritas). For a cohort of 1500 randomly sampled ECGs from that year, their DL model showed greater specificity and accuracy in triaging these ECGs, and, despite suffering only from a marginal loss in sensitivity, demonstrated potential for reducing false alarms on the ECGs by $\frac{1}{2}$. Recently, van de Leur\textsuperscript{et al.}\textsuperscript{45} also developed a model to triage ECGs, but using a dataset orders of magnitude larger and additionally incorporated a gradient-based ‘saliency feature mapping’, which leverages how the output of a model changes with small changes to different regions of the input signal,\textsuperscript{46} to identify important features investigated by the model for different types of presentations. Similar to the models developed by Smith\textsuperscript{et al.}, these models retain high specificity (0.88 to 0.98 for different classes) despite low sensitivity, highlighting their use in rapid escalation of care for those flagged by the model.

Beyond these private datasets, there were three open datasets that met the inclusion criteria for database size: Computing in Cardiology (CINC) 2017, CINC 2015, and CPSC2018 (later merged into the CINC 2020).\textsuperscript{39} In the CINC 2017 competition, which

### Table 1: Publicly available ECG datasets

| Name                        | Year   | Number of leads | Number of ECGs (patients) | ECG length | Labels                                           |
|-----------------------------|--------|-----------------|---------------------------|------------|--------------------------------------------------|
| MIMIC-III                   | 2017   | Variables       | 67,830                    | Variable   | None                                             |
| Computing in Cardiology 2017| 2017   | 1               | 12,186                    | 30s        | Atrial fibrillation classification               |
| Computing in Cardiology 2020| 2020   | 12              | 6,887                     | 30s        | ECG abnormalities\textsuperscript{a}              |
| Computing in Cardiology 2011| 2011   | 12              | 2,000                     | 10s        | ECG quality                                      |
| Computing in Cardiology 2018| 2018   | 1               | 1,985                     | Hours      | Sleep arousal classification                      |
| Computing in Cardiology 2015| 2015   | 2               | 1250                      | 5 min      | False arrhythmia classification                   |
| Chinese Cardiovascular Disease Database | 2010 | 12 | 1,000 | 10s | Beat classification, ECG abnormalities          |
| Computing in Cardiology 2014| 2014   | 1               | 700                       | 10 min     | QRS beat classification                           |
| PTB diagnostic ECG         | 1995   | 16              | 549                       | 2 min      | Diagnosis (MI, CHF, BBB, Arrhythmia, HCM, VHD, normal) |
| SHAREEE                     | 2015   | 3               | 139                       | 24h        | Adverse vascular event prediction                |
| Long-term ST DB            | 2003   | 2               | 86                        | 21–24h     | ST-segment events                                 |
| MIT-BIH supraventricular arrhythmia | 1990 | – | 78 | 30 min | Beat classification, ECG abnormalities\textsuperscript{a} |
| St. Petersburg INCART DB   | 2008   | 12              | 75                        | 30 min     | Beat labelling                                    |
| MIT-BIH arrhythmia DB      | 2001   | 2               | 48                        | 30 min     | Beat classification, ECG abnormalities\textsuperscript{a} |
| MIT-BIH ST change DB       | 1999   | –               | 28                        | Variable   | Beat labelling                                    |
| MIT-BIH atrial fibrillation DB | 1983 | 2 | 25 | 10h | Rhythm annotation (AFib, Aflutter, AV junctional rhythm, N) |
| Sudden cardiac death DB    | 1989   | –               | 23                        | ~24h       | VF                                               |
| MIT-BIH malignant ventricular ectopy DB | 1986 | – | 22 | 30 min | SVT, VF, VFib                                    |
| MIT-BIH normal sinus rhythm DB | 1999 | 18 | 18 | Long-term | Beat labelling                                |
| BIDMC CHF DB               | 1986   | 2               | 15                        | 20h        | Beat classification                               |
| MIT-BIH arrhythmia database P-wave annotations | 2018 | 2 | 12 | 30 min | P-wave labels                                  |

This table lists all publicly available ECG datasets present that were the focal point and source of ECG-based data-driven modelling prior to these new, large, privately curated datasets.

ECGs, electrocardiograms.

\textsuperscript{a}AFib, AVB, LBB, NSR, PAC, PVC, RBB, STD, and STE.
Table 2 Applications of ECGs using deep learning

| Citation | Category          | Prediction task      | Dataset                        | Number of ECGs | Number of patients | Architecture          |
|----------|------------------|----------------------|--------------------------------|----------------|-------------------|-----------------------|
| Parvaneh et al. (2018) | Arrhythmias | Atrial fibrillation | CINC 2017                     | 12 186         | 12 186            | CNN + RNN            |
| Xiong et al. (2018) | Arrhythmias | Arrhythmia           | CINC 2017                     | 12 186         | 12 186            | CNN                  |
| Ribeiro et al. (2019) | Arrhythmias | Arrhythmia           | Telehealth network of Minas Gerais | 15 584 415 | 15 584 415        | Ensemble (CNN, DNN)  |
| Attia et al. (2019) | Arrhythmias | Paroxysmal AF        | Mayo Clinic                   | 649 931        | 180 922           | CNN + GBM            |
| Wang et al. (2019) | Arrhythmias | Arrhythmia           | CINC 2017                     | 193 690        | 193 690           | CNN                  |
| Hannun et al. (2019) | Arrhythmias | Arrhythmia           | iRhythm                       | 91 232         | 53 549            | CNN                  |
| Brisk et al. (2019) | Arrhythmias | Arrhythmia           | CINC 2017                     | 12 186         | 12 186            | CNN                  |
| Wasserlauf et al. (2019) | Arrhythmias | Atrial fibrillation | CINC 2017                     | 7500           | 7500              | CNN + LSTM + SVM     |
| Ivanovic et al. (2019) | Arrhythmias | Atrial fibrillation | Serbia                        | 1097           | 1097              | CNN                  |
| Smith et al. (2019) | Arrhythmias | Arrhythmia           | Cardiolog                     | 1473           | 1473              | CNN                  |
| Mousavi et al. (2020) | Arrhythmias | Arrhythmia           | CINC 2015                     | 1250           | 1250              | CNN (DDDN)           |
| Van De Leur et al. (2020) | Arrhythmias | Arrhythmia triage in the ED | University Medical Center Utrecht | 336 835 | 142 040           | Residual CNN         |
| Oster et al. (2020) | Arrhythmias | Atrial fibrillation | UK Biobank                    | 77 202         | 75 778            | CNN                  |
| Wang et al. (2020) | Arrhythmias | Arrhythmia           | Tiandi competition            | 20 036         | 20 036            | CNN/HMM + GBM        |
| Chen et al. (2020) | Arrhythmias | Arrhythmia           | CPS2018                       | 6877           | 6877              | CNN + GBM            |
| Cai et al. (2020) | Arrhythmias | Atrial fibrillation | Chinese PLA General Hospital, wearable ECGs, CPS2018 | 16 557 11 994 | 36 186 | 36 186 | CNN |
| Tison et al. (2020) | Cardiomyopathy | Heart failure, PAH, MVP | UCSF                          | 36 186         | 36 186            | Ensemble (CNN, DNN)  |
| Kwon et al. (2020) | Cardiomyopathy | Heart failure | Mediplex Sejong Hospital      | 55 163         | 22 765            | CNN                  |
| Attia et al. (2020) | Cardiomyopathy | Heart failure | Mayo Clinic                   | 3 874          | 3 874             | CNN + LSTM + SVM     |
| Attia et al. (2020) | Cardiomyopathy | Heart failure | Mayo Clinic                   | 97 829         | 97 829            | CNN                  |
| Kwon et al. (2020) | Cardiomyopathy | Left ventricular hypertrophy | Sejong General Hospital, Mediplex Sejong Hospital; Korea | 21 286 | 21 286 | CNN |
| Yoon et al. (2020) | Extracardiac | Noise detection      | Ajou University Hospital; Korea | 3000 | 3000 | CNN |
| Ko et al. (2020) | Cardiomyopathy | Hypertrophic cardiomyopathy | Mayo Clinic | 67 001 | 67 001 | CNN + RNN |
| Attia et al. (2020) | Extracardiac | Age, Sex            | Mayo Clinic                   | 774 783        | 774 783           | CNN                  |
| Galloway et al. (2020) | Extracardiac | Hyperkalaemia       | Mayo Clinic                   | 16 384 546 | 44 580 | CNN |
| Lin et al. (2020) | Extracardiac | Hyperkalaemia       | Tri-Service General Hospital; Taiwan | 66 321 | 19 321 | CNN |
| Wang et al. (2020) | Extracardiac | Pre-diabetes        | Beijing, China                | 2914           | 2914              | CNN                  |
| Noseworthy et al. (2020) | Extracardiac | Racial Bias         | Mayo Clinic                   | 97 829         | 97 829            | CNN                  |
| Raghunath et al. (2020) | Extracardiac | Mortality           | Geisinger Hospital System     | 13 385 762 | 42 311 | CNN |
| Kwon et al. (2020) | Extracardiac | Pulmonary hypertension | Sejong General Hospital, Mediplex Sejong Hospital; Korea | 59 844 | 23 376 | CNN |
| Han et al. (2020) | Extracardiac | Nase, Adversarial attack | CINC 2017                 | 12 186         | 12 186            | CNN                  |
| Tadesse et al. (2020) | Ischaemia | Myocardial infarction (STEMI, NSTEMI) | GGH | 21 214 | 21 214 | CNN |
| Kwon et al. (2020) | Extracardiac | Aortic stenosis     | Sejong General Hospital, Mediplex Sejong Hospital; Korea | 39 371 | 39 371 | CNN |
| Kwon et al. (2020) | Valvulopathy | Mitral regurgitation | Sejong General Hospital, Mediplex Sejong Hospital; Korea | 70 709 | 38 241 | CNN + RNN |

This table highlights the 31 applications found during the literature search for ECG analysis, with information about the dataset source, sample size (by unique ECGs and unique patients) present for training and testing, task at hand, and neural network architecture used. Because these studies do not use the same metrics or the same validation protocol to evaluate each model's performance and because the authors firmly believe that comparison of models is tenuous without greater context beyond what this table can provide, these measures have been omitted from being reported in the table. CNN, convolutional neural network; ECGs, electrocardiograms; LSTM, long-short-term memory; RNN, recurrent neural network.
provided contestants with a training set of 8,528 single-lead ECGs for diagnosis of AF vs. NSR, other arrhythmias, and noise, the winner of the competition used an LSTM stacked with an XGBoost classifier (a tree-based ML algorithm). Oster et al. helped externally validate the second-place winner of this competition on 450 four-lead ECGs from the UK Biobank. As expected, the ML algorithm did not generalize well to this novel dataset (F1-score 58.9%); however, a DL model (CNN + LSTM) that was reported after the challenge concluded demonstrated close to a 30% improvement (F1-score 74.1%). In another unique application, a deep CNN trained from AliveCor ECG data, which was the source of the CINC 2017 challenge dataset, was deployed on a single-lead recorder system (KardiaBand, Apple Watch) to continuously monitor for AF in 24 patients. When compared with annotated reports from an insertable cardiac monitor (ICM), the model achieved an encouraging performance (episode sensitivity 97.5% and duration sensitivity 97.7%) on 24 patients, highlighting the utility of DL in creating an inexpensive, non-invasive approach to AF surveillance and management.

For the CPSC2018 challenge, Cai et al. added data from additional sources (hospital, ambulatory ECG monitoring device) and trained a DenseNet-inspired CNN to reach state-of-the-art performance on this multi-centre test set, with an AUC of 0.994 and a sensitivity of 99.1% for the three-label classification task (AF, normal, other arrhythmias). Furthermore, the authors explored the parameter weights of the first convolutional layer of their DNN and found the model to learn, as expected by the premise of DL models, low-level features like peaks, troughs, and upward/downward slopes in the signal, which suggests the model’s efforts to remove baseline shifts and identify key landmarks (i.e. P-waves) in diagnosis.
Ultimately, tackling arrhythmias is the most classical of pattern recognition problems around the ECG. While their diagnosis has been addressed heavily, few works have investigated the direct role of these inpatient management. To our knowledge, only a few have assessed the characteristics of the ECG that are significant for diagnosis. Further work may be undertaken to integrate and assess the role of these DL solutions in direct clinical care, in application towards screening and diagnosis of less prevalent disease states (e.g., congenital long QT syndrome), in more accurately diagnosing arrhythmias, like complex atrioventricular block and wide-complex QRS tachyarrhythmia, which may be difficult to discern clinically, and in providing insights to predicting outcomes after interventional procedures (e.g., AF ablation).

**Valvulopathy**

While ECG lacks sensitivity to diagnose valve disease from traditional clinical frameworks, subtle structural changes in response to long-standing valvular disease may be discovered by a DL model to diagnose these pathologies. Indeed, Kwon et al. demonstrate use of an ensemble model, which combines a CNN classifier operating on raw, 12-lead ECG signals and a fully connected network that incorporates demographic information and numeric ECG-derived features (HR, QT interval, QRS duration, QTc, etc.), for classification of severe aortic stenosis (AS) (<1.5 cm² or mean pressure gradient ≥20 mm Hg, as confirmed by echocardiography). Notably, the authors validated this model on 10,865 patients from a secondary hospital centre, with encouraging AUC of 0.884. The authors also perform a saliency analysis to identify features on the ECG that were most heavily used for AS prediction, identifying the model’s focus on the T-wave in V1–V4, which has been linked with a delayed repolarization from AS-related ventricular hypertrophy. However, the specificity of diagnosing AS relative to other cardiomyopathies was not evaluated in this article, which is an important drawback given that the model may instead be learning to distinguish possible non-specific structural changes secondary to AS, rather than AS itself.

With the same motivation, Kwon et al. replicated the above study on patients with significant MR (valve regurgitant orifice area ≥0.2 cm², regurgitation volume ≥30 mL, regurgitation fraction ≥30%, and MR grade II–IV). In this architecture, they instead opted for a CNN-type network only with raw ECG data as the input and trained on 56,670 ECGs from 24,202 patients in one hospital system. The external validation test set was composed of 10,865 ECGs from another hospital, to which the model had a high sensitivity and NPV at 0.87 on a test set from another hospital centre. The model was benchmarked against cardiologists assessing for LVH using the Sokolov–Lyon criteria and outperformed them on sensitivity, while operating at the same specificity level, by 177%. A saliency analysis revealed that the model focused particularly heavily on the QRS complex during an ‘easy’ diagnosis for LVH, in line with clinical criteria, but concentrating on P wave morphology in V1–V3 and T-wave in I and aVR during more difficult cases, for which clinical criteria are generally absent.

On a different use case, Attia et al. were the first to report the use of DL to predict low EF (<35%) by training a cohort of 35,970 patients on a simple CNN and achieving an AUC of 0.93 on the test set of 52,870 patients. Of significance, the model’s performance remained agnostic to age and sex unlike BNP, which is sensitive to these patient factors and has been proposed as a marker for low EF despite its lower AUC (0.60). A follow-up study included an additional 6,008 patients who had ECGs for non-cardiac clinical indications but were found to have echocardiograms within a year of this ECG indicative of systolic dysfunction. With high AUCs on this external validation set (0.918), these results are encouraging and suggest, in combination with a BNP level >150, the model and lab test can be excellent candidates in screening for systolic dysfunction. Noseworthy et al. further assessed this model’s robustness by investigating the impact of different race and ethnic groups on the model’s performance. Notwithstanding the challenges of binning patient ethnicities into a social construct such as race, the authors demonstrated the model’s invariance in predicting LVEF across various races and ethnicities, retaining AUCs >0.93 for each ethnicity. Additionally, the model demonstrated some inherent ability to predict race from an ECG as well (AUCs 0.76–0.84), though this may be falsely elevated study combining elements from DL and ML, Tison et al. trained a modified CNN architecture (U-Net) on a dataset utilizing publicly available and institutional data to automate ECG segment classification (e.g., P wave, PR segment, QRS complex). Rather than opting for an end-to-end DL architecture, the authors subsequently generated a feature vector from a DL model, fed it into a more classical ML algorithm on a set of 35,466 ECGs to predict the presence of pulmonary hypertension, HCM, amyloid detection, and mitral valve prolapse in patients and achieved encouraging AUROCs, as low as 0.78 for MVP prediction and notably at 0.91 for HCM detection.

For HCM, Ko et al. at the Mayo Clinic report the use of a CNN to train 12-lead ECGs from ~47K patients to diagnose HCM. Remarkably, their models achieved extremely high AUCs of 0.96 on the test set, and though suffering from a relatively low PPV of 31%, concomitantly strong model NPVs and sensitivity suggest its use as a screening tool in clinically suspected patients. A secondary analysis showed that their model responded to a patient who underwent septal myomectomy by lowering its diagnostic probability of HCM from 72% before the operation to 2.5% after. Furthermore, this model retained its high performing AUC in a subgroup of patients with left ventricular hypertrophy (LVH), demonstrating its ability to distinguish true HCM (disease) vs. non-HCM LVH (physiologic).

Further demonstrating the adaptability of DL architectures to different problems, Kwon et al. extend their architecture for AS classification and apply it to detecting LVH. Training their ensemble classifier leveraging both raw ECG waveforms in a CNN and structured patient data from 35,694 ECGs from 12,648 patients, their model achieves respectable AUCs of 0.87 on a test set from another hospital centre. The model was benchmarked against cardiologists assessing for LVH using the Sokolov–Lyon criteria and outperformed them on sensitivity, while operating at the same specificity level, by 177%. A saliency analysis revealed that the model focused particularly heavily on the QRS complex during an ‘easy’ diagnosis for LVH, in line with clinical criteria, but concentrating on P wave morphology in V1–V3 and T-wave in I and aVR during more difficult cases, for which clinical criteria are generally absent.

Cardiomyopathy

With respect to cardiomyopathies, both HCM and LV systolic dysfunction have been the focus of multiple research groups. In a unique
given that the model suffers from severe class imbalances (overrepresentation of non-Hispanic whites) in the training set.

Kwon et al.65,66 greatly extended this demonstration for prediction of reduced EF (EF < 40% and EF < 50% as the primary and secondary study outcomes, respectively) by adding a fully connected neural network trained on both patient-level demographic and ECG-derived data from 13,486 patients to their CNN. The authors report an encouraging model performance (AUC = 0.889 and 0.850 for primary and secondary outcome for external validation set) on an internal and external validation set of ~10,000 ECGs. It is worth noting that logistic regression and random forest (RF), two fundamental ML techniques, both performed only marginally worse relative to the DL model (AUC = 0.853 and 0.847 for LR and RF, respectively, P < 0.001), which may highlight the limited advantage of DL models on tabular data over statistical or ML techniques. By perturbing input values for different features and analysing the impact on the model’s AUC, the authors identified that the most salient features for the DL model were surprisingly in agreement with those found with logistic regression (e.g. HR, T-wave axis, QRS duration, sex, age), suggestive of the more complex and non-linear interplay between these variables (as able to be represented by their architecture) than a simply linearly weighted one. Future directions include utilizing DL with ECG for early identification for understanding or differentiating other cardiomyopathies that are clinically less well understood, such as heart failure with preserved EF (HFpEF) or cardiac amyloidosis.

Ischaemia
Though myocardial ischaemia is one of the most classical areas of cardiovascular research focus, the literature search only revealed one paper that investigated this domain of cardiovascular disease using ECGs and DL. Tadesse et al.62 used a popular framework known as transfer learning, where a model that has been trained on one task (i.e. classifying real-world objects from photos) is partially retained on a completely new, but structurally similar, dataset to solve another task. By transforming the ECGs into the Fourier space (which simply changes the representation of an ECG signal from a signal intensity vs. time to signal intensity vs. wave frequency) and spatially stacking all 12-leads together (to form a 2D-image), they trained a pre-existing, state-of-the-art image classification model, GoogLeNet,31 on an openly available Chinese ECG Challenge dataset,63 and a private curated dataset of ~17,000 ECGs from patients in Southern China with MI (STEMI and NSTEMI), attaining a respectable accuracy of 86% on the private dataset. However, their model performs notably worse with an accuracy of 49% on the Challenge dataset. Furthermore, despite highlighting an interesting technical method for performing DL on the ECG, the authors fail to disclose appropriate sensitivity, specificity, and AUC analyses, leaving room for another research effort to establish precedence for the use of DL on ECGs for patients with ischaemic cardiac disease. Future directions may involve detection of subclinical CAD along, or prior to, the ischaemic heart disease spectrum (e.g. stable angina, unstable angina, etc.).

Extracardiac
Outside the immediate realm of cardiological disease, though certainly not without an impact on the heart, DL has been applied to ECGs in two major areas: identifying electrolyte abnormalities and prognosticating health status. Physiologically, deviations from baseline in either electrolytes or mental illness (i.e. anxiety) have been reported to show short-term and long-term effects on cardiac structure and function, which encourages the study of ECGs to identify the underlying disease state even more.

The sensitivity for diagnosing hyperkalaemia from ECGs, though classically characterized on the ECG by T-wave peaks, PR shortening, QRS prolongation, remains low (34–43%).67 With this in mind, Galloway et al.64,65 conducted a multi-centre study on patients from various Mayo Clinic sites in the US to identify the presence of hyperkalaemia in chronic kidney disease patients using 2- and 4-lead ECGs. Despite low specificity for hyperkalaemia, their model achieved respectable accuracies and sensitivities on these external validation sets, suggesting the role of ECGs for hyperkalaemia screening. Lin et al.68-70 extended this study to predict either hypo- or hyperkalaemia with a single-centre database of 66,321 ECGs to all patients (irrespective of kidney disease) and attained better sensitivity, specificity, and accuracy on their test set when benchmarked against emergency physicians and cardiologists. Unlike the Mayo Clinic, this model retained high specificity (0.92) at the expense of low sensitivity (0.67), which is more akin to its application as a diagnostic tool instead of a screening one. Notably, the authors additionally performed a saliency analysis of the features, which showed a greater focus on the ST segment in those cases of hyperkalaemia that were more difficult to clinically identify (i.e. low sensitivity and high inter-rater variability). In addition to hyper/hypokalaemia, other electrolytes such as magnesium and calcium levels can be assessed here, notably to predict, in real-time, the likelihood of impending arrhythmias like Torsades de Pointes.

Beyond prediction of clinical disease and lab values reflective of disease severity, ECGs, as biometric data points over time, have the potential to capture measures of overall health as well. For example, the epitome of an elderly individual maintaining a prime state of health is captured by that individual having a ‘young heart’. Thus, the idea of an ‘ECG age’ vs. biological age can be inspired and is addressed in another piece by Attia et al.67 which sought to predict patient age using ECG. Subgroup analysis of this study revealed those cases with the largest error in prediction were found to have significantly more instances of systolic dysfunction, hypertension, and CAD, whereas those individuals in which the prediction accuracy was higher (i.e. less error) were found to have fewer cardiovascular incidents at follow-up. Though there are certain implications of overinterpreting this information, since this error could capture both the severity of cardiac disease (e.g. higher age) and also random error in model training, these results encourage the belief that an ECG may be used as a composite biomarker to track general health over time.

In further corroboration of this possible role, Raghunath et al.68 report prediction of 1-year mortality from age, sex, and baseline ECGs using a convolutional framework with a hazard ratio of 9.5 over the two predicted dead/alive groups, further corroborating the prognostic role of an ECG in a patient’s global health. The authors also employ the use of a gradient-based class activation mapping to assess feature importance and note that the model discerned ST-elevations in certain patients as notable contributors to prediction of mortality within 1-year. However, given that these ECGs were retrieved from
a hospital setting, care must be taken not to apply this model, which is prone to a heavy selection bias, on the general population.

Conclusions

When applied to large datasets that contain hidden but valuable relationships, DL has delivered groundbreaking performance. ECGs, laden with information-rich spatial and/or temporal views of the cardiac conduction system, have been amenable to having these hidden associations with cardiovascular pathologies (arrhythmias, cardiomyopathies, valvulopathies, and ischaemia) unravelled, as demonstrated by the original research articles contained within this review. Their role is certainly apparent in future endeavours, as multiple clinical trials have been created to prospectively collect ECG data for not only understanding more about their respective heart disease of interest but also validating existing DL models on these newly collected datasets in the form of a randomized, control trials. Nevertheless, difficulties in data access and model sharing, as well as limited flexibility of pre-existing IT infrastructures, are barriers that must be addressed before these algorithms can be deployed to other hospital systems.

Despite its promise, the shortcomings of these endeavours are readily apparent in the incongruence between model design, model validation, and model interpretation. For example, utilizing DL for feature extraction and performing ML on those features in series is in concept an interesting idea, but certainly carries with it the perils of not abiding by the fundamental hierarchical tenets of DL. Similarly, rigorous practices to ensure an appropriate validation of the model are of crucial importance. Because most datasets thus far have been curated from a single centre, they run the risk of overfitting and generalizing poorly to other hospital systems and other datasets, which not only may have different machines that could have slight variations in the underlying noise that may not be readily filtered for by the model. By extension, adversarial (i.e. simulated noise) training would take advantage of generative adversarial networks (GANs), which are DL models trained to discriminate random generated inputs vs. true dataset inputs and subsequently generate new samples that are more resilient to noise, that have made great strides in improving model performance when additionally trained with subtle but key noisy artefacts. Additionally, no central framework exists for comparing the performance of these various models from one institution with another. An open framework to permit such an exchange of ideas, datasets, and pre-trained model weights is not a trivial task, but can foster an environment for collaboration between what are apparent institutional silos of development.

While every original research article covered in this paper offers encouraging results for the value of DL in interpreting ECGs, only a handful offer insight into the model’s learning representation of the ECG for the respective task. Without explaining what these DL models are sensing on the ECG to perform their specific task in an interpretable way, developers of these tools run a strong risk of souring the clinician, who needs to understand how these models work before entrusting them to augment their practice, to adopting these tools. Methods to open the ‘black box’ of DL have been elucidated in detail elsewhere, offering more than a handful of techniques to evaluate both input feature importance and layer-wise information retention. Such techniques may not only make reduction of these algorithms in clinical practice more palatable but may also offer hypotheses on the pathophysiology of disease that may improve its understanding and possibly reduce the barriers to reduction to practice. Additionally, the trials and tribulations for model selection are not apparent in the methodologies for many papers, which does not instill confidence in the rigor of the model development that is otherwise heavily and rightfully emphasized by the computer science community. The question to be asked is not whether DL can solve a task, but which DL method and why can best tackle the task.

Adherence to these suggested principles of research reporting may create cohesion in the research field by virtue of models and datasets being more amenable to each other, which could in turn foster improved collaboration between research groups. For example, in diagnosing valvulopathies, it is difficult to know, given the current findings in this space, how much of the model is dependent on the effect of the continued altered flow mechanics that create subclinical perturbations in the ECG signal vs. long-standing changes to the heart, which may or may not be specific for that pathology. Performance of classifiers predicting relevant physiological cardiomyopathies or augmenting the original dataset with data from patients with non-valvular cardiomyopathy could help improve the robustness of these original seminal works in DL.

In conclusion, though the emerging literature evaluating the role of DL in ECG analysis has shown great promise and potential, with continued improvement, generalization, refinement, and standardization of methods and data to improve the short-term drawbacks in reduction to clinical practice, DL offers the ability to improve a novel way of diagnosing and managing heart disease. The concurrent development of wearable technologies and accessible platforms for deploying pre-trained DL models offers a unique and scalable opportunity to screen for and intervene early in different cardiovascular disease states.

Conflict of interest: S.S. is a co-founder of and owns equity in Monogram Orthopedics. G.N.N. has received consulting fees from AstraZeneca, Reata, BioVie, and GLG Consulting; has received financial compensation as a scientific board member and advisor to RenalytixAI; and owns equity in RenalytixAI and Pensieve Health as a cofounder. All remaining authors have declared no conflicts of interest.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
2. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.
3. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.
4. Cardio AI - Artery. Arterys. https://arterys.com/cardio-ai/ (16 May 2020, date last accessed).
5. Miotti R, Fang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2018;19:1236–46.
6. Miotti R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6:1–10.
7. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M et al. Scalable and accurate deep learning with electronic health records. npj Digital Med 2018;1:1–10.
8. Johnson KW, Torres Soto J, Glicksberg BS, Shannar K, Miotto R, Ali M et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018;71:2668–79.

9. Johnson KW, Shannar K, Glicksberg BS, Readhead B, Sengupta PP, Björkgren JLM et al. Enabling precision cardiology through multiscale biology and systems medicine. JACC Basic Trans Sci 2017;2:311–27.

10. Shannar K, Johnson KW, Glicksberg BS, Dudley JT, Sengupta PP. Machine learning in cardiac medicine: are we there yet? Heart 2018;104:1156–64.

11. Rim B, Sung N-J, Min S, Hong M. Deep learning in physiological signal data: a survey. Sensors 2020;20:969.

12. Minchale A, Camps J, Lyon A, Rodriguez B. Machine learning in the electrocardiogram. J Electrocardiol 2019;57:561–4.

13. Parvaneh Saman, Jonathan Rubin, Rahman Asif, Brian Conroy, and Saeed Babaezadeh. “Densely Connected Convolutional Networks and Signal Quality Analysis to Detect Atrial Fibrillation Using Short Single-Lead ECG Recordings.” In: 2017 Computing in Cardiology Conference (CinC). Computing in Cardiology. 2017.10.2249/cin17.2017.160.170.

14. Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. Deep learning for healthcare applications based on physiological signals: a review. Comput Methods Programs Biomed 2018;161:1–13.

15. Minchale A, Rodriguez B. Artificial intelligence for the electrocardiogram. Nat Med 2019;25:22–3.

16. Kashou AH, May AM, Noseworthy PA. Artificial intelligence-enabled ECG: a modern lens on an old technology. Curr Rev Repat Med 2020;22:57.

17. Lee SM, Seo JB, Yun J, Cho Y-H, Vogel-Claussen J, Schiebler ML et al. Deep learning applications in chest radiography and computed tomography: current state of the art. J Thorac Imaging 2019;34:75–85.

18. Feng R, Badgley M, Mocco J, Oermann EK. Deep learning guided stroke management: a review of clinical applications. J Neurointervent Surg 2018;10:358–62.

19. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press, 2016.

20. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods 2018;15:233–4.

21. Olthausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 1999;391:607–9.

22. Narula S, Shannar K, Salem Omar AM, Dudley JT, Sengupta PP. Machine-learning algorithms to automate morphological and functional assessments in 2D echocardiography. Am J Cardiol 2016;118:2287–93.

23. Anik SO, Pfister T. TabNet: Attentive Interpretable Tabular Learning. arXiv preprint arXiv:1908.07442. 20 August 2019.

24. Abutbul A, Elidan G, Katir J, El-Yanir Y. DNF-Net: A Neural Architecture for Tabular Data. arXiv preprint arXiv:2006.06465. 11 June 2020.

25. Popov S, Morozov S, Babenko A. Neural Obvious Disobedience Decisions for deep learning on tabular data. arXiv preprint arXiv:1909.06312. 13 September 2019.

26. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asrivatham SJ, Deshmukh AJ, Gersh BJ et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet 2019;394:861–7.

27. Wang F, Kaushal R, Khullar D. Should health care demand interpretable artificial intelligence or accept “black box” medicine? J Am Coll Cardiol 2020;75:57.

28. Liu H, Cocea M. Semi-random partitioning of data into training and test sets in cross-validation. In: 2017 Computing in Cardiology Conference (CinC). Computing in Cardiology. 2017.10.2249/cin17.2017.160.170.

29. Popov S, Morozov S, Babenko A. Neural Obvious Disobedience Decisions for deep learning on tabular data. arXiv preprint arXiv:1909.06312. 13 September 2019.

30. Krizhevsky A, Hinton GE. ImageNet classification with deep convolutional neural networks. 2012. arXiv:1202.2795.

31. Charlie G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ et al. Deep learning: a primer for radiologists. Radiographics 2017;37:2111–31.
62. Tadesse GA, Zhu T, Liu Y, Zhou Y, Chen J, Tian M et al. Cardiovascular disease diagnosis using cross-domain transfer learning. Conf Proc IEEE Eng Med Biol Soc 2019;2019:4262–5.

63. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. PhysioNet/CinC Challenges. https://physionetchallenges.github.io/2020/ (21 May 2020, date last accessed).

66. Lin C-S, Lin C, Fang W-H, Hsu C-J, Chen S-J, Huang K-H et al. A new deep-learning algorithm (ECG12Net) for detecting hypokalemia and hyperkalemia by electrocardiography: algorithm development. JMRI Med Infor 2020;8:e15931.

67. Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G et al. Age and sex estimation using artificial intelligence from standard 12-lead ECGs. Circ Arrhythm Electrophysiol 2019;12:e007284.

68. Raghunath S, Ulloa Cerna AE, Jing L, vanMaanen DP, Stough J, Hartzel DN et al. Development and validation of a deep-learning model to screen for hyperkalemia from the ECG. Ann Emerg Med 1991;20:1229–32.

69. Yap J, Lim WK, Sahlin A, Chin CW-L, Chew K-C, Davila S et al. Harnessing technology and molecular analysis to understand the development of cardiovascular diseases in Asia: a prospective cohort study (SingHEART). BMJ Cardiovascular Disord 2019;19:2659.

70. Yao X, McCoy RG, Friedman PA, Shah ND, Barry BA, Behnken DM et al. A deep-learning algorithm (ECG12Net) for detecting hypokalemia and hyperkalemia by electrocardiography: algorithm development. JMRI Med Infor 2020;8:e15931.

71. Sammani A, Jansen M, Linschoten M, Bagheri A, N de J, Kirkels H et al. Identification of patients with atrial fibrillation: a big data exploratory analysis of the UK Biobank. Physiol Meas 2020;41:025001. 10.1088/1361-6579/ab6f9a

72. Bundy JD, Heckbert SR, Chen LY, Lloyd-Jones DM, Greenland P. Evaluation of risk prediction models of atrial fibrillation (from the Multi-Ethnic Study of Atherosclerosis [MESA]). Am J Cardiol 2020;125:55–62.

73. Melero-Alejano JL, Cascon M, Romero A, Vara PP, Barreira-Peres M, Vicente-Palacios V et al.; SALMANTICOR study. Rationale and design of a population-based study to identify structural heart disease abnormalities: a spatial and machine learning analysis. BMJ Open 2019;9:e024605.

74. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ 2020;368:l6927.

75. Han X, Hu Y, Foschini L, Chintz L, Jankelson L, Ranganath R. Deep learning models for electrocardiograms are susceptible to adversarial attack. Nat Med 2020;26:360–3.

76. Zeller MD, Fergus R. Visualizing and Understanding Convolutional Networks. In European Conference on Computer Vision 2014 Sep 6 (pp. 818-833). Springer, Cham.

77. Xiong Z, Stiles M, Zhao J. Robust ECG Signal Classification for the Detection of Atrial Fibrillation Using Novel Neural Networks. Computing in Cardiology 2017, 10.22489/cincl2017.066-138.

78. Ke Wang E, Xi L, Sun R, Wang F, Pan L, Cheng C et al. A new deep learning model for assisted diagnosis on electrocardiogram. Mathematical Biosciences and Engineering 2019;16:2481–91. 10.3934/mbe.2019124

79. Brisk R, Bond R, Banks E, Padio A, Finlay D, Mclaughlin J et al. Deep learning to automatically interpret images of the electrocardiogram: Do we need the raw samples?. Journal of Electrocardiology 2019;57:565–9. 10.1016/j.jelectrocard.2019.09.018

80. Ivanovic M D, Vladimir A, Alexei S, Lupco H, Malucovk A. Deep Learning Approach for Highly Specific Atrial Fibrillation and Flutter Detection Based on RR Intervals.” Conference Proceedings:. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society Conference 2019:1780–3.

81. Oster J, Hopewell J C, Ziberna K, Wijesurendra R, Cassm C F, Cassde B et al. Identification of patients with atrial fibrillation: a big data exploratory analysis of the UK Biobank. Physiol Meas 2020;41:025001. 10.1088/1361-6579/ab6f9a

82. Chen T-M, Huang C-H, Shih E S, Hu Y-F, Hwang MJ. Detection and Classification of Cardiac Arrhythmias by a Challenge-Best Deep Learning Neural Network Model. iScience 2020;23:100886. 10.1016/j.isci.2020.100886

83. Yoon D, Lim H S, Jung K, Kim T Y, Lee S. Deep Learning-Based Electrocardiogram Signal Noise Detection and Screening Model. Healthc Inform Res 2019;25:201.10.4258/hir.2019.25.3.201

84. Russak AJ, Chaudhry F, De Fretas J K, Baron G, Chaudhry F F, Bienstock S et al. Machine Learning in Cardiology—Ensuring Clinical Impact Lives Up to the Hype. J Cardiovasc Pharmacol Ther 2020;25:379–90. 10.1177/1074248420928651