ESC European Society of Cardiology

European Heart Journal (2019) **00**, 1–15
doi:10.1093/eurheartj/ehz056

CLINICAL REVIEW
Frontiers in cardiovascular medicine

Deep learning for cardiovascular medicine: a practical primer

Chayakrit Krittanawong[1,2]*, Kipp W. Johnson[3], Robert S. Rosenson[2], Zhen Wang[4,5], Mehmet Aydar[6], Usman Baber[2], James K. Min[7], W.H. Wilson Tang[8,9,10], Jonathan L. Halperin[2], and Sanjiv M. Narayan[11]*

[1]Department of Internal Medicine, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA; [2]Department of Cardiovascular Diseases, Icahn School of Medicine at Mount Sinai, Mount Sinai Hospital, Mount Sinai Heart, New York, NY 10029, USA; [3]Department of Genetics and Genomic Sciences, Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [4]Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN 55905, USA; [5]Division of Health Care Policy and Research, Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA; [6]Department of Computer Science, Kent State University, Kent, OH 44240, USA; [7]Department of Radiology, New York-Presbyterian Hospital and Weill Cornell Medicine, New York, NY 10065, USA; [8]Department of Cardiovascular Medicine, Heart and Vascular Institute, Cleveland Clinic, OH 44195, USA; [9]Department of Cellular and Molecular Medicine, Lerner Research Institute, Cleveland, OH 44195, USA; [10]Center for Clinical Genomics, Cleveland Clinic, Cleveland, OH 44195, USA; and [11]Cardiovascular Institute and Department of Cardiovascular Medicine, Stanford University Medical Center, Stanford, CA 94035, USA

Received 29 September 2018; revised 2 November 2018; editorial decision 10 January 2019; accepted 22 January 2019

Deep learning (DL) is a branch of machine learning (ML) showing increasing promise in medicine, to assist in data classification, novel disease phenotyping and complex decision making. Deep learning is a form of ML typically implemented via multi-layered neural networks. Deep learning has accelerated by recent advances in computer hardware and algorithms and is increasingly applied in e-commerce, finance, and voice and image recognition to learn and classify complex datasets. The current medical literature shows both strengths and limitations of DL. Strengths of DL include its ability to automate medical image interpretation, enhance clinical decision-making, identify novel phenotypes, and select better treatment pathways in complex diseases. Deep learning may be well-suited to cardiovascular medicine in which haemodynamic and electrophysiological indices are increasingly captured on a continuous basis by wearable devices as well as image segmentation in cardiac imaging. However, DL also has significant weaknesses including difficulties in interpreting its models (the 'black-box' criticism), its need for extensive adjudicated ('labelled') data in training, lack of standardization in design, lack of data-efficiency in training, limited applicability to clinical trials, and other factors. Thus, the optimal clinical application of DL requires careful formulation of solvable problems, selection of most appropriate DL algorithms and data, and balanced interpretation of results. This review synthesizes the current state of DL for cardiovascular clinicians and investigators, and provides technical context to appreciate the promise, pitfalls, near-term challenges, and opportunities for this exciting new area.

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

Introduction

The practice of cardiovascular medicine routinely requires management of conditions as complex as heart failure with reduced (HFrEF) or preserved (HFpEF) ejection fraction, multivessel coronary disease, complex arrhythmias, sudden cardiac arrest, cardiovascular diseases (CVDs) during pregnancy, or congenital heart disease. Despite advances in each of these areas, significant clinical challenges remain. Many challenges relate to the complexity of integrating data from multiple modalities, making actionable predictions and distilling these solutions to individual patients with heterogeneous phenotypes.[1–3] Deep learning (DL) is a branch of artificial intelligence (AI) that combines computer science, statistics and decision theory to find patterns in complex and often voluminous data. In general, DL is a type of machine learning (ML) that typically utilizes multi-layered neural networks (*Figure 1*). Deep learning has been already shown to outperform experts and other ML strategies in areas as diverse as voice recognition, image classification, commerce and game playing,[4–7] and there is anticipation in that DL could similarly disrupt clinical decision-making by integrating complex data streams, making 'intelligent inferences' and ultimately personalizing therapy.

* Corresponding authors. Tel: +1 212 523 4000, Fax: +1 212 523 8605, Email: chayakrit.krittanawong@mountsinai.org; Tel: +1 650 724 1850, Fax: +1 650 725 7568, Email: sanjiv1@stanford.edu

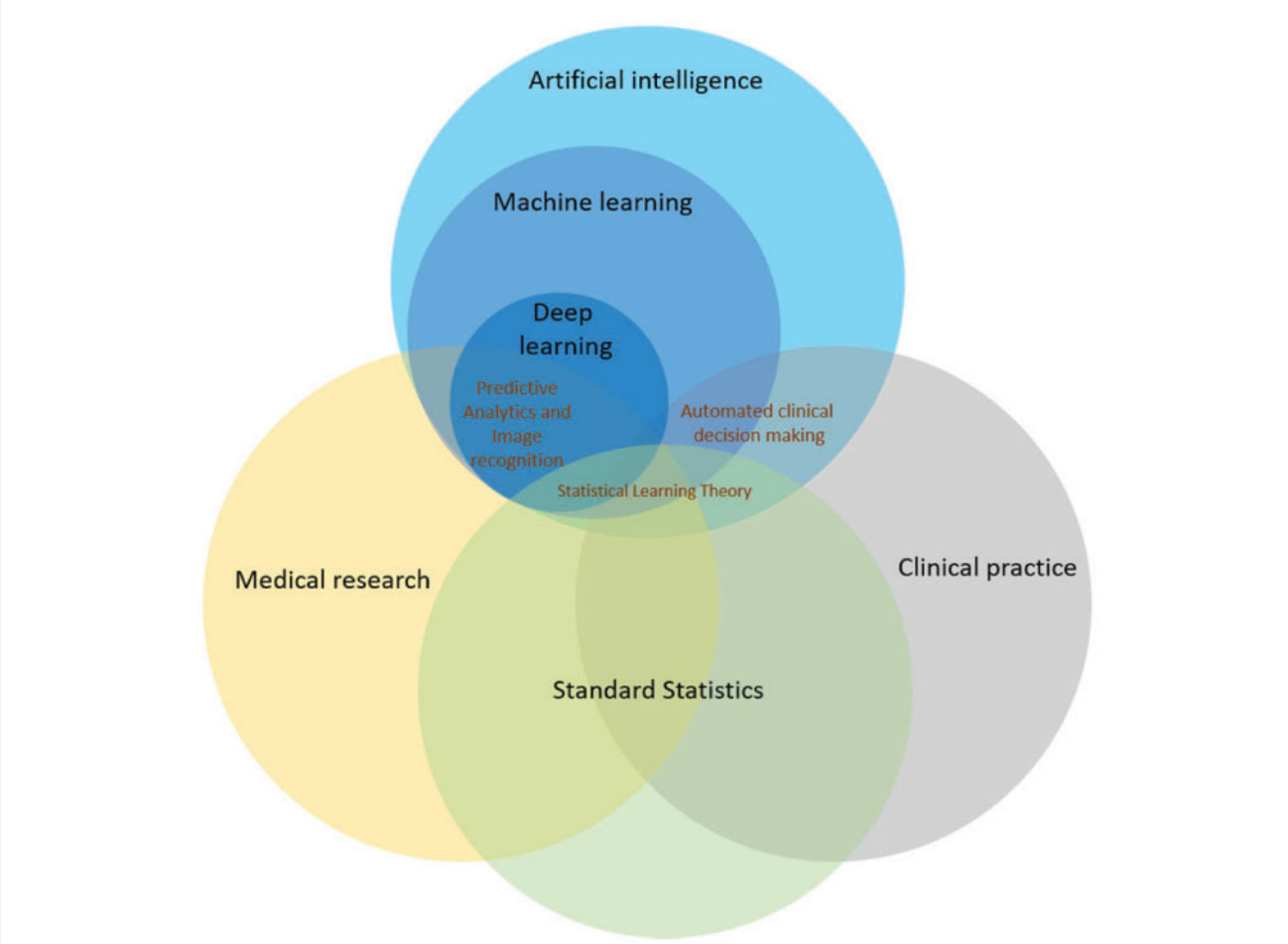


Figure 1 Relationship of deep learning to clinical and translational medicine. Venn diagrams show deep learning as one type of machine learning, within the scope of artificial intelligence. Statistical methods are applied across clinical and translational science, and the form known as statistical learning theory has overlap with machine learning. Automated decision making is often used in clinical practice. Deep learning may extend statistical approaches in some key areas by analysing large multivariate datasets, which often show complex interactions, in which simple hypotheses are difficult to formulate. Deep learning has been successful in medical image recognition (e.g. electrocardiogram, echocardiogram, and magnetic resonance imaging) and holds the promise of enhancing clinic decision making.

There are several parallels, as well as divergences, between traditional statistical methods, ML and DL. Statistical approaches typically test hypotheses or estimate parameters and emphasize inference based upon statistical sampling of a population. In these cases, traditional statistics can be as effective as DL even in large 'big data' applications; for instance in genome-wide association studies (GWAS), a single loci meeting a Bonferroni-adjusted *P*-value (e.g. $P < 5 \times 10^{-8}$) can identify important traits. However, if simple hypotheses are less readily formulated due to complex interactions, for instance if GWAS yields multiple concurrent 'hits', ML may be better suited because it does not require specific hypotheses, and can analyse varied data elements with complex interactions. Deep learning strategies generally attempt to use as much information as is available in a dataset (e.g. every pixel in echocardiography images) in order to generate novel features to be used for downstream analysis. Statistical methods and DL are both influenced by aberrations in sample data and may suffer from overfitting. While statistical methods include approaches to evaluate the possibility of overfitting, a limitation of DL is that it relies to a much greater extent on empirical validation.

To date, cardiovascular applications of DL have been promising[8–10] although many challenges remain. In particular, it is critical to select the right tools for each specific problem and dataset in CVD. This practical review is designed to enable the reader to understand and evaluate applications of DL to cardiovascular medicine or research. We discuss the historical development of DL, definitions, review the current literature to recognize optimal applications, summarize the design and interpretation of DL studies, discuss current challenges and pitfalls, and future directions.

History and definitions

Artificial intelligence is the field of computer science broadly focused on teaching computers to learn complex tasks and make predictions. Early AI applications focused on hand-developing complex decision

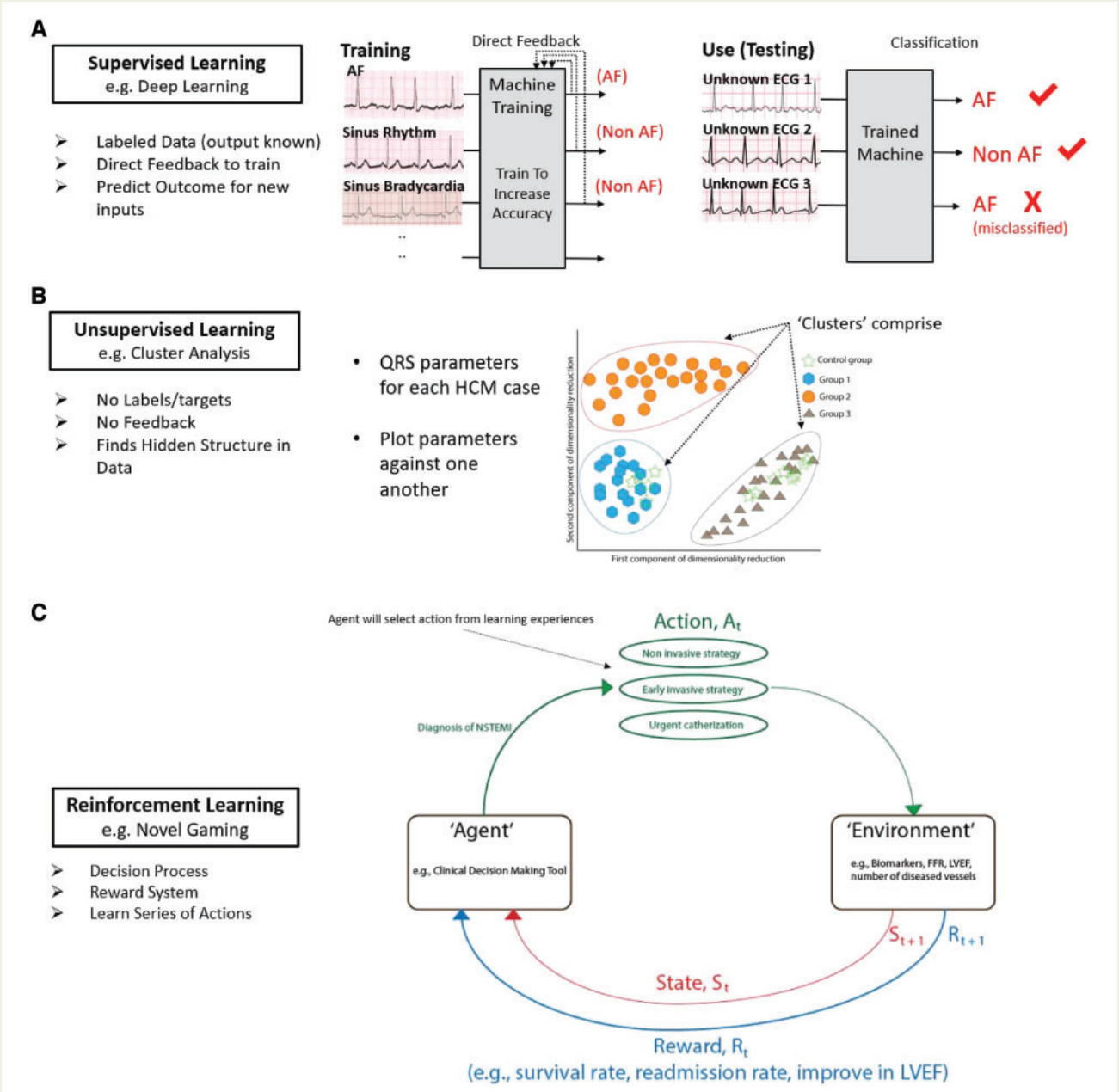


Figure 2 Types of machine learning in cardiovascular science. (*A*) Supervised learning uses inputs (e.g. electrocardiograms) each with a label ('ground truth', and a diagnosis of atrial fibrillation or not atrial fibrillation). Machines are iteratively 'trained', using direct feedback for multiple inputs, until their output matches the ground truth. Trained machines can then classify unknown (test) electrocardiograms. One misclassification is shown. (*B*) Unsupervised learning uses unlabelled data, ideally in large quantities, to identify novel patterns. In this example, QRS indices identified novel phenotypes ('clusters') for hypertrophic cardiomyopathy with distinct outcomes (Ref: Lyon *et al.*[26]). (*C*) Reinforcement learning uses models developed from psychological training applied to gaming, but infrequently to medicine. An agent, e.g. a clinical decision-making tool, performs an action A_t (e.g. which therapy for non-ST-segment elevation myocardial infarction best reduces mortality? (1) non-invasive, (2) early invasive, and (3) mixed) that alters the environment (e.g. biomarker response or patient outcomes). A R_t reward is then given (e.g. higher survival rate) that alters the state S_t. This process is iterated with the intention of moving State S_{t+1} closer to the desired goal (i.e. improved outcomes).

rules for computers to follow but, due to the complexity of human decision making, this was by-and-large not successful. Instead ML, and particularly DL, have emerged as more promising. Machine learning analyses data in ways that automate the construction of analytical models and decision rules, developing systems that learn from data and can identify patterns and make decisions. Ideally, this can happen with minimal human intervention. Machine learning can be further subdivided into supervised, unsupervised, or reinforcement learning. *Figure 2* summarizes categories of ML and DL in the context of emerging applications in cardiovascular science. Supervised learning

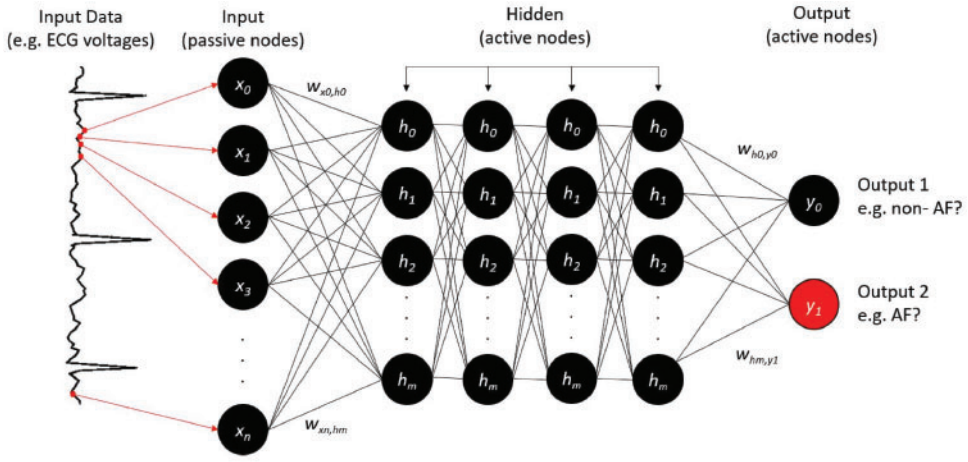


Figure 3 Neural network design to classify atrial fibrillation from the electrocardiogram. Continuous electrocardiogram voltage points (red dots, arrows) are fed to 'input neurons' (x_0, x_1, x_2, . . ., x_m), which are coded as software objects. Hidden neurons within this three-layer network (h_0, h_1, h_2, . . ., h_n) connect input and output layer neurons (here, two neurons) by numerical weights (w). Deep learning typically uses multiple hidden layers, as shown here. The output indicates atrial fibrillation (y_1; correct, red) or non-atrial fibrillation (y_0). If the output is correct for that electrocardiogram input, weights are strengthened; else they are reduced. This process is iterated during training on multiple input electrocardiograms. The trained network can then be tested on new (unseen) electrocardiograms. Other designs could accept categorical variables (age, gender) or mixed data types.

identifies patterns in large amounts of data that are typically annotated ('labelled') by humans, such as the presence or absence of reduced systolic function on an echocardiogram or atrial fibrillation (AF) on an electrocardiogram (ECG). Supervised learning may be implemented using neural networks, long-used for medical pattern recognition in cardiology,[11,12] neuroscience,[13–15] and other fields yet still limited in clinical use. Unsupervised learning, on the other hand, analyses large amounts of unlabelled data to identify hidden patterns or natural structure in data,[16] which greatly increases the volume of data that can be analysed (e.g. from large electronic medical records) at the potential cost of data quality and interpretability. Reinforcement learning trains software to make decisions that maximize a 'reward' function,[17] which may address a clinical problem (e.g. improve ST-segment elevation myocardial infarction outcomes, or reduce error in ECG diagnosis).

Deep learning is a specific type of ML inspired by the way that the human brain processes data, and enabled by hardware advances such as graphics processing units (GPUs),[16,18,19] vast catalogues of labelled data, and advances in computer science theory. To date, most implementations in medicine have used convolutional neural networks (CNNs). Following the 'AI winter' from the 1980s, when early rule-based and neural network applications were limited by hardware and algorithmic constraints, DL has accelerated supervised, unsupervised, and reinforcement learning. In 2016, DeepMind's AlphaGo Zero[4,5] beat the world champion in the ancient Chinese game of Go, and deep Q-learning[6] proved as accurate as a professional human player in 49 interactive video games. In 2018, DL applications by DeepMind rivalled humans in the 3D multiplayer videogame Quake III Arena.[7] *Figure 3* provides a schematic of how neural networks, a common basic architecture for DL, could be used to classify an ECG.

Implementing deep learning in cardiovascular applications

Hardware and software considerations

Historically, ML was computationally expensive and performed by scientists using supercomputers or high-end workstations with multicore processors. However, due to the highly parallelizable nature of DL algorithms, GPUs designed for gaming have enabled DL to be performed on desktop machines. Although professional-quality GPUs are still relatively expensive, DL can be performed using cloud services by services such as Amazon AWS or Google Cloud. Deep learning software packages are almost uniformly open-source, which means they are freely available with few constraints for academic research. Furthermore, many DL pre-trained models can be downloaded and repurposed for new tasks, including models such as AlexNet,[20] VGG Network,[21] InceptionNet,[22] and ResNet,[23] and will be further described in following sections. In fact, downloading pre-trained models and repurposing them for new datasets avoids much of the time consuming and computationally expensive steps of DL. *Table 1* demonstrates step-by-step an example DL process in cardiovascular medicine.

Selecting a deep learning software package and modelling strategy

The first practical step for DL is to choose an appropriate software package to work with such as Keras, Tensorflow, or others. Keras is often used as a starting point, as it can be used in a relatively straightforward fashion with high-level programming languages, most commonly Python. Supplementary material online, *Table S1* summarizes some platforms and related programming languages for

Table 1 A guide to approach of the deep learning applications in cardiovascular medicine research and clinical practice

Step	How to
Identify research questions and outcomes	A supervised learning problem (e.g. predict labelled outcomes) vs. unsupervised learning problem (e.g. identify or classify new phenotypes of HFpEF or new genotypes of PAD) • If supervised, is it a regression task (e.g. prediction of the cost of care for a PCSK9 inhibitor treatment) or classification (e.g. predict if a given patient has a disease or not)
Data selection	Public databases vs. EHR databases vs. Registry • Identify limitations of databases and attempt to replace with rational variables (i.e. no lab values, no vital signs between admission and discharge, no medications, no specific ICD codes for MitraClip) • Identify appropriate methodologies for database (e.g. do we realistically have enough data to attempt DL? If not, are we better off selecting an alternate approach to our problem?)
Hardware selection	Computer cluster vs. workstation with GPUs vs. cloud computing services • Can we build or buy a computer cluster or workstation with GPUs? In the long run, this will be much cheaper but with much greater expense upfront. Should we instead use cloud computing services (i.e. Google Cloud or Amazon Web Services), which have little-to-no upfront cost but can cost more in the long run
Data preparation	(1) De-identify data if needed (2) Quality control of data—assessment of missingness and verification that our dataset contains what it should. Identify mechanism of missing data and then data imputations for non-ignorable missing data (3) Denoising of images/video/textures or variant calling in NGS data (4) Exploratory analysis (summarization, visualization, identify structure of data/relationship between variables)
Feature selection	In general, DL requires little a priori feature selection. If the input dataset is highly multidimensional, strategies such as vector embedding may be required first in order to pass features to other DL models
Data splitting	Design and justify the proportion of training, validation and testing in the dataset (i.e. 70/10/20 or 80/10/10 or 60/20/20)
Modelling selection	(1) One should always see if the task can be accomplished with a simple model or standard statistical approaches (can we simply apply linear/logistic regression or polynomial regression to our dataset? If so, does it perform adequately?) (2) If simpler algorithms do not work, more complex strategies such as DL may be needed. In general, reuse of pre-trained models using transfer learning is preferred since many pre-trained models are well validated in a large database and their performance characteristics and limitations are known (3) If no pre-trained models are available for specific type of data or research questions, develop new algorithms based on highest achievement model for specific type of data
Technical details for model	Know some DL technical terms to communicate with data scientists or programmers and understand the process (learning rate selection, tuning hyperparameter, batch dropout and normalization, regularization strategies, loss function selection, and network optimization)
Evaluation of model discrimination	Report ROC curve, C-statistics, NPV, PPV, sensitivity, and specificity
Evaluation of model calibration	Compare with standard statistical approaches (i.e. multivariable regression), goodness-of-fit, calibration plots, or the decision-curve analysis
Ground truth	Compare with human experts (cardiologists, electrophysiologist, primary care physicians)
Publication and transparency	Share codes with journal (i.e. online supplements) or public space (i.e. Githut, bioRxiv). DL methodologies should be clearly explained in details. Consider strategies for computational anonymization
Generalization and replication results	Test with different datasets in a different population
Clinical trials	To minimize risks or errors, testing prediction models in clinical trials is recommendated
Meta-analysis	Meta-analysis of DL is needed to assess publication bias and heterogeneity
AI Guidelines	With clinical trial and meta-analytic results, professional societies have to develop guidelines to regulate DL in clinical decision making or predictive analytics in clinical practice
Implement in clinical practice	Start implementing in clinical practice and monitor the results closely

AI, artificial Intelligence; DL, deep learning; EHR, electronic health record; GPU, graphics processing unit; HFpEF, heart failure with preserved ejection fraction; NGS, next-generation sequencing; NPV, negative predictive value; PAD, peripheral arterial disease; PCSK9, proprotein convertase subtilisin/kexin type 9; PPV, positive predictive value; ROC curve, receiver-operating characteristic curve.

DL. Second, one should choose a DL model appropriate for the problem under consideration, which may be pre-existing (pre-trained models) or require development of a custom model (novel models). In general, due to time restrictions, and computational requirement, pre-trained models are mostly used first, and then tailored to investigators' problems/datasets by modifying outcomes related layers [i.e. last few layers or final layer (softmax layer)] to optimize their results. This concept is called transfer learning.[24] For example, if a pre-trained model was trained to accurately predict diastolic dysfunction using global longitudinal strain in HFrEF

patients in a large database, we might be able to use resultant knowledge to conduct the prediction of diastolic dysfunction in HFpEF patients. There are several pre-trained models such as AlexNet,[20] which recognizes visual patterns directly from pixel images with minimal preprocessing, the VGG Network[21] that performed well in the ImageNet Large Scale Visual Recognition Challenge in 2014, or other models (e.g. InceptionNet[22]). When validated and pre-trained models are unavailable for specific applications, which may be common in cardiovascular medicine at this early stage, custom models may be required. Each will require design (e.g. the number of layers, nodes, learning rate, and so on in *Figure 3*), validation and tuning.[4,25] Examples are provided in Supplementary material online, *Table S2*. However, transfer learning is challenging in medicine as there are differences in datasets, i.e. sources of data (non-medical videos vs. echocardiographic videos), data quality, vendor or softwares. There is no standardized guideline for when it is appropriate to use transfer learning between datasets, and this often requires empirical trial-and-error approaches.

Learning rate is an important concept that determines how the network adjusts weights during training based on correct or incorrect decisions. *Figure 4* shows appropriate learning rates in a well-trained network, and how inappropriate learning rates (too high) may impede network training. *Figure 5* shows how complex cardiovascular data that are not readily separated by a simple cut point or threshold (i.e. cannot be partitioned by a linear classifier) can be transformed and now readily separated.

Finally, a major strength of DL is its ability to infer and classify data outside its original training set (i.e. generalization). To generalize well, DL strategies must avoid overfitting to the training data. Unlike statistical methods, for which strategies exist to avoid overfitting, in DL, this is often empirical and trial-and-error. The first and best approach is usually to increase power by adding data. Another common technique in DL is data augmentation, i.e. increasing the number of training samples using the same raw data. In image analysis, for example, it is common to rotate, invert, or skew an image if it does not distort data or change its output class (e.g. rotation would not alter hypokinesis on echocardiography or the location of a space-occupying lesion), but may be unacceptable in traditional biostatistics. Other strategies include cross-validation of training data, and transformation to reduce data complexity. In complex DL architectures (e.g. several hidden layers in a neural network, *Figure 3*), reducing the numbers of neurons or layers may be effective. The ultimate test of generalization is prediction in a new dataset.

Data preparation

Data preparation must be tailored to each specific DL architecture to optimize performance and requires feature selection and imputation of missing data.

Feature selection

This involves transforming raw data to features that reduce dimensionality of the data, but that still represent the relevant clinical or physiological question, and may reduce the risk of overfitting. In *Figure 3*, input data were digitized ECG voltage-time series, while in *Figure 2B* features were used instead such as QRS duration and the intrinsicoid deflection.[26] Since a key advantage of DL is to learn complex features, overly complex preprocessing of data into features may also impede performance. One general goal of DL may be summarized as letting the algorithm automatically perform feature selection for the user, instead of the user attempting to manually engineer features.

Missing data imputation

Similar to statistical tasks, the performance of DL can be highly sensitive to missing data. The cut-off to discard features where entries are missing remains debated. Decisions on how to treat missing data can be made by evaluating if the presence or absence of specific elements correlate with desired outcomes or predictors. Those data that are correlated are 'non-ignorable', those that are not correlated may be 'ignorable'.[27] Ignorable missing data includes missing completely at random (no relationship to any variables) and missing at random (a systematic relationship between missing values and observed data). Identifying these types of missing data is a crucial before data imputation.[28] Several strategies exist for imputation. One approach to account for systematic data omission is to insert a label for those elements. For instance, since individuals too sick (e.g. with elevated serum creatinine) to receive a cardiac computed tomography (CT) may have worse outcome than those who do undergo cardiac CT, a label that indicates 'too sick' would likely still be prognostic for outcome in a model missing CT data in sick patients. Here, CT data is non-ignorable, but training can proceed with this label. In other cases, training can proceed with insertion of an imputed value, multiple imputations (i.e. MICE or missForest R packages), or expected values from the literature. One class of DL algorithms called autoencoders have been shown to produce best-in-class results for missing data imputation.[29]

Training and expected results

Datasets should be separated into distinct partitions for training, testing, and validation. One area of future improvement in DL is in considering data containing repeated measures or correlations (time series, or where observations cluster by patient). However, this is an active area of research within the DL community.[30,31] More information on sampling by patient for correlated data (i.e. repeated measures) should also be included. The specific training, testing, and validation proportions used in DL often vary depending upon the task at hand, and in general selecting one strategy is more of an 'art' than 'science'. 70/10/20 or 80/10/10 or 60/20/20 splits are common, but no standard methodologies exist to determine optimum proportions although each should come from similar data distributions to prevent mismatched proportions between sites. Training strategies are thus largely empirical with no standardized approaches.

Novel deep learning techniques

Anticipated advances in DL may mitigate some concerns of empiricism and lack of a theoretical framework. Capsule-based neural network,[32,33] meta-learning (learning how the network learns),[34–36] DeepMath (learn mathematical proofs),[37] or self-play (two agents learn by win and loss)[34] are promising innovations. Studies to optimize learning rate, network architectures, and activation functions may improve the learning efficiency of DL. There are efforts to adapt

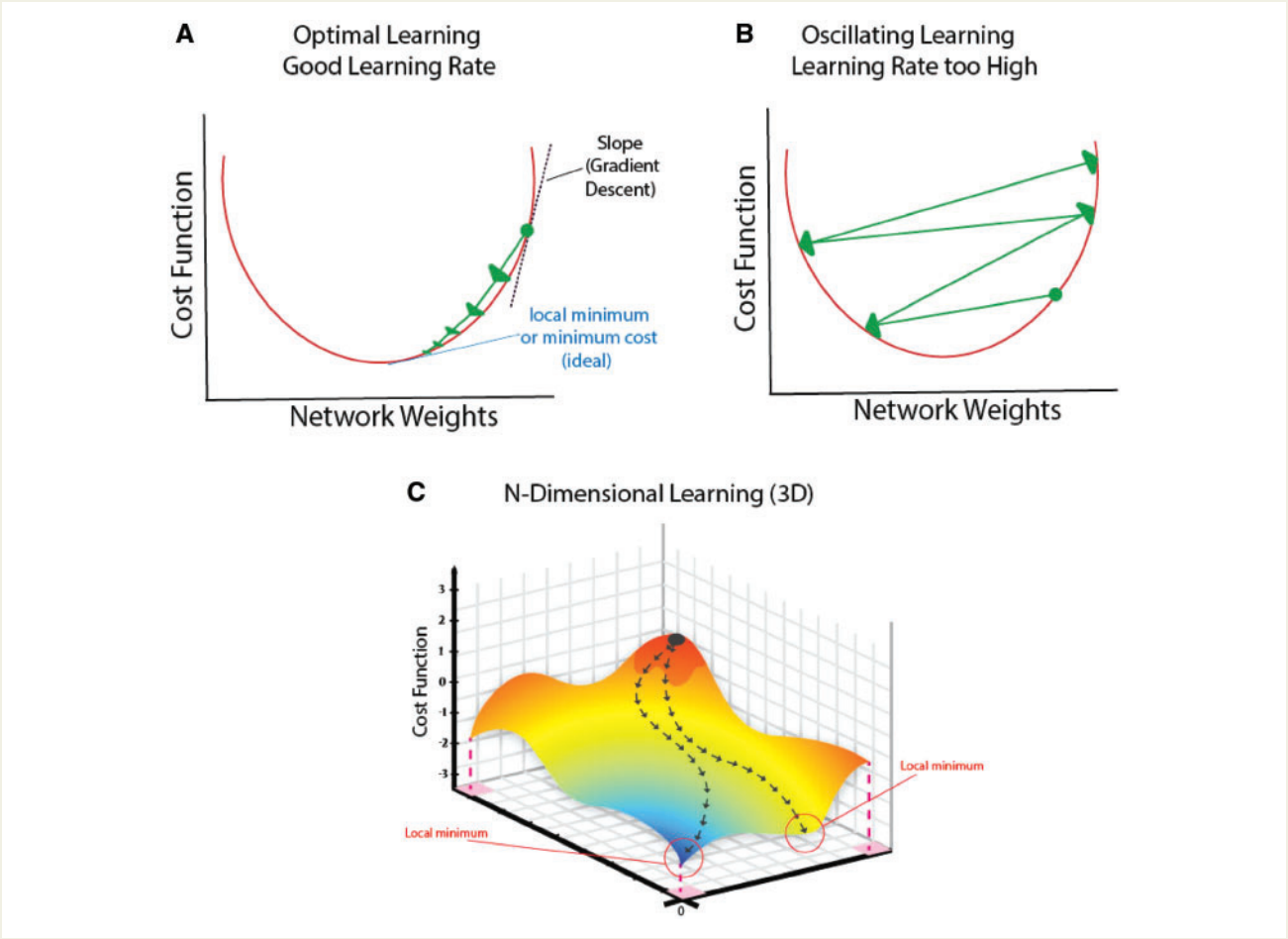


Figure 4 Impact of deep learning design on learning: effect of learning rate. (*A*) Efficient learning. Cost function (network error) gradually descends ('Gradient descent') to achieve the optimal point (called local minimum) as a function of weight. (*B*) Learning rate is too high, so that the cost function overshoots the minimum and oscillates. This network design may not be trained effectively for this problem. (*C*) Gradient descent examining two variables on cost function simultaneously.

several non-neural network-based methods for DL (Supplementary material online, *Table S2*). Support vector machines, for instance, are effective for high-dimensional data[38,39] and could be useful in cardiovascular DL with diverse datasets.[40,41] Enriching the theoretical framework for DL may improve our ability to interpret their conclusions.

Deep learning applications in cardiovascular medicine

Table 2 summarizes early DL applications in cardiovascular medicine,[42–51] while Supplementary material online, *Table S3* lists DL studies in other disciplines.

Imaging for ischaemic and structural heart disease

Deep learning has been used to classify images in many medical specialties,[52–54] recently extended to cardiac imaging (*Table 2*). By interpreting cardiac images rapidly and consistently, DL may circumvent clinical limitations of fatigue or distraction, variable inter- and intra-observer interpretation, and time-consuming interpretation of large datasets.

Zreik *et al.*[55] applied DL to automatically identify significant coronary artery stenosis in rest coronary CT angiograms in 166 patients. Compared with matched invasive fractional flow reserve (FFR) measurements, the network produced a c-statistic of 0.74 ± 0.02 with specificities of 77%, 71%, and 59% at sensitivities of 60%, 70%, and 80%, respectively, providing a possible alternative to invasive FFR.[55] Betancur *et al.*[56] applied DL to single-photon emission computed tomography myocardial perfusion imaging in 2619 consecutive patients at exercise (38%) or pharmacological stress. After 3.2 ± 0.6 years follow-up, DL better predicted major adverse cardiac events (MACE; 9.1% of patients overall) for DL using imaging with stress test data than imaging data alone (area under the receiver-operating characteristic curve: 0.81 vs. 0.78; $P < 0.01$), and both were superior to existing assessments. Motwani *et al.*[57] used DL on CT angiography in 10 300 patients with suspected coronary disease to improve prediction of 5 year all-cause mortality over existing CT or

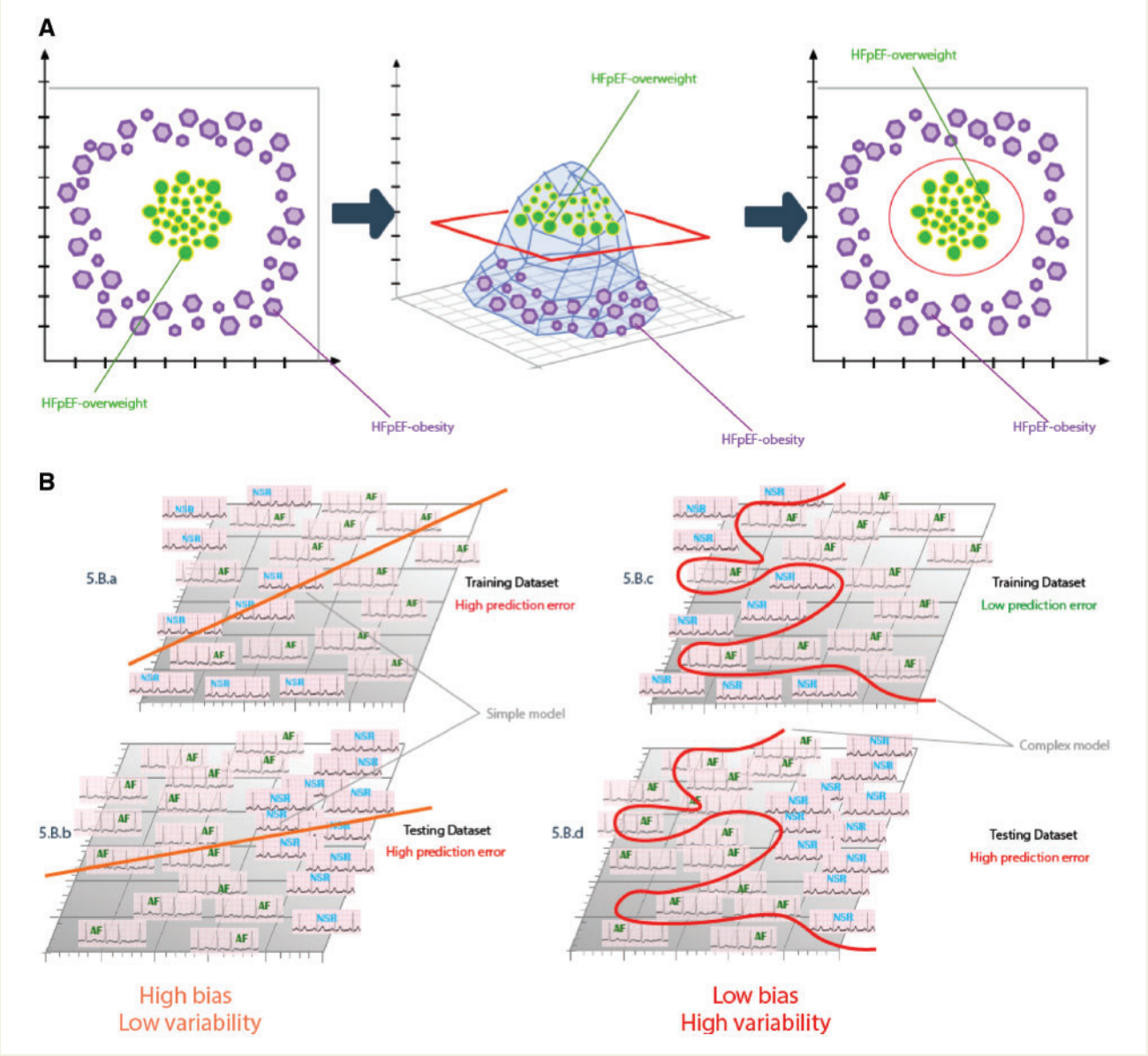


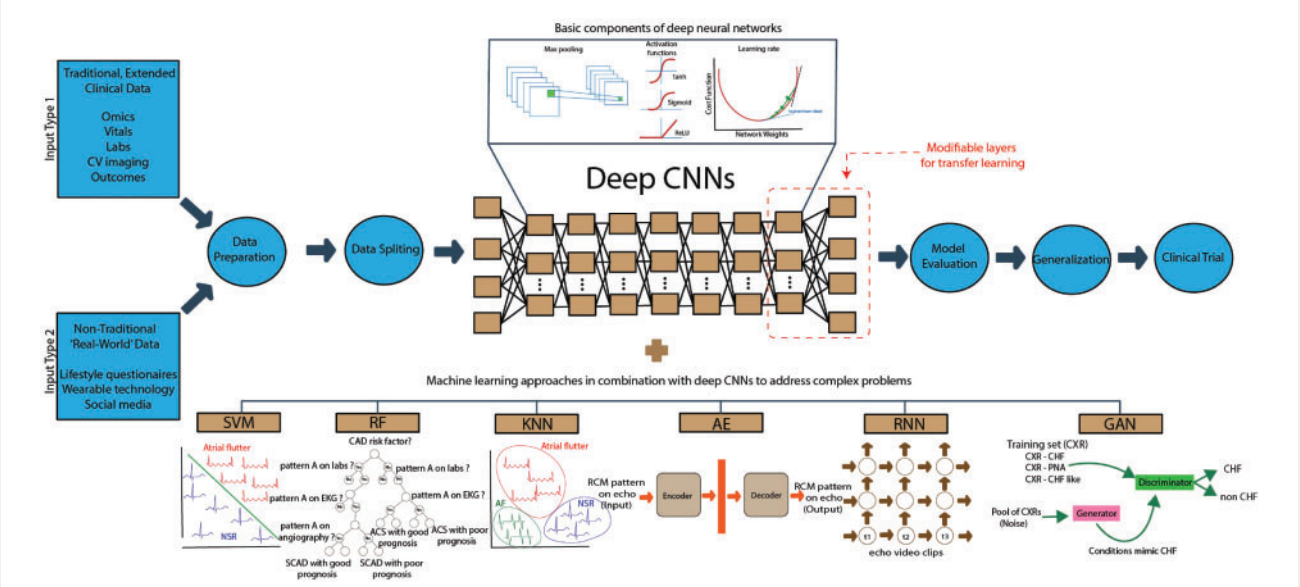
Figure 5 Classifying complex data. (*A*) Transforming data to enable linear separation of non-linearly separable raw data. Raw non-linear data are transformed by mapping functions that may include time, frequency, or other operations. This projects them into higher-dimensional parameters space in which they are now linearly separable. One example is classifying patients with heart failure with preserved ejection fraction whose response to beta-blockers may vary due to obesity, atrial fibrillation, left ventricular hypertrophy, diabetes, or other factors. Data transformation to a higher-dimensional space now enables a simple partitioning process. (*B*) Bias–variance tradeoff. Model with high bias (straight line), when a straight line could not classify appropriately (here, between atrial fibrillation and normal sinus rhythm) in both training dataset (5.B.a) and testing dataset (5.B.b). This leads to prediction errors on other datasets (low variance - frequent errors). In contrast, model with low bias (i.e. due to overtraining) when data is fitted well in training set (5.B.c), but not in testing set (5.B.d), leading to reduced generalization (high variability due to difference between training and validation sets).

clinical indices as supported by others.[58] In a novel study applying DL to retinal fundus photographs trained on 284 335 patients, DL predicted MACE with c-statistic 0.70 in two validation populations (12 026 and 999 patients).[59] Several reports have used DL to define structural heart disease from echocardiography and MR imaging (*Table 2*). Deep learning can diagnose structural disease from limited echocardiographic views.[44] DeepVentricle,[71] a DL application for cardiac MRI images, received FDA clearance for clinical use.[72] Bai *et al.*,[62] applied DL to automatically segment 93 500 labelled MR images in 4875 subjects from the UK Biobank with similar accuracy to experts for segmenting left and right ventricles on short-axis images and left and right atrium on long-axis images.

Table 2 Deep learning applications in cardiovascular medicine

First author	Disease application	Images/slides (*N*)	Type of images	Results	Algorithms
Arrhythmias					
Bumgarner *et al.*[60]	AF detection	169 records	A rhythm strip from Kardia mobile phone	93% sensitivity and 84% specificity for AF detection	Kardia Band automated algorithm
Pyakillya *et al.*[50]	ECG classification	8528 records	ECG records	86% accuracy for classification results	Deep CNN
Hannun *et al.*[49]	Arrhythmias detection (i.e. VT, Mobitz I, Mobitz II)	64 121 ECG records	ECG records	The model outperforms the average cardiologist performance on most arrhythmias (measured by sequence level accuracy and set level accuracy)	Deep CNN
Tison *et al.*[8]	AF detection	9000 ECGs	ECG from smartphone	On recumbent ECG, c-statistic 0.97; on ambulatory ECG, c-statistic 0.72	Deep CNN
Xia *et al.*[48]	AF detection	123 848 samples	ECG signal (STFT and SWT)	98.29% accuracy for STFT and 98.63% accuracy for SWT	Deep CNN
Cardiac MRI					
Avendi *et al.*[63]	Image segmentation (LV shape detection)	45 MRI datasets	Cardiac MRI images	90% accuracy (in terms of the Dice metric)	Deep CNN and stacked autoencoders
Bai *et al.*[62]	Image segmentation (LV and RV)	93 500 images	Cardiac MRI images	Reported as Dice metric, mean surface distance, and hausdorff surface distance	Deep CNN (a fully CNN)
Luo *et al.*[64]	Image segmentation (LV volumes)	1140 subjects	Cardiac MRI images	High accuracy in LV volumes prediction (measured accuracy by root mean squared error)	Deep CNN
Oktay *et al.*[65]	Image quality (i.e. LV segmentation, motion tracking)	1233 healthy adult subjects	Cardiac MRI images	Directly compare image quality and segmentation results (i.e. LV cavity volume differences or surface-to-surface distances)	Deep CNN (super resolution approaches)
Oktay *et al.*[47]	Image segmentation (cardiac pathologies)	1200 images	Cardiac MRI images	Reported as Dice metric, mean surface distance, and hausdorff surface distance	Deep CNN
Echocardiography					
Dong *et al.*[66]	LV segmentation	60 subjects	Echocardiography images	Reported as Dice metric, mean surface distance, and hausdorff surface distance	Deep GAN (VoxelAtlasGAN)
Gao *et al.*[67]	8 echo views classification	432 video images	Echocardiography images	92.1% accuracy for classification results	Deep CNN
Knackstedt *et al.*[68]	Assessment of LV volumes and EF	432 video images	Echocardiography images	92.1% accuracy for classification results	Deep CNN
Luong *et al.*[43]	Assessment of echo image quality feedback	6916 images	Echocardiography images	The average absolute error of the model compared with manual scoring was 0.68 ± 0.58	Deep CNN
Madani *et al.*[44]	15 echo views recognition	223 787 images	Echocardiography images	91.7% accuracy for classification results (F-score 0.904 ± SD 0.058)	Deep CNN
Heart failure					
Nirschl *et al.*[69]	Clinical heart failure classification	209 patients	H&E stained whole-slide images	> 93% accuracy for both training and testing datasets vs. 75% accuracy for pathologists	Deep CNN
Seah *et al.*[70]	Prediction of CHF	103 489 images	Chest X-rays	Model achieved an AUC of 0.82 (At a cut-off BNP of 100 ng/L)	Deep GAN
Islam *et al.*[105]	Pulmonary oedema detection	7284 images	Chest X-rays	The same architecture does not perform well across all abnormalities	Deep CNN
Myocardial perfusion imaging					
Betancur *et al.*[56]	Prediction of obstructive CAD	1638 patients	Fast SPECT MPI	AUC for disease prediction by deep learning was higher than for total perfusion deficit	Deep CNN

AF, atrial fibrillation; AUC, the area under the receiver operating characteristic curve; CAD, coronary artery disease; CHF, congestive heart failure; CNN, convolutional neural network; ECG, electrocardiogram; EF, ejection fraction; GAN, generative adversarial network; HCM, hypertrophic cardiomyopathy; LV, left ventricle; MPI, myocardial perfusion imaging; MRI, magnetic resonance imaging; RV, right ventricle; SPECT, single-photon emission computerized tomography; STFT, short-term Fourier transform; SWT, stationary wavelet transform.



Take home figure Deep learning process flow for cardiovascular medicine. Deep learning has the ability to produce actionable clinical information from diverse datasets. Such data may span (i) comprehensive, traditional clinical data; (ii) non-traditional 'real-world' data such as near-continuous streams from wearable devices but also questionnaires or online forms. The deep learning process flow commences with designing the most appropriate model. Deep learning is usually implemented by deep neural networks with convolutional layers defined by specific parameters (e.g. max pooling, activation function, and learning rate). Several algorithms traditionally applied in machine learning (i.e. SVM, RF, KNN, RNN, AE, GAN) can be combined to address complex problems. Data is selected and pre-processed (curated), and missing elements are imputed. Training proceeds until the deep learning machine converges at acceptable accuracy. The deep learning machine is then ready to be applied to unseen test data.

Deep learning for outcomes prediction in heart failure

Several studies have applied ML to predict outcomes in heart failure (HF).[41,73,74] Choi *et al.*[75] applied neural networks to detect new onset HF from electronic health records in 3884 patients who developed incident HF and 28 903 who did not, linking time-stamped events (disease diagnosis, medication and procedure orders). Networks provided a c-statistic for incident HF of 0.78 (12-month observation) and 0.88 (18-month observation), both significantly higher than the best baseline method. Medved *et al.*[76] compared the International Heart Transplantation Survival Algorithm developed using DL training, with the Index for Mortality Prediction After Cardiac Transplantation (IMPACT), in transplant recipients from 1997 to 2011 from the UNOS registry. In 27 705 patients, using those before 2009 for training and those from 2009 for validation, DL provided a c-statistic for 1-year survival of 0.654 for IHTSA, which reduced error by 1 compared with the IMPACT model (c-statistic 0.608). These results, while modest, show promise for DL beyond current clinical indices. Future studies may apply DL to multi-variable data (e.g. histopathology, echo, ECG, labs, multi-omics, wearable technology) to study HF outcomes. The recent BIOSTAT-CHF trial, a large registry for risk prediction for HF in 11 European countries, has multi-level data that could be used to reclassify HF patients.[77]

Deep learning for arrhythmia detection and phenotyping

Several studies have used DL to diagnose AF from the ECG. Tison *et al.*[8] trained a DL network on 9750 ambulatory smartwatch ECGs, then applied it to 12-lead ECGs. The network performed well in 51 recumbent patients before cardioversion (c-statistic 0.97 vs. 0.91 for current ECG algorithms), but less well in a cohort with ambulatory ECGs (c-statistic 0.72, sensitivity 67.7% and specificity 67.6%). A separate study used AI to diagnose AF from electrical ECG sensors in a smartphone case (Kardia), or a watch-strap which communicates via Bluetooth to a smartphone (Kardia Band).[60] In 100 patients with 169 simultaneous wearable and traditional ECGs, 57 recordings were uninterpretable. For interpretable ECGs, the device identified AF with a K coefficient of 0.77 (sensitivity 93%, specificity 84%) although physician interpretation improved results further.[60] Thus, while these data are promising, further advances in analytic algorithms and sensor technology are needed for automatic use. Emerging sensors beyond optical sensors (photoplesthysmography) in the iWatch[8] include changes in facial reflectance to visible light,[78] bioimpedance in weighing scales[79] and others. The accuracy of each sensor needs validation since, in recent comparisons against gold standards, wearable sensors had acceptable accuracy for resting heart rate yet not for ambulatory exercise heart rates nor energy expenditure.[80] Current ESC guidelines provide a Class I recommendation for the opportunistic screening for silent AF in patients >65 years of age by pulse or ECG rhythm

strip, based on evidence on cost effectiveness.[81] Therefore, integrating DL into wearable technology for intermittent screening for silent AF may be cost effective by preventing sequelae such as stroke.

Finally, ML shows promise in identifying novel arrhythmia phenotypes, using unsupervised learning (since labelled data do not exist, by definition). For AF, Inohara *et al.*[82] analysed the large clinical ORBIT-AF database to identify clusters labelled atherosclerotic-comorbid, tachy-brady device, low comorbidity, and younger behavioural disorder. The cluster approach did not improve CHADS2VASc, ORBIT, and ATRIA scores for endpoints of stroke or bleeding; however, there was a slight improvement in combination (c-statistics 0.67–0.72). In hypertrophic cardiomyopathy, Lyon *et al.*[26] identified four clusters of high risk of sudden cardiac arrest from ECG (primary T wave inversion) and echocardiographic (septal and apical hypertrophy) features. The clinical utility of novel machine learned phenotypes should be validated in independent populations compared with traditional clinical classification.

Challenges for, and limitations of, deep learning

To date, cardiovascular results of DL are promising but still modest, and several challenges must be overcome. First, and most importantly, DL is often criticized in the clinical context as a black box which cannot easily be explained. Interpretability may be enabled by capsule based networks, or strategies that systematically censor inputs to define those that most affect classification. Meta-analyses of several DL algorithms applied to the same data may increase confidence in results. A number of techniques may enable 'model-agnostic' metrics for interpretability of complex models.[18] Marblestone *et al.*[83] hypothesized analogies between DL and human cognitive functioning, proposing that integrating heterogeneous 'cost functions' over time may simplify learning. Thus, speculatively, insights into human cognition may ultimately provide insights to interpret DL models.

Second, all ML may suffer from overfitting (*Figure 5*) if data is limited and/or algorithms complex. Indeed, in some clinical studies DL provided similar results to statistical models (e.g. logistic regression).[84] This may simply mean that different analyses are better suited to different types of data. Future studies may integrate DL with statistical classification.

Third, DL faces recent ethical criticisms if biased or poor-quality data lead to biased predictions or, worse, facilitate manipulation of results. Adversarial examples, cases in which slight modifications to input data cause a major change in DL classification, are a significant concern for DL with potentially serious medical sequelae.[85–87] Some methodologies have been proposed to prevent adversarial examples (i.e. reactive strategies), but they remain ineffective.[88]

Fourth, DL studies must enable replication by other groups since differences in algorithms, initiating conditions, or parameter tuning may alter results.[89] One replication study,[90] for example, demonstrated different results from another[91] using the same algorithms. Thus, a standardized approach to perform and validate AI-related clinical studies is needed. One initial step would be to require investigators to deposit their data and a link to the code for their DL model. Medicine lags behind computer science in this respect.

Fifth, DL in cardiovascular medicine have thus far compared c-statistics (area under the receiver-operating characteristic curve), which has several limitations. Problems are that the c-index is only a measure of discrimination, ignores calibration indices and has no single universally accepted c-statistic cut-off or range of acceptable c-statistics.[92,93] This and other limitations can be addressed by calibration: dividing underlying continuous variables representing the diagnostic into partitions (e.g. deciles) and testing diagnostic ability in each. It is important to use multiple metrics of accuracy since, for example, c-statistics discriminate true outcomes (e.g. high risk from low risk patients), but are insensitive to systematic errors and do not identify whether the model is anchored at the right level of absolute risk across the spectrum of observable risks. This should be applied to machine learned models, but has yet to be done. In addition, receiver-operating characteristic curve is an ordinal technique which assumes that the underlying biological process is monotonic, yet this is often not true. For instance, blood pressure at extreme high or low levels are of disproportionate importance. Machine learning has the advantage that it is not constrained by a monotonic assumption. Reporting a comparison between DL with traditional statistical results (CNN vs. logistic regression), the Brier score, the goodness of fit,[94] calibration plots,[95] standardized checklist/strobe diagrams of prediction models[96] or the decision-curve analysis[97] would be helpful.

Sixth, there is likely a positive publication bias in medical studies of ML and DL. However, since the number of such studies is currently limited, a meta-analysis of such studies may not be worthwhile, so that funnel plots or other indices of publication bias or heterogeneity (I^2) would be difficult to quantify. As discussed in several of the specific examples, negative results are indeed discussed. These studies should be used to refine DL methods, which then need to be tested via external independent multicentre validation.

Seventh, DL and standard statistical methods may often be used for the same problems, although in addition to parallels between the methods, divergences exist. Both are influenced by aberrations in sample data, and can suffer from overfitting. This may be more predictable for statistical methods than for ML which relies on empirical validation. Cardiovascular problems which can be stated as a clear hypothesis may be equally addressed by traditional statistics or DL. Conversely, if simple hypotheses are less readily formulated due to complex interactions, DL may be advantageous. Thus, both techniques are complementary tools.

Finally, a theoretical framework to guide big-data and DL designs is urgently needed. Rather than ask simply if data *quantity* is sufficient for study, it is pertinent to ask if data *quality and diversity* are sufficient to span the parameter space necessary to address the question. Adding data dimensions increases the chance of alpha error, i.e. finding chance associations in traditional biostatistics, yet it may enrich training for DL. Are data reliable ('garbage in, garbage out')? In a large big-data arrhythmia study comparing genotype with phenotype in 2022 patients with long QT and Brugada syndromes,[98] variability in genetic testing compromised its results. A final consideration is how to organize data given that most data structures are generic and there are few which are standardized for cardiovascular medicines.

Training programmes in deep learning

Educational programmes should incorporate classes on DL, given its already ubiquitous presence. Such classes should cover the rationale,

solutions, technical, and ethical challenges it poses in medicine. At the undergraduate and graduate level, such training may focus on its complementary role to biostatistics, and on detailed software programming and hardware aspects. In medical education, implementation of a broad AI curriculum is likely to enrich understanding of many conditions in cardiovascular medicine with heterogeneous aetiologies and/or phenotypes such as HFpEF, AF, and hypertension.[99,100] The medical curriculum should also discuss ethical and legal challenges, and their potential to shape medical practice. A significant barrier to implement DL more broadly is the need for at least some programming familiarity. This may be less problematic for newer generations of trainees. Deep learning is an excellent vehicle to foster interdisciplinary teams of engineers, physicians, businessmen, legal, and ethical teams.[100] It may be helpful to borrow engineering approaches such as 'hackathons', e.g. the Data Science Bowl, PhysioNet/Computing in Cardiology Challenges or Kaggle Competition. Scientists including biomedical trainees could compete at these or traditional medical conferences to analyse cardiovascular sample datasets (imaging, ECGs) using DL.

Funding opportunities for DL outside medicine are increasing, but funding from ESC/AHA/ACC/NIH are increasingly needed.[101] Crowdfunding has been an alternative for DL funding outside medicine and, although rare thus far in cardiovascular research, the potential of crowdfunding for cardiovascular DL research is intriguing.[102]

Guidelines must be developed to standardize broad applications of AI in medicine. This will require complex discussions between multiple stakeholders including regulatory agencies in Europe and Asia, the U.S. Food and Drug Administration (FDA) and others, patient-advocacy and privacy groups, professional societies in medicine and computer science, other healthcare organizations and technology companies.

Future direction of deep learning in cardiovascular medicine

Deep learning promises to better integrate medical data sources, address the heterogeneity in patient disease types, bridge the gap between omics research and bedside phenotypes[103,104] and ultimately enable personalized medicine. This may require advances in the science to overcome current limitations including a limited theoretical foundation for design and testing, limited interpretability, and strategies to resolve overfitting.[61] Cardiovascular medicine, in particular, is well suited to benefit from smart analysis of continuous and massive data streams in this new era of wearable sensors, to integrate traditional health data with lifestyle indices (*Take home figure*). If data privacy and security concerns were satisfied, this integration would form the basis of a medical internet of things between medical devices and analytic systems. Seamless integration of diverse sources of data could enable continuous disease monitoring, risk stratification and early warnings of potential decompensation.

Conclusion

Deep learning is a rapidly developing field in the computer sciences with great promise for cardiovascular medicine. Using DL for big data analysis may not only identify hidden information in complex, heterogeneous datasets, but also may bridge the gap between disease pathogenesis, genotypes, phenotypes to enable personalized medicine. However, to transform cardiovascular care, DL will have to address challenges in obtaining extensive labelled data, in improving interpretability and robustness, and in developing standardized approaches for validation and testing. Deep learning is one of the most exciting areas of innovation in cardiovascular medicine that holds the possibility to provide more efficient care with improvement in outcomes.

Supplementary material

Supplementary material is available at *European Heart Journal* online.

Funding

S.M.N. is supported, in part, by grants from the National Institutes of Health (NIH R01 HL 83359; K24 HL103800).

Conflict of interest: S.M.N. is author of patents owned by Stanford University and the University of California Regents, and is a consultant to Beyond Limits.ai. R.S.R. reports research grants to his institution from Akcea, Amgen, AstraZeneca, Medicines Company and Regneron; consulting from Akcea, Amgen, C5, CVS Caremark; non-speaker bureau lectures from Amgen and Kowa; and stock ownership in MediMergent. J.K.M. discloses the following relationships - medical advisory board: GE Healthcare, Arineta; equity interest: Cleerly. All other authors declared no conflict of interest.

References

1. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, Falk V, Gonzalez-Juanatey JR, Harjola VP, Jankowska EA, Jessup M, Linde C, Nihoyannopoulos P, Parissis JT, Pieske B, Riley JP, Rosano GM, Ruilope LM, Ruschitzka F, Rutten FH, van der Meer P. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the task force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur J Heart Fail* 2016;**18**:891–975.
2. Regitz-Zagrosek V, Roos-Hesselink JW, Bauersachs J, Blomström-Lundqvist C, Cífková R, De Bonis M, Iung B, Johnson MR, Kintscher U, Kranke P, Lang IM, Morais J, Pieper PG, Presbitero P, Price S, Rosano GMC, Seeland U, Simoncini T, Swan L, Warnes CA;ESC Scientific Document Group. 2018 ESC guidelines for the management of cardiovascular diseases during pregnancy. *Eur Heart J* 2018;**39**:3165–3241.
3. Baumgartner H, Bonhoeffer P, De Groot NM, de Haan F, Deanfield JE, Galie N, Gatzoulis MA, Gohlke-Baerwolf C, Kaemmerer H, Kilner P, Meijboom F, Mulder BJ, Oechslin E, Oliver JM, Serraf A, Szatmari A, Thaulow E, Vouhe PR, Walma E. ESC guidelines for the management of grown-up congenital heart disease (new version 2010). *Eur Heart J* 2010;**31**:2915–2957.
4. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of go with deep neural networks and tree search. *Nature* 2016;**529**:484–489.
5. Kelleher K. Deepmind's ai is now beating humans at quake because winning in go wasn't terrifying enough. https://www.businessinsider.com/deepmind-ai-beat-humans-quake-2018-7 (5 February 2019).
6. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. *Nature* 2015;**518**:529–533.
7. Jaderberg M, Czarnecki WM, Dunning I, Marris L, Lever G, Garcia Castaneda A, Beattie C, Rabinowitz NC, Morcos AS, Ruderman A, Sonnerat N, Green T, Deason L, Leibo JZ, Silver D, Hassabis D, Kavukcuoglu K, Graepel T. Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. ArXiv e-prints. 2018.

8. Tison GH, Sanchez JM, Ballinger B, Singh A, Olgin JE, Pletcher MJ, Vittinghoff E, Lee ES, Fan SM, Gladstone RA, Mikell C, Sohoni N, Hsieh J, Marcus GM. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiol* 2018;**3**:409–416.
9. Bello GA, Dawes TJ, Duan J, Biffi C, de Marvao A, Howard LS, Gibbs JSR, Wilkins MR, Cook SA, Rueckert D. Deep learning cardiac motion analysis for human survival prediction. arXiv preprint arXiv:1810.03382. 2018.
10. Al'Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, Pandey M, Maliakal G, van Rosendael AR, Beecy AN, Berman DS, Leipsic J, Nieman K, Andreini D, Pontone G, Schoepf UJ, Shaw LJ, Chang H-J, Narula J, Bax JJ, Guan Y, Min JK. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J* 2018; doi:10.1093/eurheartj/ehy404.
11. Heden B, Ohlsson M, Rittner R, Pahlm O, Haisty WK Jr, Peterson C, Edenbrandt L. Agreement between artificial neural networks and experienced electrocardiographer on electrocardiographic diagnosis of healed myocardial infarction. *J Am Coll Cardiol* 1996;**28**:1012–1016.
12. Vasquez C, Hernandez A, Mora F, Carrault G, Passariello G. Atrial activity enhancement by wiener filtering using an artificial neural network. *IEEE Trans Biomed Eng* 2001;**48**:940–944.
13. Vos JE, Scheepstra KA. Computer-simulated neural networks: an appropriate model for motor development? *Early Hum Dev* 1993;**34**:101–112.
14. Webber WR, Litt B, Wilson K, Lesser RP. Practical detection of epileptiform discharges (EDs) in the EEG using an artificial neural network: a comparison of raw and parameterized EEG data. *Electroencephalogr Clin Neurophysiol* 1994;**91**: 194–204.
15. Narayan SM. Restricted connectivities in neural networks. Masters of Science Thesis. University of Birmingham, UK, Department of Computer Science 1990. p1–153.
16. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol* 2017;**69**:2657–2664.
17. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. *J Artif Intell Res* 1996;**4**:237–285.
18. Ribeiro MT, Singh S, Guestrin C. Model-agnostic interpretability of machine learning. The Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY.
19. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, Ashley E, Dudley JT. Artificial intelligence in cardiology. *J Am Coll Cardiol* 2018;**71**: 2668–2679.
20. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (NIPS) 2012, Harrah's Lake Tahoe, Stateline, NV.
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.
22. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. 2017. International Conference on Learning Representations 2016, San Juan, Puerto Rico.
23. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV.
24. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010; **22**:1345–1359.
25. Afshar P, Mohammadi A, Plataniotis KN. Brain tumor type classification via capsule networks. ArXiv e-prints 2018. 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece.
26. Lyon A, Ariga R, Minchole A, Mahmod M, Ormondroyd E, Laguna P, de Freitas N, Neubauer S, Watkins H, Rodriguez B. Distinct ECG phenotypes identified in hypertrophic cardiomyopathy using machine learning associate with arrhythmic risk markers. *Front Physiol* 2018;**9**:213.
27. Rubin DB. Inference and missing data. *Biometrika* 1976;**63**:581–592.
28. Siddique J, Harel O, Crespi CM. Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal clinical trial. *Ann Appl Stat* 2012;**6**:1814–1837.
29. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med Inform* 2018;**6**:e11.
30. Falissard L, Fagherazzi G, Howard N, Falissard B. Deep clustering of longitudinal data. arXiv preprint arXiv:1802.03212. 2018.
31. Karch J, A machine learning perspective on repeated measures. http://edoc.hu-berlin.de/18452/18293 (5 February 2019).
32. Hinton GF. A parallel computation that assigns canonical object-based frames of reference. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, BC, Canada. Vol. 2. 1981. p683–685.
33. Sabour S, Frosst N, E Hinton G. Dynamic routing between capsules. Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, CA.
34. Al-Shedivat M, Bansal T, Burda Y, Sutskever I, Mordatch I, Abbeel P. Continuous adaptation via meta-learning in nonstationary and competitive environments. ICLR 2018: International Conference on Learning Representations, Vancouver, BC, Canada.
35. Peng XB, Andrychowicz M, Zaremba W, Abbeel P. Sim-to-real transfer of robotic control with dynamics randomization. 2018. IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia.
36. Frans K, Ho J, Chen X, Abbeel P, Schulman J. Meta learning shared hierarchies. ICLR 2018: International Conference on Learning Representations, Vancouver, BC, Canada. eprint arXiv:1710.09767. 2017:arXiv:1710.09767.
37. Alemi AA, Chollet F, Een N, Irving G, Szegedy C, Urban J. Deepmath—deep sequence models for premise selection. 2016 Advances in Neural Information Processing Systems (NIPS 2016), Barcelona, Spain. ArXiv e-prints.
38. Bisgin H, Bera T, Ding H, Semey HG, Wu L, Liu Z, Barnes AE, Langley DA, Pava-Ripoll M, Vyas HJ, Tong W, Xu J. Comparing SVM and ANN based machine learning methods for species identification of food contaminating beetles. *Sci Rep* 2018;**8**:6532.
39. Jodas DS, Marranghello N, Pereira AS, Guido RC. Comparing support vector machines and artificial neural networks in the recognition of steering angle for driving of mobile robots through paths in plantations. *Proc Comput Sci* 2013;**18**: 240–249.
40. Abdullah A, Veltkamp RC, Wiering MA, An ensemble of deep support vector machines for image categorization. 2009 International Conference of Soft Computing and Pattern Recognition (SoCPaR 2009), Malacca, Malaysia.
41. Li Y, Zhang T. Deep neural mapping support vector machines. *Neural Netw* 2017;**93**:185–194.
42. Rubin J, Abreu R, Ganguli A, Nelaturi S, Matei I, Sricharan K. Recognizing abnormal heart sounds using deep learning. The 25th International Joint Conference on Artificial Intelligence IJCAI16, New York, NY. arXiv preprint arXiv:1707.04642. 2017.
43. Luong C, Abdi A, Jue J, Gin K, Fleming S, Abolmaesumi P, Tsang T. Automatic quality assessment of echo apical 4-chamber images using computer deep learning. 2016. The American Heart Association's Scientific Sessions 2017 in Anaheim, CA.
44. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med* 2018;**1**:6.
45. Bernard O, Bosch JG, Heyde B, Alessandrini M, Barbosa D, Camarasu-Pop S, Cervenansky F, Valette S, Mirea O, Bernier M, Jodoin P-M, Domingos JS, Stebbing RV, Keraudren K, Oktay O, Caballero J, Shi W, Rueckert D, Milletari F, Ahmadi S-A, Smistad E, Lindseth F, van Stralen M, Wang C, Smedby O, Donal E, Monaghan M, Papachristidis A, Geleijnse ML, Galli E, D'hooge J. Standardized evaluation system for left ventricular segmentation algorithms in 3D echocardiography. *IEEE Trans Med Imaging* 2016;**35**:967–977.
46. Keraudren K, Oktay O, Shi W, Hajnal JV, Rueckert D. Endocardial 3D ultrasound segmentation using autocontext random forests. In: Proceedings MICCAI Challenge on Echocardiographic Three-Dimensional Ultrasound Segmentation (CETUS), Boston, MA. 2014. p.41–48.
47. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, Cook SA, de Marvao A, Dawes T, O'Regan DP, Kainz B, Glocker B, Rueckert D. Anatomically constrained neural networks (ACNNS): application to cardiac image enhancement and segmentation. *IEEE Trans Med Imaging* 2018;**37**: 384–395.
48. Xia Y, Wulan N, Wang K, Zhang H. Detecting atrial fibrillation by deep convolutional neural networks. *Comput Biol Med* 2018;**93**:84–92.
49. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, Ng AY. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine* 2019;**25**: 65–69.
50. Pyakillya B, Kazachenko N, Mikhailovsky N. Deep learning for ecg classification. *J Phys Conf Ser* 2017;**913**:012004.
51. Shashikumar SP, Shah AJ, Li Q, Clifford GD, Nemati S, A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. In: *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. Orlando, FL. 2017. p.141–144.
52. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak JAWM, Hermsen M, Manson QF, Balkenhol M, Geessink O, Stathonikos N, van Dijk MC, Bult P, Beca F, Beck AH, Wang D, Khosla A, Gargeya R, Irshad H, Zhong A, Dou Q, Li Q, Chen H, Lin H-J, Heng P-A, Haß C, Bruni E, Wong Q, Halici U, Öner MÜ, Cetin-Atalay R, Berseth M, Khvatkov V, Vylegzhanin A, Kraus O, Shaban M, Rajpoot N, Awan R, Sirinukunwattana K, Qaiser T, Tsang Y-W, Tellez D, Annuscheit J, Hufnagl P, Valkonen M, Kartasalo K, Latonen L, Ruusuvuori P, Liimatainen K, Albarqouni S, Mungal B, George A, Demirci S, Navab N, Watanabe S, Seno S, Takenaka Y, Matsuda H, Ahmady Phoulady H, Kovalev V, Kalinovsky A, Liauchuk V, Bueno G, Fernandez-Carrobles MM, Serrano I, Deniz O, Racoceanu D, Venâncio R. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;**318**:2199–2210.

53. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**542**:115.
54. Mundhra D, Cheluvaraju B, Rampure J, Rai Dastidar T, Analyzing microscopic images of peripheral blood smear using deep learning. In: Jorge Cardoso (ed.), *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2017. p178–185.
55. Zreik M, Lessmann N, van Hamersvelt RW, Wolterink JM, Voskuil M, Viergever MA, Leiner T, Išgum I. Deep learning analysis of the myocardium in coronary CT angiography for identification of patients with functionally significant coronary artery stenosis. *Med Image Anal* 2018;**44**:72–85.
56. Betancur J, Commandeur F, Motlagh M, Sharir T, Einstein AJ, Bokhari S, Fish MB, Ruddy TD, Kaufmann P, Sinusas AJ, Miller EJ, Bateman TM, Dorbala S, Di Carli M, Germano G, Otaki Y, Tamarappoo BK, Dey D, Berman DS, Slomka PJ. Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study. *JACC Cardiovasc Imaging* 2018;**11**:1654–1663.
57. Motwani M, Dey D, Berman DS, Germano G, Achenbach S, Al-Mallah MH, Andreini D, Budoff MJ, Cademartiri F, Callister TQ, Chang HJ, Chinnaiyan K, Chow BJ, Cury RC, Delago A, Gomez M, Gransar H, Hadamitzky M, Hausleiter J, Hindoyan N, Feuchtner G, Kaufmann PA, Kim YJ, Leipsic J, Lin FY, Maffei E, Marques H, Pontone G, Raff G, Rubinshtein R, Shaw LJ, Stehli J, Villines TC, Dunning A, Min JK, Slomka PJ. Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis. *Eur Heart J* 2017;**38**:500–507.
58. Li X, Liu H, Yang J, Xie G, Xu M, Yang Y. Using machine learning models to predict in-hospital mortality for ST-elevation myocardial infarction patients. *Stud Health Technol Inform* 2017;**245**:476–480.
59. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering* 2018;**2**:158–164.
60. Bumgarner JM, Lambert CT, Hussein AA, Cantillon DJ, Baranowski B, Wolski K, Lindsay BD, Wazni OM, Tarakji KG. Smartwatch algorithm for automated detection of atrial fibrillation. *J Am Coll Cardiol* 2018;**71**:2381–2388.
61. Cherkassky V. The nature of statistical learning theory. *IEEE Trans Neural Netw* 1997;**8**:1564.
62. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, Lee AM, Aung N, Lukaschuk E, Sanghvi MM, Zemrak F. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *J Cardiovasc Magn Reson* 2018;**20**:6.
63. Avendi M, Kheradvar A, Jafarkhani H. A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac MRI. *Med Image Anal* 2016;**30**:108–119.
64. Luo G, Dong S, Wang K, Zuo W, Cao S, Zhang H. Multi-views fusion CNN for left ventricular volumes estimation on cardiac MR images. *IEEE Trans Biomed Eng* 2018;**65**:1924–1934.
65. Oktay O, Bai W, Lee M, Guerrero R, Kamnitsas K, Caballero J, de Marvao A, Cook S, O'Regan D, Rueckert D. Multi-input cardiac image super-resolution using convolutional neural networks. MICCAI 2016, the 19th International Conference on Medical Image Computing and Computer Assisted Intervention, Athens, Greece.
66. Dong S, Luo G, Wang K, Cao S, Mercado A, Shmuilovich O, Zhang H, Li S. Voxelatlasgan: 3D left ventricle segmentation on echocardiography with atlas guided generation and voxel-to-voxel discrimination. MICCAI 2018, the 21st International Conference on Medical Image Computing and Computer Assisted Intervention, Granada, Spain. arXiv preprint arXiv:1806.03619. 2018.
67. Gao X, Li W, Loomes M, Wang L. A fused deep learning architecture for viewpoint classification of echocardiography. *Information Fusion* 2017;**36**:103–113.
68. Knackstedt C, Bekkers SC, Schummers G, Schreckenberg M, Muraru D, Badano LP, Franke A, Bavishi C, Omar AM, Sengupta PP. Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the FAST-EFs multicenter study. *J Am Coll Cardiol* 2015;**66**:1456–1466.
69. Nirschl JJ, Janowczyk A, Peyster EG, Frank R, Margulies KB, Feldman MD, Madabhushi A. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. *PLoS One* 2018;**13**:e0192726.
70. Seah JCY, Tang JSN, Kitchen A, Gaillard F, Dixon AF. Chest radiographs in congestive heart failure: visualizing neural network learning. *Radiology* 2018;**290**: 514–522.
71. Lieman-Sifry J, Le M, Lau F, Sall S, Golden D, Fastventricle: cardiac segmentation with ENET. In: *International Conference on Functional Imaging and Modeling of the Heart*. Toronto, ON, Canada, 2017. p127–138
72. Pyrkov TV, Slipensky K, Barg M, Kondrashin A, Zhurov B, Zenin A, Pyatnitskiy M, Menshikov L, Markov S, Fedichev PO. Extracting biological age from biomedical data via deep learning: too much of a good thing? *Sci Rep* 2018;**8**:5210.
73. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, Bonow RO, Huang CC, Deo RC. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation* 2015;**131**:269–279.
74. Miller K, Hettinger C, Humpherys J, Jarvis T, Kartchner D. Forward thinking: building deep random forests. arXiv preprint arXiv:1705.07366. 2017.
75. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017;**24**: 361–370.
76. Medved D, Ohlsson M, Höglund P, Andersson B, Nugues P, Nilsson J. Improving prediction of heart transplantation outcome using deep learning techniques. *Sci Rep* 2018;**8**:3613.
77. Voors AA, Anker SD, Cleland JG, Dickstein K, Filippatos G, van der Harst P, Hillege HL, Lang CC, Ter Maaten JM, Ng L, Ponikowski P, Samani NJ, van Veldhuisen DJ, Zannad F, Zwinderman AH, Metra M. A systems biology study to tailored treatment in chronic heart failure: rationale, design, and baseline characteristics of BIOSTAT-CHF. *Eur J Heart Fail* 2016;**18**:716–726.
78. Couderc JP, Kyal S, Mestha LK, Xu B, Peterson DR, Xia X, Hall B. Detection of atrial fibrillation using contactless facial video monitoring. *Heart Rhythm* 2015; **12**:195–201.
79. Díaz DH, Ó C, Pallas-Areny R. Heart rate detection from single-foot plantar bioimpedance measurements in a weighing scale. *Conf Proc IEEE Eng Med Biol Soc* 2010;**2010**:6489–6492.
80. Shcherbina A, Mattsson CM, Waggott D, Salisbury H, Christle JW, Hastie T, Wheeler MT, Ashley EA. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med* 2017;**7**:2.
81. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, Castella M, Diener H-C, Heidbuchel H, Hendriks J, Hindricks G, Manolis AS, Oldgren J, Popescu BA, Schotten U, Van Putte B, Vardas P, Agewall S, Camm J, Baron Esquivias G, Budts W, Carerj S, Casselman F, Coca A, De Caterina R, Deftereos S, Dobrev D, Ferro JM, Filippatos G, Fitzsimons D, Gorenek B, Guenoun M, Hohnloser SH, Kolh P, Lip GYH, Manolis A, McMurray J, Ponikowski P, Rosenhek R, Ruschitzka F, Savelieva I, Sharma S, Suwalski P, Tamargo JL, Taylor CJ, Van Gelder IC, Voors AA, Windecker S, Zamorano JL, Zeppenfeld K. 2016 ESC guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *Eur Heart J* 2016;**37**:2893–2962.
82. Inohara T, Shrader P, Pieper K, Blanco RG, Thomas L, Singer DE, Freeman JV, Allen LA, Fonarow GC, Gersh B, Ezekowitz MD, Kowey PR, Reiffel JA, Naccarelli GV, Chan PS, Steinberg BA, Peterson ED, Piccini JP. Association of atrial fibrillation clinical phenotypes with treatment patterns and outcomes: a multicenter registry study. *JAMA Cardiol* 2018;**3**:54–63.
83. Marblestone AH, Wayne G, Kording KP. Toward an integration of deep learning and neuroscience. *Front Comput Neurosci* 2016;**10**:94.
84. Jansche M, Maximum expected f-measure training of logistic regression models. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada, 2005. p.692–699.
85. Yuan X, He P, Zhu Q, Li X. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst 2019;doi: 10.1109/TNNLS.2018.2886017.
86. Elsayed GF, Shankar S, Cheung B, Papernot N, Kurakin A, Goodfellow I, Sohl-Dickstein J. Adversarial examples that fool both computer vision and time-limited humans. 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada. ArXiv e-prints 2018.
87. Papernot N, McDaniel P, Goodfellow I, Jha S, Berkay Celik Z, Swami A. Practical black-box attacks against machine learning. ASIA CCS '17 Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates. ArXiv e-prints 2016.
88. Meng D, Chen H. Magnet: a two-pronged defense against adversarial examples. CCS '17 Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, Texas, USA. ArXiv e-prints 2017.
89. Olson RS, La Cava W, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput* 2018;**23**:192–203.
90. Voets M. Deep learning: from data extraction to large-scale analysis. Master's thesis. UiT Norges arktiske universitet, 2018.
91. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016; **316**:2402–2410.
92. Steyerberg EW, Harrell FE Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001;**54**:774–781.
93. Hand DJ. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning* 2009;**77**:103–123.
94. Gong G, Quante AS, Terry MB, Whittemore AS. Assessing the goodness of fit of personal risk models. *Stat Med* 2014;**33**:3179–3190.
95. Gerds TA, Andersen PK, Kattan MW. Calibration plots for risk prediction models in the presence of competing risks. *Stat Med* 2014;**33**:3191–3203.

96. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): explanation and elaboration. *Ann Intern Med* 2015;**162**:W1–73.
97. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;**35**: 1925–1931.
98. Van Driest SL, Wells QS, Stallings S, Bush WS, Gordon A, Nickerson DA, Kim JH, Crosslin DR, Jarvik GP, Carrell DS, Ralston JD, Larson EB, Bielinski SJ, Olson JE, Ye Z, Kullo IJ, Abul-Husn NS, Scott SA, Bottinger E, Almoguera B, Connolly J, Chiavacci R, Hakonarson H, Rasmussen-Torvik LJ, Pan V, Persell SD, Smith M, Chisholm RL, Kitchner TE, He MM, Brilliant MH, Wallace JR, Doheny KF, Shoemaker MB, Li R, Manolio TA, Callis TE, Macaya D, Williams MS, Carey D, Kapplinger JD, Ackerman MJ, Ritchie MD, Denny JC, Roden DM. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA* 2016;**315**:47–57.
99. Krittanawong C, Bomback AS, Baber U, Bangalore S, Messerli FH, Wilson Tang WH. Future direction for using artificial intelligence to predict and manage hypertension. *Curr Hypertens Rep* 2018;**20**:75.
100. Krittanawong C. Future physicians in the era of precision cardiovascular medicine. *Circulation* 2017;**136**:1572–1574.
101. Lardinois F. Andrew ng officially launches his $175m ai fund. https://techcrunch.com/2018/01/30/andrew-ng-officially-launches-his-175m-ai-fund (5 February 2019).
102. Krittanawong C, Zhang HJ, Aydar M, Wang Z, Sun T. Crowdfunding for cardiovascular research. *Int J Cardiol* 2018;**250**:268–269.
103. Krittanawong C, Johnson KW, Hershman SG, Tang WHW. Big data, artificial intelligence, and cardiovascular precision medicine. *Expert Rev Precis Med Drug Dev* 2018;**3**:305–317.
104. Krittanawong C, Kitai T. Identifying genotypes and phenotypes of cardiovascular diseases using big data analytics. *JAMA Cardiol* 2017;**2**:1169–1170.
105. Islam MT, Aowal MA, Minhaz AT, Ashraf K. Abnormality detection and localization in chest x-rays using deep convolutional neural networks. arXiv preprint arXiv:1705.09850. 2017.