Journal Pre-proof

Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure

Cameron R. Olsen, Robert J. Mentz, Kevin J. Anstrom, David Page, Priyesh A. Patel

PII: S0002-8703(20)30215-5

DOI: https://doi.org/10.1016/j.ahj.2020.07.009

Reference: YMHJ 6194

To appear in: *American Heart Journal*

Received date: 14 April 2020

Accepted date: 8 July 2020

Please cite this article as: C.R. Olsen, R.J. Mentz, K.J. Anstrom, et al., Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure, *American Heart Journal* (2020), https://doi.org/10.1016/j.ahj.2020.07.009

Title: Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure

Short title: Machine learning in heart failure

Cameron R. Olsen, M.D.*[1], Robert J. Mentz, M.D.[1], Kevin J. Anstrom, Ph.D., [2], David Page, Ph.D.[2], Priyesh A. Patel, M.D.[3]

[1]Division of Cardiology, Duke University Medical Center, Durham, North Carolina, USA

[2]Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA

[3]Sanger Heart and Vascular Institute, Atrium Health, Charlotte, North Carolina, USA

*Corresponding author information: Cameron R. Olsen, M.D., , 8 Searle Center Dr, Durham, NC 27710, USA, Fax: 919-668-7063 (cameron.olsen@duke.edu)

Abstract:
Machine learning and artificial intelligence are generating significant attention in the scientific community and media. Such algorithms have great potential in medicine for personalizing and improving patient care, including in the diagnosis and management of heart failure. Many physicians are familiar with these terms and the excitement surrounding them, but many are unfamiliar with the basics of these algorithms and how they are applied to medicine. Within heart failure research, current applications of machine learning include creating new approaches to diagnosis, classifying patients into novel phenotypic groups, and improving prediction capabilities. In this paper, we provide an overview of machine learning targeted for the practicing clinician and evaluate current applications of machine learning in the diagnosis, classification, and prediction of heart failure.

Keywords:
Heart failure; Machine learning; Artificial Intelligence; Deep Learning; Diagnosis; Prediction

Word Count: Abstract, 122; Text, 6982

Introduction

More than 1,000,000 Americans are diagnosed with heart failure in the United States each year, and an estimated 6.2 million individuals in the U.S. currently live with heart failure.[1] Rapid diagnosis and risk assessment of heart failure are essential to providing timely, cost-effective care[2]. Traditional risk prediction tools have modest predictive power,[3-5] and the complexity of heart failure pathophysiology and treatment produces an array of information that is challenging for clinicians and researchers to process. Artificial intelligence (AI) and machine learning (ML) techniques offer a potential solution to organizing and interpreting these complex and ever-increasing data[6-8].

AI and ML have benefited from sophisticated technology and improving computer processing speeds, leading to the creation of algorithms for autonomous driving,[9] image and facial recognition,[10, 11] automatic suggestion software, and natural language processing.[12] Within medicine, some of the most notable successes have come in areas relying on image-based diagnosis, such as the diagnosis of diabetic retinopathy from retinal fundus photographs[13], the diagnosis of skin cancer[14], and certain imaging diagnoses such as pneumonia[15] and breast cancer[16]. The goal of applying machine learning to medicine is to improve detection and classification of disease, make better predictions, and improve personalization of medicine[6, 17], and the same is true for the management of heart failure.

We herein review the current applications of machine learning within heart failure research. To begin, we will present a brief introduction of machine learning written for the practicing clinician or cardiologist, reviewing broad categories of machine learning, its uses, and some important pitfalls and considerations. Following this, we will review current literature covering the varied uses of machine learning specifically appertaining to heart failure – explicitly, those applications related to the diagnosis, classification, and prediction of heart failure. The objective of this review is to provide the practicing cardiologist a better understanding of what machine learning is, how others are implementing it, and what impact it will have on their current and future careers. We hope to impart a practical, easily accessible review for the practicing physician that is up to date with the latest clinical research[18, 19].

A brief introduction to machine learning for the clinician

Machine learning (ML) is a sub-area of artificial intelligence and consists of pattern-recognition algorithms used to define relationships and make predictions after training on selected datasets. The power of ML algorithms comes from the ability to automatically learn these patterns from large datasets for predictive analytics, allowing the user to glean knowledge from past data and apply it to future predictions. Machine learning as a field is a natural outgrowth of traditional statistics, and in many ways these have significant areas of overlap.[20] Traditional statistical models aim to make inferences about a population given a data sample and a set of assumptions about the data, whereas the focus of machine learning is to learn from prior experience to make more accurate predictions. Most biostatistical models currently used are based on parametric models that are linear in nature, which limits their ability to model complex relationships. Though the focus of statistical models is not strictly to make predictions, the models can and are often used to do so. For instance, logistic and linear regression are two statistical models that are commonly used in machine learning for predictive purposes. Many other ML algorithms are nonlinear in nature (e.g., decision trees or random forests) and are used to model more abstract relationships between the data. Likewise, the reintroduction of what are now termed deep learning methods in the early 2000's has led to the implementation of multi-

layered neural network algorithms[21], which use the data to automatically construct new, useful features from lower-level features in earlier layers (see Deep Learning section below). In addition to their ability to learn in highly complex and nonlinear settings, ML algorithms are also well suited for handling big data, and the implementation of electronic health records within health systems has led to large and valuable datasets upon which to apply these techniques[17, 22]. However, applying machine learning algorithms to health data requires careful consideration, including a knowledge of appropriate use and important drawbacks[23, 24]. Here, we will introduce important considerations for implementing machine learning algorithms and discuss several broad groups of machine learning techniques. For further study, we refer the reader to more detailed papers[22, 25-28].

Initial Considerations for Using Machine Learning Algorithms

When developing and training a machine learning algorithm, the developer faces many choices before they can put a model together. The first steps include creating a data set that is clean and accessible by the algorithm. For many algorithms, the user must define which features or variables (e.g., systolic blood pressure) are the most relevant to their problem and hence those which should be included in their algorithm (termed feature selection). Data must also be processed to combine redundant data and minimize the number of features to improve model performance and outcomes (termed feature engineering). Selecting for and engineering features allows the researcher to limit the number of features available to the model and thus avoid the "curse of dimensionality," a phenomenon where increasingly more data is needed to reach statistical significance as more dimensions, or features, are used by the model. Intentionally eliminating features may come at the cost of some level of accuracy, yet doing so can also lead to decreased model complexity and lowers the chance that a model will be over-fitted to an insufficiently sized training set[29]. In addition to selecting and engineering features, clinical databases often contain missing values or are incomplete, and clinicians and data scientists must decide how to handle the missing data (e.g. through removing incomplete records or imputing values for missing data, among many options[30, 31]) while preventing the excessive loss of data and maximizing both applicability and accuracy of their model. When possible, increasing the number of data samples often decreases prediction error and improves performance[32], but gathering more data is not always practical nor cost-effective in medical research. Once these issues have been addressed, the researcher must also decide which approach they will use in analyzing the data. When choosing between algorithms, researchers must consider the type of problem they are trying to solve (e.g. regression vs. classification), how to partition the data for training and testing purposes, and how to balance model complexity with accuracy and interpretability. Here we will introduce two broad categories of algorithms, namely supervised and unsupervised learning, and will touch on the additional topic of deep learning (which can be both supervised or unsupervised). These approaches to machine learning do not encompass all of machine learning, and there are many techniques that do not fit neatly within the supervised-unsupervised paradigm, such as reinforcement learning, active learning, and one-shot learning. As techniques such as these are not common in heart failure research, we include them for reference and refer the reader for further study[33-35].

Supervised Learning

One of the first classifications of ML algorithms separates them according to the presence or absence of supervised training, leading to two main groups of algorithms: supervised and

unsupervised. The reference to supervision indicates whether data have been "labelled" (i.e. associated) with the true response or outcome of interest, and algorithms that use labelled data are termed supervised learning algorithms[36]. For example, in a supervised algorithm aimed at predicting the presence of heart failure from a patient's echocardiogram, the echocardiographic data is labeled as having come from a patient with or without heart failure. These algorithms are commonly used for predictive analyses, which include making diagnoses and predicting outcomes. There are many different types of supervised learning algorithms, and despite their strengths and weaknesses, these algorithms tend to perform more similarly than not[37]. A few common examples of supervised algorithms include linear and logistic regression, support vector machines (SVM), decision trees, ensemble methods (e.g. random forests), k-nearest neighbor (k-NN), and naïve Bayes classifiers[37].

Similar to other predictive algorithms, ML predictive abilities are often described with a receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC)[38-40]. Most ML algorithms produce a model that can be used to rank test examples (e.g., patients) from most likely to be a positive example (e.g., most likely to suffer heart failure) to least likely; often the ranking results from the model assigning to each example a probability of being positive. An ROC curve is created using the true positive rates (TPR, or sensitivity) and the false positive rates (FPR, or 1 – specificity) calculated at differing cuts in the ranking, e.g. at different probability thresholds above which examples are predicted to be positive and below which negative.

Though the AUC has become nearly universal in evaluating ML algorithms, one must be careful when using it in situations where there is an imbalance between the number of positive and negative cases, as the AUC may inflate the true predictive capabilities.[41, 42] For example, a diagnostic test or predictive model might have an AUC of 0.97, but if the event is very rare then the positive predictive value (PPV, or precision) could still be under 0.1; this phenomenon has been observed from mammography to internet search. To get a sense of how this can happen, consider a condition occurring in only ten of the 10,000 patients we see. Suppose at one threshold we have a TPR of 0.9 and an FPR of only 0.01; then at that threshold we correctly identify the condition in 9 of the ten patients having the condition, but we falsely claim that 100 of the patients without the condition have it. Therefore the PPV or precision is only 9/109. Even though in ROC space the point having TPR of 0.9 and FPR of 0.01 looks quite good, less than 10% of patients flagged for the condition actually have it.

Motivated by the above discussion, another measure of performance is the area under the precision-recall curve (AUPRC). Here, recall is another term for TPR, whereas precision is equal to the positive predictive value (PPV). Though it has been used less frequently than the ROC, one advantage of using a PRC is that it can give a more realistic assessment in situations where the dataset is imbalanced (e.g. with a low number of true positives)[41, 42]. Both a strength and weakness of PRCs is that they are sensitive to class imbalance and hence show its effect; this is the reason they are widely used in problems such as internet search where the number of relevant web pages for most searches is extremely low. It has been shown that for a fixed task or data set, and hence fixed class imbalance, one can convert between points in ROC space and points in PR space, and hence between ROC curves and PR curves[42]. AUCPR is also equivalent to Average Precision (PPV averaged over all recall values) if dense enough sampling of recalls is used. Different specific operating points (corresponding to different thresholds) in ROC or PR space also yield other commonly used measures such as F_1 score or F_β. F_1 is the geometric mean of precision and recall, and other values of the subscript weight precision and recall differently in

taking the mean. Like AUCPR, for better or worse F scores are also altered by the magnitude of class imbalance. Other measures of performance exist within machine learning such as the Matthews correlation coefficient or the cost curve, but as these were not commonly implemented in the articles reviewed herein, we refer the reader to other sources for further information[41, 43]. Regardless of the metric chosen, it is of critical importance that the performance measurement should not be done on the data used in training the algorithm (i.e. the training data set) but on a held-aside test set (a.k.a. validation set) or by cross-validation. Doing so provides a form of internal validation or testing; even better would be performing external validation on an entirely new data set, ideally from a prospective study. Testing on separate data helps avoid the problem of over-fitting the data, an issue where the algorithm fits the training data so well that it loses generalizability to prospective data unseen by the trained model.

Unsupervised Learning

As opposed to supervised algorithms, unsupervised learning algorithms use "unlabeled" data, meaning there is no outcome or prediction label applied to the dataset[44]. Since there is no truth label assigned to the data, unsupervised algorithms are not used to make predictions or classify patients to predefined groups. Rather, they are used on large datasets where the goal is to discover and analyze relationships between hidden groups within the data. Here, the model itself discovers the patient grouping as opposed to pre-labeling them with known classes. Common applications of unsupervised algorithms include clustering analysis, where data is grouped according to similar characteristics; density estimation, where data is analyzed to estimate its probability distribution, as is done for anomaly detection; and dimensionality reduction, where the number of features or variables are reduced to a set of core features that capture the essence of the data while reducing linearity or redundancy between features. The last of these uses, dimensionality reduction, plays an important role in both data visualization and in circumventing the curse of dimensionality. In many instances, a researcher may use both unsupervised and supervised techniques to analyze a dataset, yet unsupervised algorithms can also be used independently, such as is common in heart failure research with the use of clustering algorithms. In such instances, patients are clustered into separate groups, which groups can be analyzed for their differences in baseline characteristics, laboratory data, or outcomes. Lastly, some "semi-supervised" learning algorithms use a mixture of labeled and unlabeled data, or use "weakly labeled" data (i.e. data labeled by some noise-prone, often automated, process), and hence can make predictions without a complete set of labeled data.[45, 46]

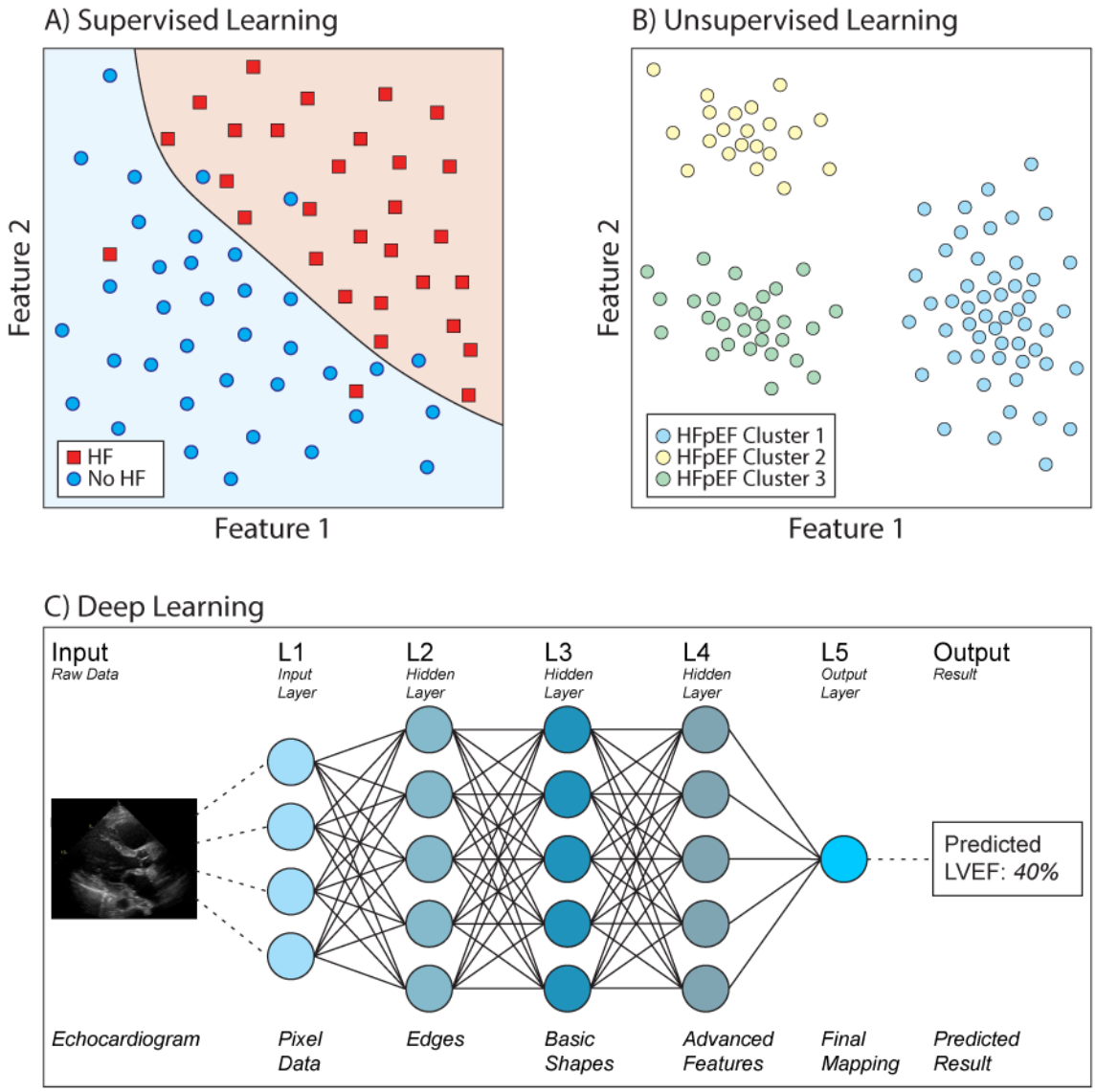

Figure 1. Sample visualizations of machine learning. A) Model of a supervised learning algorithm. Supervised algorithms train using data that has been labeled with the outcome of interest. Once trained, the algorithm will predict where a new data sample lies. In this example, if the two features used in this algorithm places a data point in the red area, the algorithm will predict that the patient will have heart failure (HF). B) Model of an unsupervised learning algorithm. Unsupervised algorithms train using data with no labeling and instead look for patterns or groups of data. In this example plot, patients with heart failure with preserved ejection fraction (HFpEF) are further classified into new phenotypes by an unsupervised clustering algorithm. C) Model of a deep learning algorithm. Deep learning algorithms incorporate multi-layer processing, and by definition have more than one layer between the input and output layers (i.e. hidden layers). This model demonstrates how each successive layer builds increasingly complex relationships before producing the final prediction. Wording along the bottom of the figure gives an example of what each layer could represent in an algorithm that analyzes pixel data from an echocardiogram to predict a patient's left ventricular ejection fraction (LVEF).

Deep Learning

Recent developments in computer and data mining technology have led to the reintroduction of what have been termed deep learning algorithms[47], algorithms studied as early the 1980s but which fell out of favor due to the lack of the computational power available at the time[20]. "Deep learning" refers to having more than one computational layer between the inputs and outputs of the algorithm (i.e. hidden layers, see Figure 1), which allows for greater abstraction at the cost of speed and computational power[44]. Most deep learning methods learn using multi-layer neural networks, though other methods (such as restricted Boltzmann machines, a probabilistic graphical model) are available as well[27]. Using multi-layer processing (i.e. deep learning techniques) is common in both supervised and unsupervised approaches.

In many instances, deep learning techniques offer a powerful alternative to conventional ML, enabling the user to perform more complex analysis with less user input[44, 48]. A downside to deep learning techniques is they often require very large datasets for training and are computationally expensive[44]. Without validated and tuned model parameters, deep learning models often suffer from overfitting, especially when an adequately large data set is not available. One way in which researchers address small sample sizes is through the implementation of transfer learning, a type of learning often used in image-processing algorithms (which are typically deep learning algorithms). In transfer learning, an algorithm is trained on one dataset before being subsequently trained on a separate but related dataset (the dataset of interest) with the goal of improving model performance. For instance, a deep learning algorithm trained to analyze echocardiograms could first be trained on datasets with general images, where it could learn to identify common image features such as edges, shapes, and patterns, and then apply that learning to better characterize the echocardiogram database. Another downside to deep learning is that the resulting models can be more difficult to interpret than non-deep methods, as the complexity of the neural networks produces relationships that become unrecognizable[24]. This results in a model that is effective in predicting the desired outcome but fails to explain why such a result occurred. In many instances, the optimal algorithm is less complicated and more interpretable while still achieving a similar performance. As such, it is often advisable to begin analysis with a more conventional ML algorithm (i.e. non-deep) before moving on to a more complicated model (e.g. a deep learning algorithm). For example, tabular structured data often works better with gradient boosting machines (a form of supervised, non-deep learning), while computer vision is more suited for deep learning techniques.[11, 49]

Applications of Machine Learning

In the following sections, we explore the current applications of machine learning in heart failure. All studies discussed in this paper will be displayed for summary in Table 1, which represents a portion of the total literature on this topic. For a more complete table, we refer the reader to the table found in the supplementary information (Supplementary Table 1).

Diagnosing heart failure

Current methods for diagnosing heart failure rely upon a patient's history, their physical exam, and both laboratory and imaging data.[50] ML-based methods aim to improve diagnosis through leveraging data found from each of these areas, including electrocardiography, echocardiography, EHR data, and other sources.

Electrocardiography

The importance of electrocardiography (ECG) in the workup of heart failure is well established, though its main use is in identifying underlying etiologies or complicating factors, such as arrhythmias.[50, 51] As changes in electrical activity are due to physiologic changes, several groups have hypothesized that applying ML to these tracings will lead to more accurate detection of heart failure. Such an approach would facilitate screening for heart failure in both the outpatient setting and in the emergency department.

Some of the earlier studies using ECGs created algorithms to differentiate between healthy patients and those with heart failure, most using the same publicly-available, online ECG databases[52-55]. These databases consist of a small (156 patients total) and narrow set of patients that were either healthy and in normal sinus rhythm or had a known diagnosis of heart failure. One recent representative paper in this group of studies used a deep long short-term memory (LSTM) network to identify those patients with heart failure[56]. This study separated the available data into two databases based on severity of disease, following which it produced accuracies of 98.9% and 87.6% for detecting the presence or absence heart failure, which results were in line with other studies using the same databases. The focus of these studies was often the improvement and validation of ML methods, and the size and nature of these databases limit their application to clinical uses.

Recently, other groups have evaluated larger and more varied datasets for the detection of heart failure. In 2019, Attia et al. trained and validated a convolutional neural network on ECG data from 44,959 ECG-TTE pairs to screen for patients with asymptomatic left ventricular dysfunction and subsequently tested it on a separate set of 52,870 ECG-TTE pairs.[57] Using echocardiography as the standard for assessing ventricular function, the ML model predicted the presence of ventricular dysfunction, defined by an ejection fraction less than 35%, with an AUC of 0.93 and with a sensitivity, specificity, and accuracy of 86.3%, 85.7%, and 85.7%, respectively. In addition, patients who were falsely identified as having an abnormal ECG were shown to have a fourfold risk of developing dysfunction in the future as compared to those who had a normal ECG. The authors did not directly compare the predictive ability of this ECG-based model with that of the B-type natriuretic peptide (BNP), yet their results suggest that using machine learning on ECG data may outperform the predictive ability of BNP in the detection of LV dysfunction when compared with data reported in literature.[58] In the same year, Kwon et al. reported using a similar algorithm for detecting either HFrEF or HFmrEF.[59] They incorporated a total of 55,163 ECG-TTE pairs from two separate hospitals, using the first hospital for both training and internal validation (44,673 pairs) and the second hospital for external validation (10,490 pairs). For detection of HFrEF, they report AUCs for the internal and external validation sets of 0.843 and 0.889, which outperformed both LR (0.800 & 0.847) and a random forest algorithm (0.807 & 0.853). Results for these studies may be limited by selection bias due to the retrospective inclusion of patients with both a recent ECG and TTE, which population likely differs from a standard screening population.

Echocardiography

Echocardiography has been a standard tool for heart failure characterization and management for decades.[50, 60] However, echocardiography requires a skilled user and interpreter, and the interpretation of echocardiographic images has varying levels of subjectivity and inter-reader reliability.[61] ML techniques have shown excellent ability in medical image processing and analysis, such as in chest radiography[62], and there has been considerable interest in applying these same successes to echocardiography. In this section, we will focus our review to those

studies that are more applicable to heart failure (e.g. detecting HF, automated EF calculation) rather than those covering all aspects of echocardiography, such as view classification.

In 2018, Zhang et al. created a model that automatically evaluated over 14,000 echocardiographs to measure cardiac structure and function using a convolutional neural network.[63] Among many features of their algorithm, the model automatically calculated left ventricular ejection fraction (LVEF) and longitudinal strain (LS), which showed respective median absolute differences of 6% and 1.4% when compared with manual tracings. However, these measurements were subject to wide variability in how closely EF approximated the cardiologist-reported value. As is common in other methods for automated EF estimation, the authors utilized the biplane area-length method, which relies on detection of the endocardial borders. In a different approach, Asch et al. hypothesized that they could estimate EF while circumventing standard border detection techniques, incorporating a deep learning algorithm to do so[64]. The algorithm was trained and validated on >50,000 echocardiographic studies, whereas the testing set consisted of 99 studies that were each manually evaluated by three experts using the biplane technique, which average LVEF was taken as the agreed value. Automatically calculated EFs showed a mean absolute deviation of 2.9% and excellent agreement with the expert-derived values ($r = 0.95$). In addition, the algorithm was tested to detect EF $\leq$ 35%, which it did so with a sensitivity and specificity of 0.93 and 0.87.

Machine learning has also been applied to standard features in echocardiography reports for the detection and evaluation of heart failure, including HFpEF[65-67] (Table 1). For example, Reddy et al. used ML techniques to identify clinical and echocardiographic variables that were the most predictive of having a diagnosis of HFpEF.[65] Echocardiographic variables under consideration included features such as the estimated pulmonary artery filling pressure and E/e' ratio. Using 6 variables selected by their ML algorithm, they created a risk score that resulted in an AUC of 0.841 for the diagnosis of HFpEF. This was notably higher than the performance of the European Society of Cardiology's 2016 algorithm for HFpEF detection on the same set of patients (AUC of 0.672), which algorithm relies on variables such as BNP and other echocardiographic or imaging features.[50] Creating a risk score based on ML-selected features allows for easier incorporation into clinical use, as these scores often are based on easy-to-obtain features. However, they face similar requirements for external validation and generally have lower performance than a full-feature algorithm.

Identification of HF from EHR

With the widespread incorporation of electronic health records (EHR), physicians and health systems have access to an ever-increasing amount of clinical data. Improved identification of patients with heart failure, particularly those without a prior clinical diagnosis, could provide opportunities to detect patients early on and appropriately target resources and education to better management.

In 2016, Blecker et al. compared several supervised machine learning algorithms in identifying patients currently hospitalized with heart failure using patient demographics, labs, and other clinical data.[68] Their ML model outperformed the ability of multiple non-ML algorithms, which included current problem list, having a BNP of 500 pg/mL or higher, and a logistic regression model of routine clinical variables. They reported an AUC of 0.974, with a sensitivity of 83% and a positive predictive value of 90% for the identification of hospitalized HF patients. Other groups have incorporated natural-language processing (NLP) algorithms for identification of hospitalized HF patients,[69] where NLP is a subfield of AI that often relies on

supervised and deep learning techniques to process free-text (such as the text in unstructured medical reports). By incorporating NLP techniques into their identification algorithm, Evans et al. improved detection of hospitalized HF patients, leading to a sensitivity of 95.3% and PPV of 97.5%, an increase of greater than 10% sensitivity when compared with their previous algorithm using structured clinical data alone.[69]

Other Applications of ML for the Diagnosis of Heart Failure

Beyond the applications above, machine learning has also been implemented in other methods for the diagnosis of heart failure. For instance, Nirschl et al. instituted a deep learning algorithm to evaluate H&E stained images of cardiac tissue in post-mortem patients for the presence of heart failure[70]. Both their DL and RF algorithms outperformed the results of 2 clinical pathologists, with the DL algorithm producing an AUC of 0.989, as compared to an AUC of 0.960 for the RF algorithm. Liu et al. analyzed digital heart sound recordings to discriminate between HFpEF and healthy patients, with their best algorithm producing an accuracy, sensitivity, and specificity of 0.963, 0.955, and 0.971, respectively.[71] More recently, Marcinkiewics-Siemion et al. measured fasting serum metabolites from 67 patients with HFrEF[72]. Using a random forest algorithm, they identified metabolites that were the most predictive of HFrEF and used the top 8 to create a panel for detecting the disease. Despite their ML-led process, this panel did not perform any better than BNP alone (AUC of 0.85 vs. AUC of 0.92). Notwithstanding, such a result highlights the feasibility of ML to aid in producing such panels in the future.

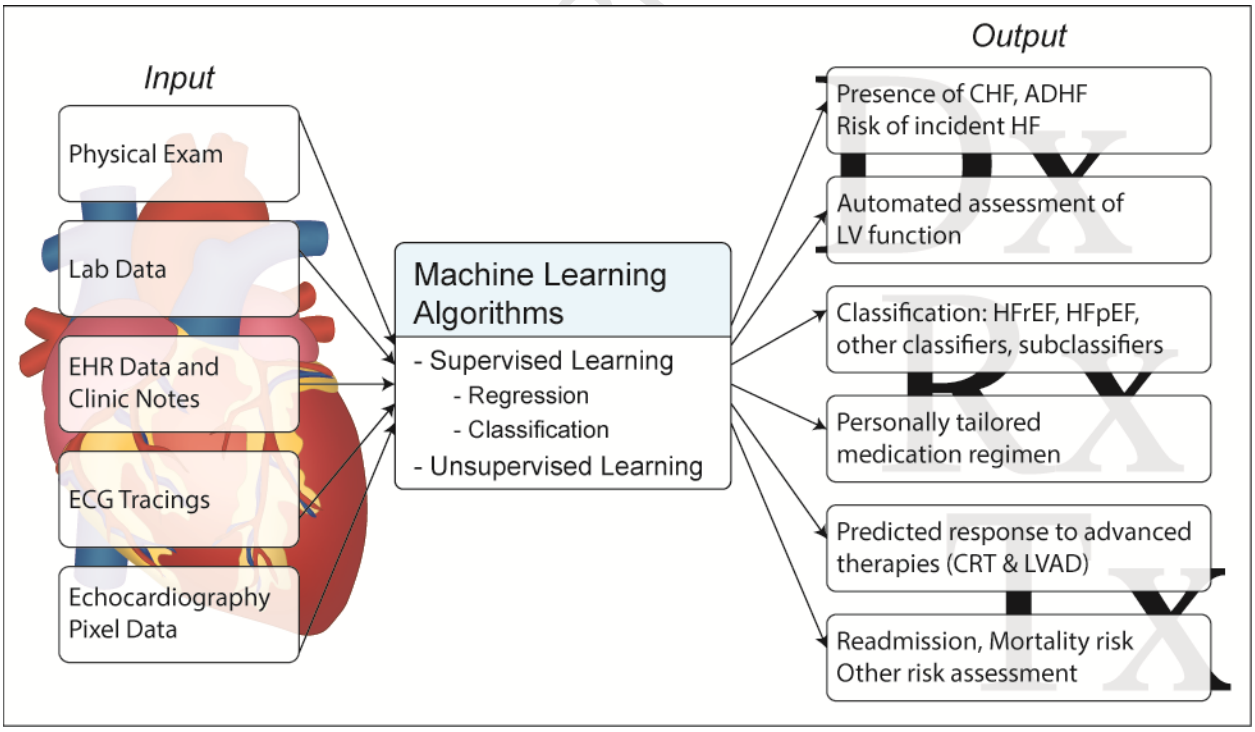

Figure 2. Machine learning in heart failure. This figure demonstrates the flow of data through machine learning algorithms to predict the diagnosis, classifications, and ideal treatments of patients with heart failure. In current literature, any combination of input data types is used with the proper machine learning algorithm to predict the specified outcome. Under the supervised learning category are included two common modes of supervised learning: regression and classification. Regression models are used for predicting a continuous variable (e.g. automatically assessing left-ventricular ejection fraction (LVEF) from an echocardiogram), whereas classification is used for predicting a specific class label, or categorical variable (e.g. the presence or

absence of heart failure using a patient's ECG tracing). Other examples of current uses of machine learning in heart failure are listed under the "output" column.

Precision medicine in heart failure

In addition to improving the diagnosis of heart failure, researchers have sought to leverage ML techniques to better classify and understand HF patients, with the goal of being able to tailor therapy to individual patients for improved outcomes[73-78] (Table 1). Areas in which heart failure research seeks to improve precision include stratifying patients with heart failure and identifying those who will benefit from advanced treatments, such as CRT and LVAD therapy.

Subtype classification

Traditionally, heart failure patients have been stratified according to LVEF, and effective treatments have been found for patients classified as having heart failure with reduced ejection fraction (HFrEF)[60] while the same success has evaded researchers for patients with HFpEF.[50, 79] In 2015, Shah et al. was one of the first to use an unsupervised algorithm to cluster patients into unique HFpEF phenotypes.[73] With standard lab, ECG, and echocardiographic data, they found that the patients with HFpEF settled into three groups based on cardiac structure and function, invasive hemodynamics, and outcomes. The validity of these three groups were tested on a separate patient cohort, which showed similar outcomes after patients were matched to the pre-defined groups. Subgrouping HFpEF has also been attempted by groups aiming to identify proteomic correlates of clusters[78] as well those incorporating pre- and post-exercise imaging data for further characterizations of groups[77]. Similar studies have likewise been performed on patients with HFrEF and acute decompensated heart failure.[75, 76] Each of these studies showed unique groups of patients with graduated outcomes depending on cluster assignment. Whether there is any clinical utility of these clusters in each instance has yet to be determined, and the value remains in whether changes in outcomes can be achieved by targeting therapies or prevention strategies to individual groups.

Using a different approach, Ahmad et al. sought to identify subgroups of HF patients with distinct phenotypes and survival outcomes independent of LVEF, ignoring the traditional classifications of HFrEF and HFpEF.[74] Using a cohort of 44,886 patients with heart failure, they implemented a supervised algorithm to identify the 8 most predictive features of 1-year survival, after which they implemented a clustering algorithm on those features to group patients into 4 distinct clusters. Assignment to a cluster had better discrimination for 1-year survival when compared to survival based on LVEF (AUC of 0.68 vs. 0.52). In addition, Ahmed et al. found that responses to pharmacotherapies differed across survival clusters. For example, patients in cluster 4 had a greater benefit from beta-blocker therapy than patients in cluster 1 (HR of 0.56 compared to 0.82), while all patient benefitted equally from angiotensin-converting enzyme inhibitors (ACE-I). These results suggest that it may be possible to apply ML algorithms to identify individuals who will benefit from tailored therapies, and for whom prognostication can be more accurately described, yet the utility of studies like this one will remain unknown until their results are verified by prospective study.

Response to advanced therapies

Cardiac resynchronization therapy (CRT) has been shown to decrease morbidity and mortality in appropriately selected patients, yet some patients show limited or no response to therapy.[80, 81] Since this discovery, interest has spiked in attempting to identify responders to

CRT,[80, 82, 83] including those using machine learning techniques (Table 1). Cikes et al. used a k-means clustering algorithm to compare the treatment effect of cardiac resynchronization therapy defibrillators (CRT-D) on the combined endpoint of all-cause death or any HF event.[84] Following analysis, four distinct patient groups were identified, two of which were found to have better responses to CRT-D, with hazard ratios for the primary outcome of 0.35 and 0.36, as compared to HRs of >0.8 for the other 2 groups. Another group of investigators used a supervised approach to predict mortality or readmission at 12 months following cardiac resynchronization therapy pacemaker (CRT-P) implantation.[85] They classified patients into 4 risk categories, where the highest risk category showed a hazard ratio of 7.96 for 12-month survival compared to the lowest risk category. Survival prediction in this study was notably better than prediction based on bundle branch block morphology and QRS duration, which combined did not reach significance for the hazard ratio (HR of 1.42, p=0.21). In 2019, Feeny et al. tested the ability of multiple ML algorithms (including SVM, RF, and LR) to identify responders to CRT and predict survival, comparing these to the American Heart Association/American College of Cardiology's 2013 heart failure guidelines[86]. The ML-based algorithms improved both response classification (AUC 0.70 vs. 0.65) and survival prediction (0.61 vs. 0.56) compared to the guidelines alone.

Left ventricular assist devices (LVADs) have likewise been shown to improve survival and increase quality of life.[87, 88] Similar to CRT, response to therapy is varied and current risk prediction tools for adverse outcomes such as RV failure and survival have modest predictive power. Kanwar et al. used a naïve-Bayes classifier (a supervised learning algorithm) to predict survival 1, 3, and 12 months following LVAD implantation.[89] Their model achieved an AUC of ~0.70, outperforming the HeartMate II Risk Score in 90-day and 1-year survival prediction (AUC ~0.60).[89, 90] Loghmanpour et al. likewise evaluated LVAD patients post-implant to explore predictors of right ventricular failure (RVF).[91] Their model outperformed previously published risk scores for RVF, including the right ventricular failure risk score (RVFRS)[92] and the Drakos score[93].

Incidence and outcome prediction

A major component of precision medicine includes making accurate predictions for patients based off personal characteristics and clinical data. Predicting the incidence and outcome of patients with heart failure is difficult, notwithstanding many risk scores reported in recent decades.[3, 4] In this section, we explore the application of machine learning to improving the prediction of heart failure onset and outcomes.

Predicting incidence of Heart Failure

Understanding factors leading to when a patient first experiences heart failure is key to providing preventative care and early treatment. In 2017, Ambale-Venktesh et al. used a random survival forest algorithm to predict cardiovascular outcomes, including incident heart failure, among patients who had been enrolled in the Multi-Ethnic Study of Atherosclerosis (MESA) study.[94] They identified the top-20 predictors for each outcome included in the study, which for the prediction of HF onset resulted in an AUC of 0.84, a mild improvement over the MESA-HF risk score (AUC of 0.80). Similarly, Segar et al. sought to predict incident HF hospitalization in patients with diabetes mellitus[95]. They trained and validated a random forests model on an internal cohort of 8,756 patients, which outperformed their best-performing Cox-based method (AUC of 0.77 to 0.73). Using the top variables from the RF model, they produce a risk score

with an AUC of 0.72, with an AUC of 0.70 following evaluation on an external validation cohort. In 2016, Choi et al. compared the effectiveness of deep learning techniques with logistic regression and conventional supervised algorithms in predicting heart failure onset.[96] Using 12-months of data prior to prediction, their deep-learning model mildly outperformed conventional models, producing an AUC of 0.78, compared to 0.75 for logistic regression and 0.76 for the best-performing conventional ML model. Chen et al. studied how the quantity and diversity of data affected the predictive performance of several different models, targeting the prediction of future HF status as the comparison tool[97]. As might be expected, increasing the amount and diversity of data led to improved predictive powers, with the full data set leading to an AUC of 0.791 for their recurrent neural network (RNN) model. This improvement was also seen in the other models, with the LR model performing similarly with an AUC of 0.786. Such small increases in performances may not be worth the increase in model complexity that comes with deep methods.

Mortality/Survival

For patients who already have a diagnosis of HF, mortality and survival predictions can help with understanding prognosis and targeting interventions. Many groups have investigated ML techniques as tools to improve survival predictions in HF patients[98-101] (Table 1). In 2015, Panahiazar et al. compared several ML models with the SHFM for survival prediction in patients with HFrEF.[98] Using EHR data from standard clinical encounters, their model resulted in an improved performance over the SHFM for predicting 1, 2, and 5-year survival, with an improvement in AUC of 11%. However, no machine learning method proved superior to logistic regression. Following this, Samad et al. hypothesized that by including a large panel of echocardiographic measurements, they would be better able to predict patient survival following echocardiography.[99] For the 15,492 heart failure patients included in their sample, 5-year survival prediction models with supervised learning techniques resulted in an improvement of 17% in AUC over the SHFM (from 0.63 to 0.80). In 2019, Kwon et al. compared a deep neural network to conventional ML, LR, and traditional scores for predicting in-hospital mortality, 1-year, and 3-year mortality of patients with acute heart failure.[102] For in-hospital mortality, their deep neural network algorithm (AUC 0.880) outperformed the best performing among multiple conventional (i.e. non-deep) ML algorithms (0.756), LR (0.720), and the GWTH-HF score (0.728). 1-year and 3-year mortality estimates also outperformed their counterparts, with AUCs of 0.782 and 0.813, respectively. Most recently, in 2020, Raghunath et al. used EKG tracings to predict 1-year all-cause mortality in any patient undergoing EKG, including those with heart failure[103]. Though their algorithm performed well across all patients (AUC 0.876, AUPRC 0.425) and outperformed their non-deep XGBoost model (AUC 0.860), it performed more modestly for patients with known heart failure (AUC 0.734, AUPRC 0.605).

Others have sought to create risk scores that are easy to use and based on ML-selected variables. For example, Adler et al. created a ML-based risk score to discriminate between high- and low-mortality risk HF patients[104]. They showed that their score has a superior ability to classify mortality risk in patients with HF compared to other risk scores (AUC 0.88 vs. best score AUC 0.78). The performance improvements may be limited, as the authors intentionally excluded patients defined as having an intermediate risk in their training cohort. However, once their risk score was developed, they compared the life expectancy for the entire cohort based on their scores, including those of intermediate risk, which resulted in a stronger correlation (-0.41)

with mortality when compared to non-ML-derived scores (IMRS, -0.31; GWTG-HF -0.25; ADHERE, -0.14).

30-day readmission

30-day readmission rates are a common measure for a hospital system's management of HF, and predicting readmission can assist in adequately meeting patient needs while minimizing resource utilization. Machine learning algorithms have been used in several approaches to better predict this outcome, yet accurate prediction remains elusive[105-107] (Table 1). In 2017, Frizzell et al. evaluated ML approaches for 56,477 patients from the Get with the Guidelines – Heart Failure registry.[105] They showed that, among multiple algorithms, none performed better in 30-day readmission predictions than logistic regression, which had a limited performance with an AUC of 0.624. In response, Awan et al. compared a multi-layer perceptron (MLP) model to both conventional ML and LR. [108] Their MLP had an improved performance compared to LR (AUCs of 0.62 vs. 0.58 with AUPRCs of 0.46 and 0.46, respectively).

Subsequently, Golas et al. compared a deep learning (DL) algorithm with conventional ML techniques for the same purpose.[107] Using over 3,500 variables from the EHR, their resultant DL model (AUC of 0.71) performed mildly better than both conventional ML (0.70) and logistic regression (0.66). In 2019, Mahajan et al. incorporated unstructured data into their ML algorithm[109]. This failed to improve AUC (0.645 with unstructured data compared to 0.645 without), though after decision curve analysis, unstructured data did improve net benefit compared to other models. The lack of success amongst all these approaches suggests it may require more than an improved model to make an accurate prediction, and that it may in part be due to reasons within the data itself. For instance, certain socio-economic factors and social support systems can affect readmission rates, which information is not always well-captured in research and hospital databases.[110, 111]

Table

Topic	Author	Year	Study Objectives	Algorithm	Comparison	Sample Size	Results
DIAGNOSIS AND IDENTIFICATION OF HEART FAILURE							
Electrocardiography							
	Attia et al.[57]	2019	Screen for patients with asymptomatic left ventricular dysfunction (LVEF ≤35%) using ECG data	CNN	BNP	163,892 ECG-TTE pairs (10,029 patients)	AUC 0.93 (CNN) vs. 0.89 (BNP)[ref]. Abnormal ECG with normal LVEF had 4x risk of developing future HF.
	Kwon et al.[59]	2019	Screen for HFrEF and HFmrEF using DL on ECG data using any patient with recent echocardiograph and ECG	DL	Conventional ML and LR	55,163 ECG-TTE pairs (22,785 patients)	0.889 (DL) vs. 0.847 (LR) and 0.853 (RF) 0.850 (DL) vs. 0.809 (LR) and 0.782 (RF)
	Wang et al.[56]	2019	Use LSTM on ECG to discriminate HF vs. healthy patients	LSTM	Literature	156 patients	Classification accuracies of 98.9% and 87.6% on two separate databases
Echocardiography							
	Zhang et al.[63]	2018	Automatically identify cardiac view and measure LVEF and LS, create predictor of HCM, PAH, and cardiac amyloid	CNN	EF and LS from echo reports	14,035 echocardiograms	Good agreement for LVEF and LS, Predictors with AUCs of 0.93, 0.87, and 0.85 respectively
	Reddy et al.[65]	2018	Form risk score for diagnosis of HFpEF using noninvasive echocardiographic and clinical data	CART, LR	ESC algorithm	514 patients	Algorithm selected 6 variables AUC 0.84 vs. 0.67 (ESC algorithm)
	Asch et al.[64]	2019	Automate LVEF quantification without volume measurements; Also detect EF ≤ 35%	DL	Human Experts	50,000 echocardiograms	r = 0.95, sensitivity of 0.90, specificity of 0.92
Identification of HF from EHR							
	Blecker et al.[68]	2016	Compare supervised algorithm to other methods for identifying hospitalized patients with HF	L1-regularized LR	LR and non-ML algorithms	46,804 hospitalizations	AUC 0.97 vs. 0.95 (LR) Sensitivity and PPV of 0.83 & 0.90 vs. 0.68 & 0.90 (LR)
	Evans et al.[69]	2016	Improve current tool for identifying hospitalized patients with HF by using NLP and supervised classifier	Prior algorithm with NLP	Prior algorithm without NLP	16,971 hospitalizations	Sensitivity and Specificity of 0.95 & 0.98 (NLP model) vs. 0.83 & 0.83 (prior model)
Other Methods of HF Detection							
	Marcinkiewics-Siemion et al.[72]	2020	Diagnose HFrEF using a panel of fasting serum metabolites	RF-selected variables	BNP	67 patients w/HFrEF 39 controls	RF Panel: AUC 0.85, Spec. 0.92, Sens. 0.73 BNP: AUC 0.92, Spec. 0.82, Sens. 0.80
	Liu et al.[71]	2019	Discriminate between healthy and HFpEF patients using heart sound characteristics	ELM	SVM	441 patients w/HFpEF 401 controls	ELM: Accuracy 0.963, Sens. 0.955, Spec 0.971 SVM: Accuracy 0.940, Sens. 0.929, Spec. 0.951
	Nirschl et al.[70]	2018	Diagnose HF in post-mortem patients using H&E stained images of cardiac tissue	DL	RF, 2 clinical pathologists	94 patients w/HF 115 controls	DL: AUC 0.989, Sens. 0.993, Spec. 0.935 RF: AUC 0.960, Sens. 0.932, Spec. 0.905 Both models outperformed pathologists on training set
PRECISION MEDICINE							
Subtype Classification							
	Shah et al.[73]	2015	Identify subgroups of HFpEF	Hierarchal clustering	---	504 patients	3 clusters identified that varied in baseline characteristics and outcomes
	Ahmad et al.[74]	2018	Identify subgroup HF patients by ML-selected features predictive of survival, independent of LVEF	RF, k-means clustering	SHFM, MAGGIC	44,886 patients	4 groups identified; Full RF model had AUC of 0.83 vs. 0.72 (SHFM) & 0.74 (MAGGIC) for survival prediction
	Przewlocka-Kosmala et al.[77]	2019	Subgroup HFpEF patients using pre- and post-cardiopulmonary exercise testing echocardiography	Hierarchal clustering	---	177 patients 51 controls	3 clusters differed in baseline chronotropic, diastolic reserve, & survival free of CV hospitalization or death
	Hedman et al.[78]	2020	Subgroup HFpEF patients and identify proteomic correlates of phenotype-based groups	Model-based clustering	---	320 patients w/HFpEF	6 groups with differences in comorbidities, outcomes, and baseline proteomics
Response to Advanced Therapies							
	Cikes et al.[84]	2018	Identify responders to CRT-D;	k-means clustering	---	1,106 patients	4 groups identified; 2 groups had a better response (HR 0.35 vs. > 0.8)
	Kalscheur et al.[85]	2018	Classify risk for mortality/readmission following CRT-P implant;	RF	BBB morphology and QRS time	595 patients	HR of 7.96 for 12-month survival (RF) vs. HR of 1.42 (ECG changes); AUC 0.74 vs. 0.67 (LR)

			algorithms	guidelines		survival prediction: AUC 0.61 vs. 0.50
Kanwar et al.[89]	2018	Predict survival following LVAD implant	TAN	HMRS	10,277 patients	AUC of 0.70 vs. 0.59 (HMRS) for 1-year survival
Loghmanpour et al.[91]	2016	Predict RV failure in LVAD patients;	TAN	Drakos score and RVFRS	10,909 patients	AUC of 0.88 vs 0.55 (Drakos) and 0.50 (RVFRS) for RV failure onset >14 days
INCIDENCE AND OUTCOME PREDICTIONS						
Incident Heart Failure and Early Detection						
Ambale-Venktesh et al.[94]	2017	Identify top-20 variables for predicting CV outcomes, including incident HF	RF	MESA-HF	6,814 patients	AUC 0.84 (RF) vs. 0.80 (MESA-HF) for HF onset
Choi et al.[96]	2016	Compare deep learning to conventional ML techniques in predicting HF onset	RNN	Conventional ML and LR	32,787 patients	AUC 0.78 (RNN) vs. 0.76 (ML) and 0.75 (LR)
Segar et al.[112]	2019	Predicts incident HF hospitalization in patients with DM	RF, RF-based risk score	Cox-based models	19,575 patients	ML-derived variables outperformed cox models AUC 0.70
Chen et al.[97]	2019	Determine how quality/diversity of data affects performance of RNN, RF, and LR in predicting future HF	RNN	RF, LR	4370 incident HF 30132 controls	Increasing data improved performance. Full dataset AUC 0.791 (RNN) vs. 0.786 (LR)
Mortality/Survival Prediction						
Panahiazar et al.[98]	2015	Compare ML models to SHFM for survival prediction in HFrEF patients	Multiple ML algorithms	SHFM; LR	5,044 patients	ML improved AUC by 11% No ML method performed better than LR
Samad et al.[99]	2018	Use echocardiographic data for survival prediction in both HF patients and all patients undergoing echocardiography	Multiple ML algorithms	SHFM; LR	331,317 echos (171,510 patients)	AUC 0.80 (RF) vs. 0.63 (SHFM) for survival in HF AUC 0.89 (RF) vs. 0.86 (LR) for survival in all patients
Adler et al.[104]	2019	Create risk score to separate HF patients with either high- or low-risk of death, ignoring those of intermediate risk	BDT	Multiple risk scores	5,822 patients for training	AUC 0.88 (BDT) vs. 0.78 (IMRS), 0.72 (GWTG-HF), and 0.63 (ADHERE), external validation AUC 0.84 and 0.81
Kwon et al.[102]	2019	Predict inpatient, 1-year, and 3-year mortality of acute heart failure with deep learning	DL	Conventional ML, LR, GWTG-HF	2,165 patients training 4,759 patients testing	In-hospital: AUC 0.88 (DL), 0.76 (RF), 0.73 (GWTH-HF) 1 year: AUC 0.78 (DL), 0.70 (RF), 0.718 (MAGGIC)
Angraal et al.[113]	2019	Predict mortality and HF hospitalization in HFpEF patients within 3 years of cohort entry	Multiple ML algorithms	Conventional ML and LR	1767 patients	Mortality: AUC 0.72 (RF), 0.66 (LR) HF Hospitalization: AUC 0.76 (RF), 0.73 (LR)
30-day Readmission						
Frizzell et al.[105]	2017	Compare supervised techniques for predicting 30-day readmission	Multiple ML algorithms	LR	56,477 patients	AUC 0.62 (TAN) vs. 0.62 (LR)
Awan et al.[108]	2019	Compares MLP for predicting 30-day HF readmission	MLP	Conventional ML and LR	10,757 patients	AUC 0.62 (MLP) vs. 0.58 (LR), AUPRC 0.46 vs. 0.46 (LR), LR > other ML methods
Golas et al.[107]	2018	Predict 30-day readmission using deep learning supervised algorithm	DL	GBM, LR	27,334 admissions for 11,510 patients	AUC 0.705 (DL) vs 0.650 (GBM) and 0.664 (LR)
Mahajan et al.[109]	2019	Predict 30-day readmission using structured and unstructured data	Structured + unstructured ML	Structured, unstructured ML	1,269 patients	Unstructured data did not increase AUC compared to structured model (0.645 vs. 0.645)

ADHERE = acute decompensated heart failure national registry; AUC = area under curve; BBB = bundle branch block; BDT = boosted decision tree; BNP = brain natriuretic peptide; CART = classification and regression tree; CNN = convolutional neural network; CRT-D = cardiac resynchronization therapy-defibrillator; CRT-P cardiac resynchronization therapy-pacemaker; CV = cardiovascular; DL = deep learning; ECG = electrocardiography; EF = ejection fraction; EHR = electronic health record; ESC = European Society of Cardiology; GBM = gradient boosted machine; HCM = hypertrophic cardiomyopathy; HF = heart failure; HFmrEF = heart failure with mid-range ejection fraction; HFpEF = heart failure with preserved ejection fraction; HFrEF = heart failure with reduced ejection fraction; HR = hazard ratio; IMRS = Intermountain risk score; LR = logistic regression; LS = longitudinal strain; LV = left ventricle; LVAD = left ventricular assist device; LVEF = left ventricular ejection fraction; MAGGIC = Meta-Analysis Global Group in Chronic Heart Failure; MESA-HF = multi-ethnic study of atherosclerosis - heart failure; ML = machine learning; MLP = multi-layer perceptron; NLP = natural language processing; PPV = positive predictive value; RNN = recurrent neural network; RV = right ventricle; RVFRS = right ventricular failure risk score; SHFM = Seattle Heart Failure Model; TAN = tree-augmented naive Bayes; TTE = transthoracic echocardiogram

Summary and Discussion

Artificial intelligence and machine learning are exciting areas of research within heart failure. Researchers have applied machine learning to many aspects of heart failure as outlined above, and each instance has been met with varying degrees of success. While many of these applications show potential for widespread incorporation in the near future, others require further testing and validation and are not ready for introduction into every-day practice.

For instance, machine learning shows promise in the diagnosis of and screening for heart failure. This may be especially so when applying machine learning to ECGs and echocardiography, where machine learning algorithms showed good performance in automating echocardiography reporting and in detecting for heart failure. Though many of these studies showed good algorithm performance, only a few were designed to be used clinically rather than simply for algorithm development[57, 59]. Other studies reviewed here are limited by their patient cohorts, which were small in size and essentially discriminated between patients with cardiac disease and those without[54, 56]. Lastly, machine learning algorithms improved the identification of patients hospitalized with heart failure, and such algorithms are already in use at several institutions.[68, 69] The widespread of these algorithms are limited to the difficulties of transferring an algorithm from one place to another, and each institution may be left to develop their own algorithms in house.

Despite initial successes of machine learning in diagnosing heart failure, other algorithms may yet be years away from standard clinical practice. For example, the classification and subgrouping of HF patients, especially those with HFpEF, resulted in interesting phenotypes with differing outcomes,[73, 74, 77, 78] yet the clinical utility and meaningfulness of these phenotypes have yet to be proven through prospective study. Similarly, investigation of the response to advanced therapies such as CRT[84-86] and LVADs[89, 91] identified certain groups who are more likely to respond to the separate therapies, but whether these groups can be used as selection criteria that improve outcomes is yet unknown. The potential for incorporation of these results into clinical practice rely on future validation and study through randomized controlled trials.

Similarly, the prediction of outcomes in heart failure continues to be challenging despite the incorporation of machine learning. In many instances, machine learning algorithms improved prediction capability over traditional risk scores[98, 99, 104]; however, machine learning often failed to provide better performance over more traditional methods (such as logistic regression)[95, 96, 98, 99, 105]. In some studies, deep learning models modestly improved performance beyond other non-deep methods,[102, 107] yet due to their complexity, the models' improvements came at a loss of interpretability in prediction results.[102] The lack of significant improvement with machine learning over traditional methods is notable, yet improvements with machine learning may still be possible through further research and development. Such results may be a reflection of the difficulty in obtaining high-quality data as well as the natural heterogeneity that exists between patients. In the meantime, as machine learning did not afford a significant advantage and introduced more model complexity, it may be such that these problems are best represented by simpler models with more inherent interpretability.[114]

Remaining Obstacles to the Widespread Utilization of Machine Learning Algorithms in Clinical Practice

As noted above, there are still many obstacles remaining before machine learning algorithms progress to becoming a part of standard practice. In some cases, these can be overcome through improved data collection and algorithm development, and others through improved

interpretability and interoperability. In addition, researchers and clinicians will need to decide how best to use these algorithms after their development and testing and determine what their appropriate roles should be. In many instances, these algorithms may prove to have greater utility as clinical decision supports, where others will help aid in identifying and tracking patients in ways clinicians have little time to do. In general, the future incorporation of these and other machine learning algorithms relies upon the satisfaction of multiple criteria, several of which include: 1) algorithms must be validated on multiple external cohorts and tested through rigorous prospective study, 2) they must have wide applicability to different hospital systems, settings, and geographic locations, 3) they must strike a balance between interpretability and model complexity and accuracy, 4) they must be able to be monitored to ensure they continue to provide accurate results given new user-driven data, and 5) they must provide enough benefit over other methods to be cost-effective in both development and in wide-spread implementation. In addition, the clinicians must ensure the proper use and application of machine learning algorithms, typically after considering questions such as the following: 1) Will this application be better served by machine learning as opposed to traditional statistical models? 2) How complicated of a model does this scenario require, and does any added complexity outweigh any drawbacks of such a model? Will this model benefit from using a simpler, conventional ML model, or does its complexity require something more, such as a deep learning method? 3) Which features or variables should be included in this model, and what needs to be done to maximize accuracy while avoiding overfitting? How many clinical parameters should be used to create the model such that gathering this information on a new patient is not overly expensive or burdensome? 4) Which outcomes matter most to the clinical scenario, and will predicting them lead to improved patient care? 5) How will this ML model be adapted into clinical practice, and what will be done to gain the trust of clinicians who will be using it when the model may not be interpretable?[115] If these questions can be appropriately met and the algorithms rigorously proven, we may soon see the heralded improvements sought by applying artificial intelligence to medicine.

Conclusion

In conclusion, machine learning has found interesting applications in the diagnosis, classification, and prediction of heart failure. Many of these methods show potential in aiding the management of heart failure, yet few are ready for broad application and incorporation into common practice. Each must be proven to meet strict criteria through further evaluation, rigorous study, and thorough validation through prospective study in order to transition these methods from being a possibility to reality. Notwithstanding the difficulties facing machine learning techniques in improving heart failure management, we believe that these methods will someday lead to more accurate diagnoses, more precise treatments, and better understanding and prediction of patient outcomes.

Funding: No extramural funding was used to support this work.

Acknowledgements: None. The authors are solely responsible for the design and conduct of this study, all study analyses, the drafting and editing of the paper and its final contents.

Conflicts of Interest: RJM receives significant research support from the National Institutes of Health (U01HL125511-01A1, U10HL110312 and R01AG045551-01A1), Akros, Amgen, AstraZeneca, Bayer, GlaxoSmithKline, Gilead, Luitpold/American Regent, Medtronic, Merck, Novartis, Otsuka, and ResMed; honoraria from Abbott, Amgen, AstraZeneca, Bayer, Janssen, Luitpold Pharmaceuticals, Medtronic, Merck, Novartis, and ResMed; and has served on an advisory board for Amgen, AstraZeneca, Luitpold, Merck, Novartis and Boehringer Ingelheim. The other authors have no relevant disclosures. KJA has received research support from NIH, Merck, Bayer, AstraZeneca, Bristol-Meyers Squibb, Eli Lilly & Company, and Boehringer Ingelheim; has served as a consultant for Abbott Vascular, AstraZeneca, Bristol-Meyers Squibb, Gilead, Janssen, Pfizer, Promedior, and GlaxoSmithKline; and has served on data monitoring committees for NIH, Boehringer Ingelheim, Forest, Pfizer, Promedior, and GlaxoSmithKline, University of Miami, University of North Carolina, University of Virginia. The other authors have no relevant disclosures to report.

References

1. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, et al. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. Circulation 2017:CIR. 0000000000000659.

2. National Academies of Sciences E, Medicine. Improving diagnosis in health care: National Academies Press; 2015.

3. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, et al. The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation 2006;113(11):1424-33.

4. Pocock SJ, Ariti CA, McMurray JJ, Maggioni A, Kober L, Squire IB, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. Eur Heart J 2013;34(19):1404-13.

5. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, et al. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. Circ Cardiovasc Qual Outcomes 2010;3(1):25-32.

6. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine 2019;380(14):1347-1358.

7. Deo RC. Machine Learning in Medicine. Circulation 2015;132(20):1920-30.

8. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology 2017;2(4):230-243.

9. Levinson J, Askeland J, Becker J, Dolson J, Held D, Kammel S, et al. Towards fully autonomous driving: Systems and algorithms. In: Intelligent Vehicles Symposium (IV), 2011 IEEE; 2011: IEEE; 2011. p. 163-168.

10. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012; 2012. p. 1097-1105.

11. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; 2016. p. 770-778.

12. Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning; 2008: ACM; 2008. p. 160-167.

13. 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(22):2402-2410.

14. 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(7639):115-118.

15. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:171105225 2017.

16. Samala RK, Chan HP, Hadjiiski L, Helvie MA, Wei J, Cha K. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Medical physics 2016;43(12):6654-6666.

17. 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 Medicine 2018;1(1):18.

18. Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Comput Struct Biotechnol J 2017;15:26-47.

19. Awan SE, Sohel F, Sanfilippo FM, Bennamoun M, Dwivedi G. Machine learning in heart failure: ready for prime time. Current opinion in cardiology 2018;33(2):190-195.

20. Bzdok D, Altman N, Krzywinski M. Points of significance: statistics versus machine learning. In: Nature Publishing Group; 2018.

21. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural computation 2006;18(7):1527-1554.

22. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE journal of biomedical and health informatics 2017;22(5):1589-1604.

23. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA internal medicine 2018;178(11):1544-1547.

24. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. Jama 2017;318(6):517-518.

25. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol 2017;69(21):2657-2664.

26. LeCun Y, Bengio Y, Hinton G. Deep learning. nature 2015;521(7553):436-444.

27. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE journal of biomedical and health informatics 2016;21(1):4-21.

28. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol 2018;71(23):2668-2679.

29. Guyon I, Elisseeff A. An introduction to variable and feature selection. Journal of machine learning research 2003;3(Mar):1157-1182.

30. Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial intelligence in medicine 2010;50(2):105-115.

31. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Computing and Applications 2010;19(2):263-282.

32. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC medical informatics and decision making 2012;12(1):8.

33. Sutton RS, Barto AG. Reinforcement learning: An introduction: MIT press; 2018.

34. Settles B. Active learning literature survey: University of Wisconsin-Madison Department of Computer Sciences; 2009.

35. Vinyals O, Blundell C, Lillicrap T, Wierstra D. Matching networks for one shot learning. In: Advances in neural information processing systems; 2016; 2016. p. 3630-3638.

36. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering 2007;160:3-24.

37. Singh A, Thakur N, Sharma A. A review of supervised machine learning algorithms. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom); 2016: Ieee; 2016. p. 1310-1315.

38. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148(3):839-43.

39. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29-36.

40. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition 1997;30(7):1145-1159.

41. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one 2015;10(3).

42. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning; 2006; 2006. p. 233-240.

43. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics 2020;21(1):6.

44. Goodfellow I, Bengio Y, Courville A. Deep Learning: MIT Press; 2016.

45. Madani A, Ong JR, Tibrewal A, Mofrad MR. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ digital medicine 2018;1(1):1-11.

46. Zhu X, Goldberg AB. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning 2009;3(1):1-130.

47. Raina R, Madhavan A, Ng AY. Large-scale deep unsupervised learning using graphics processors. In: Proceedings of the 26th annual international conference on machine learning; 2009; 2009. p. 873-880.

48. Beam AL, Kohane IS. Big data and machine learning in health care. Jama 2018;319(13):1317-1318.

49. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: ACM; 2016. p. 785-794.

50. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, et al. 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. European Journal of Heart Failure 2016;18(8):891-975.

51. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine 2019;25(1):65.

52. Chen W, Liu G, Su S, Jiang Q, Nguyen H. A CHF detection method based on deep learning with RR intervals. In: Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE; 2017: IEEE; 2017. p. 3369-3372.

53. Chen WH, Zheng LR, Li KY, Wang Q, Liu GZ, Jiang Q. A Novel and Effective Method for Congestive Heart Failure Detection and Quantification Using Dynamic Heart Rate Variability Measurement. Plos One 2016;11(11):e0165304.

54. Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. Computer Methods and Programs in Biomedicine 2016;130:54-64.

55. Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M, et al. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. Applied Intelligence 2018:1-12.

56. Wang L, Zhou X. Detection of congestive heart failure based on LSTM-based deep network via short-term RR intervals. Sensors 2019;19(7):1502.

57. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nature medicine 2019;25(1):70.

58. Costello-Boerrigter LC, Boerrigter G, Redfield MM, Rodeheffer RJ, Urban LH, Mahoney DW, et al. Amino-terminal pro-B-type natriuretic peptide and B-type natriuretic peptide in the general community: determinants and detection of left ventricular dysfunction. Journal of the American College of Cardiology 2006;47(2):345-353.

59. Kwon J-m, Kim K-H, Jeon K-H, Kim HM, Kim MJ, Lim S-M, et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. Korean circulation journal 2019;49(7):629-639.

60. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE, Drazner MH, et al. 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. Journal of the American College of Cardiology 2013;62(16):e147-e239.

61. Selmeryd J, Henriksen E, Leppert J, Hedberg P. Interstudy heterogeneity of definitions of diastolic dysfunction severely affects reported prevalence. Eur Heart J Cardiovasc Imaging 2016;17(8):892-9.

62. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology 2017;284(2):574-582.

63. Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, et al. Fully Automated Echocardiogram Interpretation in Clinical Practice: Feasibility and Diagnostic Accuracy. Circulation 2018;138(16):1623-1635.

64. Asch FM, Poilvert N, Abraham T, Jankowski M, Cleve J, Adams M, et al. Automated Echocardiographic Quantification of Left Ventricular Ejection Fraction Without Volume Measurements Using a Machine Learning Algorithm Mimicking a Human Expert. Circulation: Cardiovascular Imaging 2019;12(9):e009303.

65. Reddy YNV, Carter RE, Obokata M, Redfield MM, Borlaug BA. A Simple, Evidence-Based Approach to Help Guide Diagnosis of Heart Failure With Preserved Ejection Fraction. Circulation 2018;138(9):861-870.

66. Sanchez-Martinez S, Duchateau N, Erdei T, Kunszt G, Aakhus S, Degiovanni A, et al. Machine Learning Analysis of Left Ventricular Function to Characterize Heart Failure With Preserved Ejection Fraction. Circ Cardiovasc Imaging 2018;11(4):e007138.

67. Tabassian M, Sunderji I, Erdei T, Sanchez-Martinez S, Degiovanni A, Marino P, et al. Diagnosis of Heart Failure With Preserved Ejection Fraction: Machine Learning of Spatiotemporal Variations in Left Ventricular Deformation. Journal of the American Society of Echocardiography 2018;31(12):1272-+.

68. Blecker S, Katz SD, Horwitz LI, Kuperman G, Park H, Gold A, et al. Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. JAMA Cardiol 2016;1(9):1014-1020.

69. Evans RS, Benuzillo J, Horne BD, Lloyd JF, Bradshaw A, Budge D, et al. Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. J Am Med Inform Assoc 2016;23(5):872-8.

70. Nirschl JJ, Janowczyk A, Peyster EG, Frank R, Margulies KB, Feldman MD, et al. A deep-learning classifier identifies patients with clinical heart failure using whole-slide images of H&E tissue. Plos One 2018;13(4):e0192726.

71. Liu Y, Guo X, Zheng Y. An automatic approach using ELM classifier for HFpEF identification based on heart sound characteristics. Journal of medical systems 2019;43(9):285.

72. Marcinkiewicz-Siemion M, Kaminski M, Ciborowski M, Ptaszynska-Kopczynska K, Szpakowicz A, Lisowska A, et al. Machine-learning facilitates selection of a novel diagnostic panel of metabolites for the detection of heart failure. Scientific Reports 2020;10(1):1-11.

73. Shah SJ, Katz DH, Selvaraj S, Burke MA, Yancy CW, Gheorghiade M, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. Circulation 2015;131(3):269-79.

74. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, et al. Machine Learning Methods Improve Prognostication, Identify Clinically Distinct Phenotypes, and Detect Heterogeneity in Response to Therapy in a Large Cohort of Heart Failure Patients. J Am Heart Assoc 2018;7(8):e008081.

75. Ahmad T, Pencina MJ, Schulte PJ, O'Brien E, Whellan DJ, Pina IL, et al. Clinical implications of chronic heart failure phenotypes defined by cluster analysis. J Am Coll Cardiol 2014;64(17):1765-74.

76. Horiuchi Y, Tanimoto S, Latif AHMM, Urayama KY, Aoki J, Yahagi K, et al. Identifying novel phenotypes of acute heart failure using cluster analysis of clinical variables. International Journal of Cardiology 2018;262:57-63.

77. Przewlocka-Kosmala M, Marwick TH, Dabrowski A, Kosmala W. Contribution of Cardiovascular Reserve to Prognostic Categories of Heart Failure With Preserved Ejection

Fraction: A Classification Based on Machine Learning. J Am Soc Echocardiogr 2019;32(5):604-615 e6.

78. Hedman ÅK, Hage C, Sharma A, Brosnan MJ, Buckbinder L, Gan L-M, et al. Identification of novel pheno-groups in heart failure with preserved ejection fraction using machine learning. Heart 2020.

79. Pitt B, Pfeffer MA, Assmann SF, Boineau R, Anand IS, Claggett B, et al. Spironolactone for heart failure with preserved ejection fraction. New England Journal of Medicine 2014;370(15):1383-1392.

80. Abraham WT, Fisher WG, Smith AL, Delurgio DB, Leon AR, Loh E, et al. Cardiac resynchronization in chronic heart failure. N Engl J Med 2002;346(24):1845-53.

81. Bristow MR, Saxon LA, Boehmer J, Krueger S, Kass DA, De Marco T, et al. Cardiac-resynchronization therapy with or without an implantable defibrillator in advanced chronic heart failure. New England Journal of Medicine 2004;350(21):2140-2150.

82. Yu CM, Zhang Q, Fung JW, Chan HC, Chan YS, Yip GW, et al. A novel tool to assess systolic asynchrony and identify responders of cardiac resynchronization therapy by tissue synchronization imaging. J Am Coll Cardiol 2005;45(5):677-84.

83. Pitzalis MV, Iacoviello M, Romito R, Massari F, Rizzon B, Luzzi G, et al. Cardiac resynchronization therapy tailored by echocardiographic evaluation of ventricular asynchrony. J Am Coll Cardiol 2002;40(9):1615-22.

84. Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. Eur J Heart Fail 2019;21(1):74-85.

85. Kalscheur MM, Kipp RT, Tattersall MC, Mei C, Buhr KA, DeMets DL, et al. Machine Learning Algorithm Predicts Cardiac Resynchronization Therapy Outcomes: Lessons From the COMPANION Trial. Circ Arrhythm Electrophysiol 2018;11(1):e005499.

86. Feeny AK, Rickard J, Patel D, Toro S, Trulock KM, Park CJ, et al. Machine Learning Prediction of Response to Cardiac Resynchronization Therapy: Improvement Versus Current Guidelines. Circulation: Arrhythmia and Electrophysiology 2019;12(7):e007316.

87. Rogers JG, Pagani FD, Tatooles AJ, Bhat G, Slaughter MS, Birks EJ, et al. Intrapericardial Left Ventricular Assist Device for Advanced Heart Failure. N Engl J Med 2017;376(5):451-460.

88. Mehra MR, Naka Y, Uriel N, Goldstein DJ, Cleveland JC, Jr., Colombo PC, et al. A Fully Magnetically Levitated Circulatory Pump for Advanced Heart Failure. N Engl J Med 2017;376(5):440-450.

89. Kanwar MK, Lohmueller LC, Kormos RL, Teuteberg JJ, Rogers JG, Lindenfeld J, et al. A Bayesian Model to Predict Survival After Left Ventricular Assist Device Implantation. J Am Coll Cardiol HF 2018;6(9):771-779.

90. Cowger J, Sundareswaran K, Rogers JG, Park SJ, Pagani FD, Bhat G, et al. Predicting survival in patients receiving continuous flow left ventricular assist devices: the HeartMate II risk score. Journal of the American College of Cardiology 2013;61(3):313-321.

91. Loghmanpour NA, Kormos RL, Kanwar MK, Teuteberg JJ, Murali S, Antaki JF. A Bayesian Model to Predict Right Ventricular Failure Following Left Ventricular Assist Device Therapy. J Am Coll Cardiol HF 2016;4(9):711-21.

92. Matthews JC, Koelling TM, Pagani FD, Aaronson KD. The right ventricular failure risk score: a pre-operative tool for assessing the risk of right ventricular failure in left ventricular assist device candidates. Journal of the American College of Cardiology 2008;51(22):2163-2172.

93. Drakos SG, Janicki L, Horne BD, Kfoury AG, Reid BB, Clayson S, et al. Risk factors predictive of right ventricular failure after left ventricular assist device implantation. The American journal of cardiology 2010;105(7):1030-1035.

94. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res 2017;121(9):1092-1101.

95. Segar MW, Vaduganathan M, Patel KV, McGuire DK, Butler J, Fonarow GC, et al. Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. Diabetes care 2019;42(12):2298-2306.

96. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. Journal of the American Medical Informatics Association 2016;24(2):361-370.

97. Chen R, Stewart WF, Sun J, Ng K, Yan X. Recurrent Neural Networks for Early Detection of Heart Failure From Longitudinal Electronic Health Record Data: Implications for Temporal Modeling With Respect to Time Before Diagnosis, Data Density, Data Quantity, and Data Type. Circulation: Cardiovascular Quality and Outcomes 2019;12(10):e005114.

98. Panahiazar M, Taslimitehrani V, Pereira N, Pathak J. Using EHRs and Machine Learning for Heart Failure Survival Analysis. Medinfo 2015: Ehealth-Enabled Health 2015;216:40-44.

99. Samad MD, Ulloa A, Wehner GJ, Jing L, Hartzel D, Good CW, et al. Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning. J Am Coll Cardiol Img 2018;12(4):681-689.

100. Kwon Jm, Kim KH, Jeon KH, Park J. Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. Echocardiography 2018.

101. Hearn J, Ross HJ, Mueller B, Fan CP, Crowdy E, Duhamel J, et al. Neural Networks for Prognostication of Patients With Heart Failure: Improving Performance Through the Incorporation of Breath-by-Breath Data From Cardiopulmonary Exercise Testing. Circulation-Heart Failure 2018;11(8):e005193.

102. Kwon J-m, Kim K-H, Jeon K-H, Lee SE, Lee H-Y, Cho H-J, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PloS one 2019;14(7).

103. Raghunath S, Cerna AEU, Jing L, Stough J, Hartzel DN, Leader JB, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nature Medicine 2020:1-6.

104. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, et al. Improving risk prediction in heart failure using machine learning. European journal of heart failure 2019.

105. Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. JAMA Cardiol 2017;2(2):204-209.

106. Shameer K, Johnson KW, Yahi A, Miotto R, Li L, Ricks D, et al. Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study

using Mount Sinai Heart Failure Cohort. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017; 2017: World Scientific; 2017. p. 276-287.

107. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak 2018;18(1):44.

108. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Dwivedi G. Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. ESC heart failure 2019;6(2):428-435.

109. Mahajan SM, Ghani R. Combining Structured and Unstructured Data for Predicting Risk of Readmission for Heart Failure Patients. Studies in health technology and informatics 2019;264:238-242.

110. Kind AJ, Jencks S, Brock J, Yu M, Bartels C, Ehlenbach W, et al. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. Annals of internal medicine 2014;161(11):765-774.

111. Hu J, Gonsahn MD, Nerenz DR. Socioeconomic status and readmissions: evidence from an urban teaching hospital. Health Affairs 2014;33(5):778-785.

112. Segar MW, Patel KV, Ayers C, Basit M, Tang WW, Willett D, et al. Phenomapping of patients with heart failure with preserved ejection fraction using machine learning-based unsupervised cluster analysis. European journal of heart failure 2019.

113. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. JACC: Heart Failure 2019;8(1):12-21.

114. Akbilgic O, Davis RL. The Promise of Machine Learning: When Will it be Delivered? In: Elsevier; 2019.

115. Manlhiot C. Machine learning for predictive analytics in medicine: real opportunity or overblown hype? In: Oxford University Press; 2018.

Highlights:

- Machine learning aims to use big data to make accurate predictions.
- New applications are arising in diagnosing and understanding heart failure.
- Understanding the basics of machine learning is crucial to proper implementation.
- Machine learning has the potential to improve the way we manage heart failure.

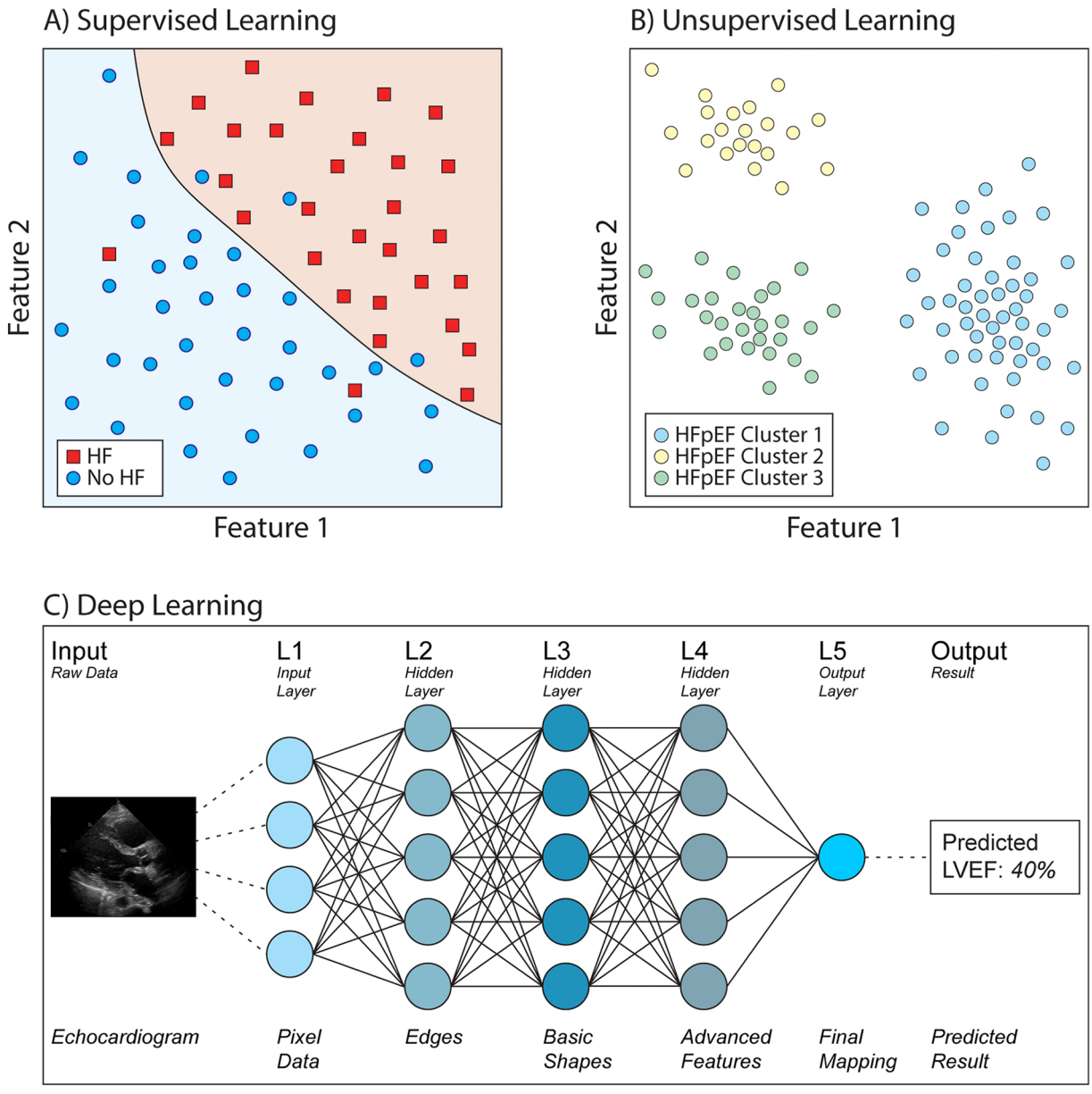
A) Supervised Learning
Feature 2
Feature 1
HF
No HF
B) Unsupervised Learning
Feature 2
Feature 1
HFpEF Cluster 1
HFpEF Cluster 2
HFpEF Cluster 3
C) Deep Learning
Input
Raw Data
L1
Input Layer
L2
Hidden Layer
L3
Hidden Layer
L4
Hidden Layer
L5
Output Layer
Output
Result
Predicted LVEF: 40%
Echocardiogram
Pixel Data
Edges
Basic Shapes
Advanced Features
Final Mapping
Predicted Result

Figure 1

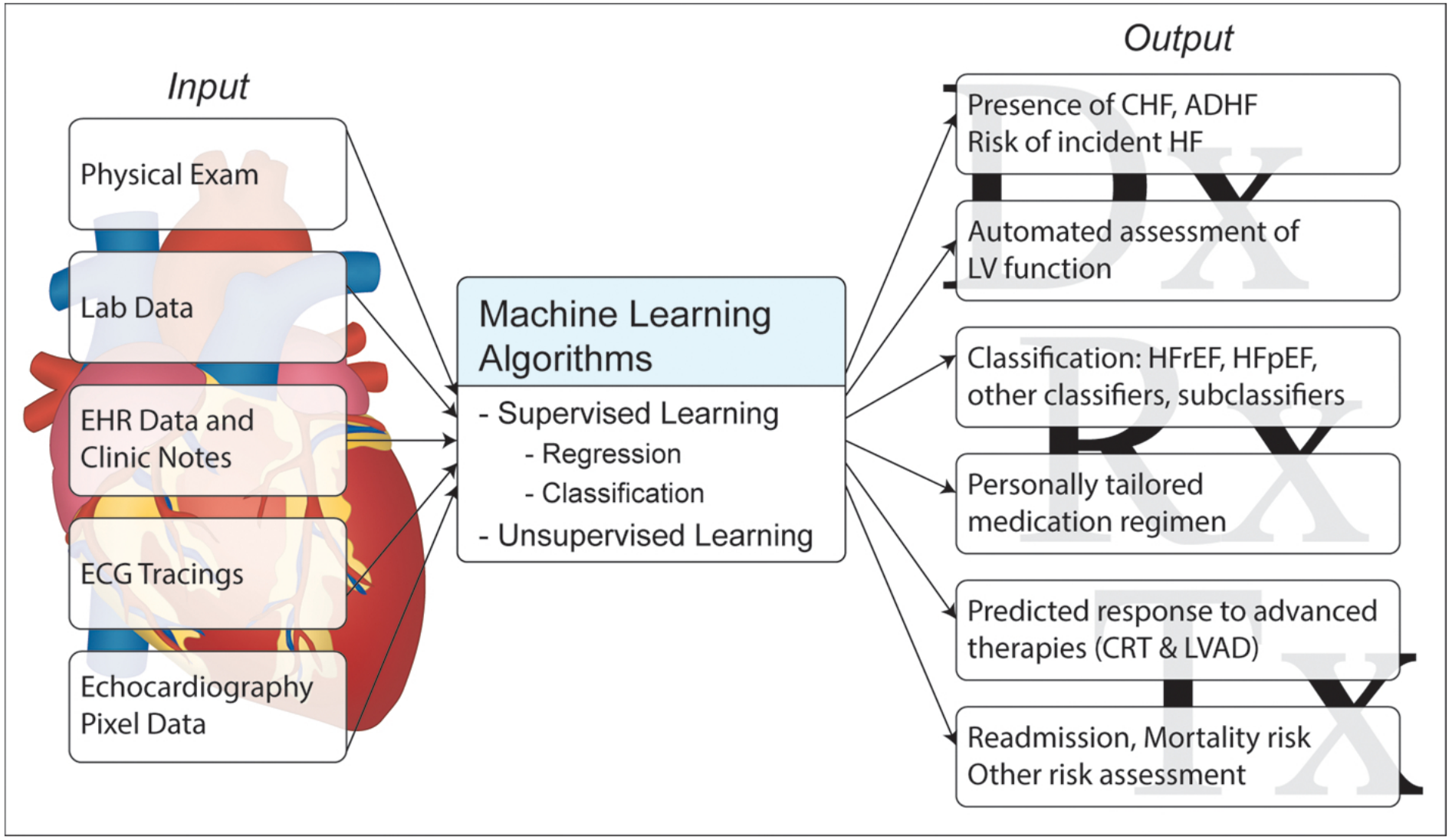
Input
Physical Exam
Lab Data
EHR Data and Clinic Notes
ECG Tracings
Echocardiography Pixel Data
Machine Learning Algorithms
- Supervised Learning
- Regression
- Classification
- Unsupervised Learning
Output
Presence of CHF, ADHF
Risk of incident HF
Automated assessment of LV function
Classification: HFrEF, HFpEF, other classifiers, subclassifiers
Personally tailored medication regimen
Predicted response to advanced therapies (CRT & LVAD)
Readmission, Mortality risk
Other risk assessment

Figure 2