Machine learning versus conventional clinical methods in guiding management of heart failure patients—a systematic review

George Bazoukis[1] · Stavros Stavrakis[2] · Jiandong Zhou[3,4] · Sandeep Chandra Bollepalli[5] · Gary Tse[6] ·
Qingpeng Zhang[3,4] · Jagmeet P. Singh[7] · Antonis A. Armoundas[5,8]

Abstract

Machine learning (ML) algorithms "learn" information directly from data, and their performance improves proportionally with the number of high-quality samples. The aim of our systematic review is to present the state of the art regarding the implementation of ML techniques in the management of heart failure (HF) patients. We manually searched MEDLINE and Cochrane databases as well the reference lists of the relevant review studies and included studies. Our search retrieved 122 relevant studies. These studies mainly refer to (a) the role of ML in the classification of HF patients into distinct categories which may require a different treatment strategy, (b) discrimination of HF patients from the healthy population or other diseases, (c) prediction of HF outcomes, (d) identification of HF patients from electronic records and identification of HF patients with similar characteristics who may benefit form a similar treatment strategy, (e) supporting the extraction of important data from clinical notes, and (f) prediction of outcomes in HF populations with implantable devices (left ventricular assist device, cardiac resynchronization therapy). We concluded that ML techniques may play an important role for the efficient construction of methodologies for diagnosis, management, and prediction of outcomes in HF patients.

Keywords Machine learning · Heart failure · Deep learning

Introduction

Heart failure (HF) is a clinical syndrome characterized by dyspnea, fatigue, and clinical signs of congestion leading to frequent hospitalizations, poor quality of life, and shortened life expectancy [1, 2]. HF is a global pandemic that affects approximately 1–2% of the adult population in developed countries [3], around 26 million people worldwide [4], rising to ≥ 10% among people > 70 years of age [3], while the considerable HF health expenditures (~ $31 billion, in the USA in 2012) [5] are expected to sharply increase with an aging population.

Despite advancements in medical, device-based, and surgical management of HF, outcomes remain non-satisfactory even in Western developed countries [6]. Evidently, emphasis

✉ Antonis A. Armoundas
    armoundas.antonis@mgh.harvard.edu

1   Second Department of Cardiology, Evangelismos General Hospital of Athens, Athens, Greece

2   University of Oklahoma Health Science Center, Oklahoma City, OK, USA

3   School of Data Science, City University of Hong Kong, Hong Kong, China

4   Shenzhen Research Institute of City University of Hong Kong, Shenzhen, Guangdong, China

5   Cardiovascular Research Center, Massachusetts General Hospital, 149 13th Street, Charlestown, Boston, MA 02129, USA

6   Laboratory of Cardiovascular Physiology, Li Ka Shing Institute of Health Sciences, Hong Kong SAR, People's Republic of China

7   Cardiology Division, Cardiac Arrhythmia Service, Massachusetts General Hospital, Boston, MA, USA

8   Institute for Medical Engineering and Science, Massachusetts Institute of Technology Cambridge, Cambridge, MA, USA

in investigating efficient research methodologies for HF management is one of the leading study directions that cannot be overlooked [7].

Recently, machine learning (ML) algorithms have used computational methods to "learn" information directly from data, and their performance has been shown to improve proportionally with the number of high-quality samples [8]. ML algorithms have been applied in different aspects of medicine [9, 10], including earlier disease detection [11, 12], improve diagnosis accuracy [13–16], identification of new physiological observations or patterns [17], development of personalized diagnostics and/or therapeutic approaches [18, 19], research purposes [20], etc.

The aim of this systematic review is to present the state of the art regarding the utility of ML techniques in comparison with conventional methods, in improving outcomes in HF patients.

Methods

This systematic review was guided by the PRISMA statement for systematic reviews and meta-analyses [21].

Machine learning architectures

Machine learning is an emerging technology paradigm that enables computers to learn patterns and insights from the data without being explicitly programmed. Details of ML algorithms adopted for managing HF patients are provided in the Online Supplement.

Search strategy

MEDLINE and Cochrane library databases were manually searched (G.B., G.T.) without year or language restriction or any other limits until May 29, 2019. The following algorithm was used: "((Machine learning OR deep learning OR bayes OR regression tree OR k means clustering OR vector machine OR artificial neural networks OR random forests OR decision trees OR nearest neighbours) AND heart failure)." Furthermore, the reference list of all the included studies as well as relevant review articles were also searched.

Study inclusion/exclusion criteria

All studies that included data about the implementation of ML techniques in HF (diagnosis, severity classification, prediction of adverse outcomes, identification of HF patients in electronic records, etc.) were considered as relevant and included in the systematic review. Review studies, studies that did not include data regarding HF patients and studies in experimental models, were excluded either at the title/abstract or at the full-text level.

Data extraction and statistical analysis

The data extraction was performed by two independent investigators (G.B., J.Z.) and any disagreement was resolved by discussion.

We used a recently proposed score by Qiao [22] for the quality assessment of ML studies (for details, please see the Online Supplement).

Results

Search results

As outlined in Supplementary Fig. 1, our search strategy revealed in total 122 relevant studies (one study provided data for two different outcomes [OS21]). Figure 1 summarizes the different areas of ML implementation in HF patients.

Classification of HF patients

Our search retrieved four streams of studies regarding the implementation of ML techniques in patient classification, pertaining to HF with reduced ejection fraction (HFrEF), HF with preserved ejection fraction (HFpEF), and in different HFpEF subtypes. The variables for HF characteristics included demographics, clinical examination, laboratory exams, medical history, electrocardiographic data, echocardiographic data, and heart rate variability (HRV) (Supplementary Table 1). All studies were classified as intermediate-high quality (intermediate: 2 studies, high: 2 studies) in the quality assessment (Supplementary Table 9). This suggests that the provided outcomes are less prone to different kinds of bias.

Modern classification methods have shown a better performance over conventional classification methods that could lead to better management in clinical practice (Table 1).

Discrimination of HF patients from subjects with no HF

Our search retrieved 30 studies regarding the discrimination of HF patients, from subjects with no HF (Supplementary Table 2). All studies were classified as intermediate-high quality (intermediate: 14 studies, high: 16 studies) in the quality assessment (Supplementary Table 9), suggesting that the provided outcomes are less prone to different kinds of bias (Table 1).

The general process of ML techniques for HF discrimination in a non-acute setting is to estimate the probability of HF based on prior clinical history of the patient, the presenting symptoms, physical examination, and resting electrocardiogram. Application of ML techniques for HF discrimination on the available data is less time consuming and more accurate
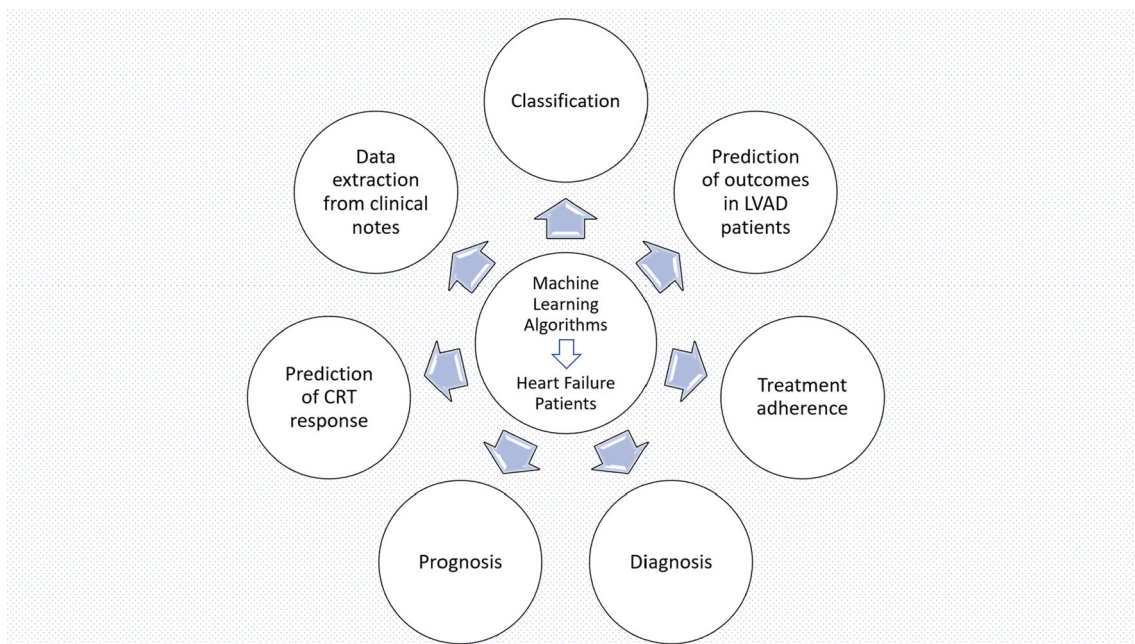
Fig. 1 Areas of application of machine learning in the management of heart failure patients

than traditionally used statistics or expert methods. Accurate HF discrimination via ML techniques allows for treatments and interventions to be delivered in a more efficient and targeted way, permits assessment of the HF patient's progress, prevents condition worsening, affects positively the patient's health, and contributes to decrease of medical costs. The main difference between the ML methods for HF discrimination lies in the different heart rate variability features employed to detect HF.

Sanchez-Martinez et al. (2017) used multiple kernel learning method to differentiate cardiac and non-cardiac cause of breathlessness and revealed processes leading to HFpEF with a specificity as high as 90.9% [OS42]. It should be noted that many ML studies found that feature selection determines the performance of the model, and thus automatic feature selection scheme is needed. Such automatic feature selection is also an advantage of the latest ML methods.

Prediction of outcomes

Our search retrieved 58 studies regarding the implementation of ML techniques in the prediction of major outcomes in HF patients. Specifically, the measured outcomes that were studied include mortality, hospitalizations, decompensations, implantable cardioverter defibrillator (ICD) implantations for secondary prevention, need for mechanical circulatory support, heart transplantation, pump failure, myocardial infarction, strokes, and ventricular assist device implantation (Supplementary Table 3). All studies were classified as intermediate-high quality (intermediate: 39 studies, high: 21 studies) in the quality assessment (Supplementary Table 9).

This suggests that the provided outcomes are less prone to different kinds of bias (Table 1).

Existing studies utilize demographic, clinical, laboratory, and electrocardiographic data (short-term or long-term HRV measures) as the main predictors and incorporate multiple classifiers such as support vector machine (SVM), classification and regression trees (CART), k-nearest neighbor algorithm (k-NN). These methods can work well separately, or collectively through certain ensemble learning techniques [23].

Identification of HF patients with similar characteristics from electronic medical records

Our search retrieved 6 studies regarding the role of ML techniques in the identification of HF patients from a pool of hospitalized patients or identification of patients with similar characteristics (Supplementary Table 4). All studies were classified as intermediate-high quality (intermediate: 2 studies, high: 4 studies) in the quality assessment (Supplementary Table 9), suggesting that the provided outcomes are less prone to different kinds of bias.

Specifically, Cikes et al. (2019) used unsupervised machine learning-based phenogrouping in HF to provide a clinically meaningful classification of a phenotypically heterogeneous HF cohort by integrating clinical parameters and full heart cycle imaging data [OS127]. Pakhomov et al. (2007) used predictive ML techniques and language processing contained in the electronic medical records, to identify patients with HF with 96% specificity [OS114]. Panahiazar et al. (2015) developed a multidimensional patient similarity

Table 1 Comparison of machine learning algorithms with traditional methods in the management of heart failure

Author	Journal	Year	Outcome	Comparison between machine learning and conventional methods		Conclusion
				Machine learning models	Conventional methods	
Classification of HF patients						
Austin PC	*Journal of Clinical Epidemiology*	2013	Discrimination HFpEF vs HFrEF	Model / AUC: Regression tree 0.683; Bagged regression tree 0.733; Random forest 0.751; Boosted regression tree (depth 1) 0.752; Boosted regression tree (depth 2) 0.768; Boosted regression tree (depth 3) 0.772; Boosted regression tree (depth 4) 0.774	Model LR / AUC 0.780	Conventional LR performed at least as well as modern methods
CRT response						
Kalscheur MM	*Circ Arrhythm Electropysiol*	2018	All-cause mortality or HF hospitalization in CRT recipients	AUC values RF model (0.74, 95% CI 0.72–0.76) Sequential minimal optimization to train a SVM (0.67, 95% CI 0.65–0.68)	AUC values Multivariate LR (0.67, 95% CI 0.65–0.69)	The improvement in AUC for the RF model was statistically significant compared to the other models, $p < 0.001$
Data extraction						
Zhang R	*BMC Med Inform Decis Mak*	2018	HF information (NYHA) extraction from clinical notes	RF, n-gram features → F-measure 93.78%, recall 92.23%, precision 95.40%, SVM → F-measure 93.52%, recall 93.21%, precision 93.84%	LR → F-measure 90.42%, recall 90.82%, precision 90.03%	ML-based methods outperformed a rule-based method. The best machine learning method was an RF
HF diagnosis						
Nirschi JJ	*PlosOne*	2018	HF diagnosis using biopsy images	AUC value RF 0.952 Deep learning 0.974	AUC value Pathologists 0.75	ML models outperformed conventional methods
Rasmy L	*J Biomed Inform*	2018	HF diagnosis	AUC value Recurrent NN 0.822	AUC value LR 0.766	ML outperformed conventional methods
Son CS	*J Biomed Inform*	2012	HF diagnosis	Rough sets based decision-making model → accuracy 97.5%, SENS 97.2%, SPE 97.7%, PPV 97.2%, NPV 97.7%, AUC 97.5%	LR-based decision-making model → accuracy 88.7%, SENS 90.1%, SPE 87.5%, PPV 85.3%, NPV 91.7%, AUC 88.8%	ML models outperformed conventional methods
Wu J	*Med Care*	2010	HF diagnosis	Boosting using a less strict cut-off had better performance compared to SVM	The highest median AUC (0.77) was observed for LR with Bayesian information criterion	LR and boosting were, both, superior to SVM
Identification of HF patients						
Blecker S	*JAMA Cardiology*	2016	Identification of HF patients	ML using notes and imaging reports → (developmental set) AUC 99%, SENS 92%, PPV 80%. (Validation SET) AUC 97%, SENS 84%, PPV 80%	LR using structured data → (developmental set) AUC 96%, SENS 78%, PPV 80%. (Validation SET) AUC 95%, SENS 76%, PPV 80%	ML models improved identification of HF patients

Table 1 (continued)

Author	Journal	Year	Outcome	Comparison between machine learning and conventional methods		Conclusion
				Machine learning models	Conventional methods	
Blecker S	*J Card Fail*	2018	Identification of HF hospitalization	ML with use of both data → (developmental set) AUC 99%, SENS 98%, PPV 43%. (Validation SET) AUC 99%, SENS 98%, PPV 34%	LR using structured data, notes, and imaging reports → (developmental set) AUC 96%, SENS 98%, PPV 14%. (Validation SET) AUC 96%, SENS 98%, PPV 15%	ML models performed better in identifying decompensated HF
Choi E	*Journal of AMIA*	2017	Predicting HF diagnosis from EHR	AUC values 12-month observation → NN model 0.777 MLP with 1 hidden layer 0.765 SVM 0.743 K-NN 0.730	AUC values 12-month observation → LR 0.747	ML models performed better in detecting incident HF with a short observation window of 12–18 months
Prediction of outcomes						
Austin PC	*Biom J*	2012	30-day mortality	AUC values Regression tree 0.674 Bagged trees 0.713 Random forests 0.752 Boosted trees—depth one 0.769 Boosted trees—depth two 0.788 Boosted trees—depth three 0.801 Boosted trees—depth four 0.811	AUC values LR 0.773	Ensemble methods from the data mining and ML literature increase the predictive performance of regression trees, but may not lead to clear advantages over conventional LR models
Austin PC	*J Clin Epidemiol*	2010	In-hospital mortality	AUC values LR models Regression trees 0.620–0.651	AUC values LR 0.747–0.775	LR predicted in-hospital mortality in patients hospitalized with HF more accurately than did the regression trees
Awan SE	*ESC Heart Failure*	2019	30-day readmissions	AUC values MLP 0.62 Weighted random forest 0.55 Weighted decision trees 0.53 Weighted SVM models 0.54	AUC values LR 0.58	The proposed MLP-based approach is superior to other ML and regression techniques
Fonarow GC	*JAMA*	2005	In-hospital mortality	AUC values CART model (derivation cohort 68.7%; validation cohort 66.8%)	AUC values LR model (derivation cohort 75.9%; validation cohort 75.7%)	Based on AUC, the accuracy of the CART model (derivation cohort 68.7%; validation cohort 66.8%) was modestly less than that of the more complicated LR model (derivation cohort 75.9%; validation cohort 75.7%)

Table 1 (continued)

Author	Journal	Year	Outcome	Machine learning models	Conventional methods	Conclusion
				Comparison between machine learning and conventional methods		
Frizzell JD	JAMA Cardiol	2016	30-day readmissions	C-statistics Tree-augmented naive Bayesian network 0.618 RF 0.607 Gradient-boosted 0.614 Least absolute shrinkage and selection operator models 0.618	C-statistics LR 0.624	ML methods showed limited predictive ability
Golas SB	BMC Med Inform Decis Mak	2018	30-day readmissions	AUC values Gradient boosting 0.650 ± 0.011 Maxout networks 0.695 ± 0.016 Deep unified networks 0.705 ± 0.015	AUC values LR 0.664 ± 0.015	Deep learning techniques performed better than other traditional techniques
Hearn J	Circ Heart Fail	2018	Clinical deterioration (i.e., the need for mechanical circulatory support, listing for heart transplantation, or mortality from any cause)	AUC values ppVo2 0.800 (0.753–0.838) Staged LASSO 0.827 (0.785–0.867) Staged NN 0.835 (0.795–0.880) BxB LASSO 0.816 (0.767–0.866) BxB NN 0.842 (0.794–0.882)	AUC values CPET risk score 0.759 (0.709–0.799)	NN incorporating breath-by-breath data achieved the best performance
Kwon JM	Echocardiography	2019	Hospital mortality	AUC values Deep learning 0.913 RF 0.835	AUC values LR 0.835 MAGGIC score 0.806 GWTG score 0.783	The echocardiography-based deep learning model predicted in-hospital mortality among HD patients more accurately than existing prediction models
Phillips KT	AMIA Annu Symp Proc	2005	Mortality	AUC levels Nearest neighbor 0.823 NN 0.802 Decision tree 0.4975	AUC values Stepwise LR 0.734	Data mining methods outperform multiple logistic regression and traditional epidemiological methods
Mortazavi BJ	Circ Cardiovasc Qual Outcomes	2016	HF readmissions	C-statistics Boosting 0.678	C-statistics LR 0.543	Boosting improved the c-statistic by 24.9% over LR
Myers J	Int J Cardiol	2014	Cardiovascular death	AUC values Artificial NN 0.72 Cox PH models 0.69	AUC values LR 0.70	An artificial NN model slightly improves upon conventional models
Panahiazar M	Stud Health Technol Inform	2015	5-year mortality	AUC values RF 62% (baseline set), 72% (extended set) Decision tree 50% (baseline set), 50% (extended set) SVM 55% (baseline set), 38% (extended set) AdaBoost 61% (baseline set), 68% (extended set)	AUC values LR 61% (baseline set), 73% (extended set)	LR and RF return more accurate models
Subramanian D	Circ Heart Fail	2011	1-year mortality	C-statistics	C-statistics	

Table 1 (continued)

Author	Journal	Year	Outcome	Comparison between machine learning and conventional methods		Conclusion
				Machine learning models	Conventional methods	
				Ensemble model using gentle boosting with 10-fold cross-validation 84%	Multivariate LR model using time-series cytokine Measurements 81%	The ensemble model showed significantly better performance
Taslimitehrani V	*J Biomed Inform*	2016	5-year survival	Precision SVM 0.2, CPXR (log) 0.721 Recall SVM 0.5, CPXR (log) 0.615 Accuracy SVM 0.66, CPXR 0.809	Precision LR 0.513 Recall LR 0.506 Accuracy LR 0.717	CPXR is better than logistic regression, SVM, random forest and AdaBoost
Turgeman L	*Artif Intell Med*	2016	Hospital readmissions	AUC values NN 0.589 (train), 0.639 (test) Naïve Bayes 0.699 (train), 0.676 (test) SVM 0.768 (train), 0.643 (test) CART decision tree 0.529 (train), 0.556 (test) Ensemble models C5 0.714 (train), 0.693 (test) CHAID decision tree 0.671 (train), 0.691 (test)	AUC values LR 0.642 (train), 0.699 (test)	A dynamic mixed-ensemble model combines a boosted C5.0 model as the base ensemble classifier and SVM model as a secondary classifier to control classification error for the minority class
Wong W	*Scientific World Journal*	2003	Mortality (365 days models)	AUC values MLP 69% Radial basis function 67%	AUC values LR 60%	NNs are able to outperform the LR in terms of sample prediction
Yu S	*Artif Intell Med*	2015	30-day HF readmissions	AUC values Linear SVM 0.65 Poly SVM 0.61 Cox PH 0.63	AUC values Industry standard method (LACE) 0.56	The ML models performed better compared to standard method
Zhang J	*Int J Cardiol*	2013	Death or hospitalization	AUC values Decision trees 79.7%	AUC values LR 73.8%	Decision trees tended to perform better than LR models
Zhu K	*Methods Inf Med*	2015	30-day readmissions	AUC values RF 0.577 SVM 0.560 Conditional LR 1 = 0.576 Conditional LR 2 = 0.608 Conditional LR 3 = 0.615	AUC values Standard LR 0.547 Stepwise LR 0.539	LR after combining ML outperforms standard classification models

Table 1 (continued)

Author	Year	Outcome	Journal	Comparison between machine learning and conventional methods		Conclusion
				Machine learning models	Conventional methods	
Zolfaghar K	2013	HF readmissions	In *2013 IEEE International Conference on Big Data*	AUC values Multicare health systems model RF 62.25%	AUC values Multicare health systems model LR 63.78% Yale model LR 59.72%	ML random forest model does not outperform traditional LR model

AUC area under the receiver operating curve, *CPET* cardiopulmonary exercise test, *HF* heart failure, *LR* logistic regression, *ML* machine learning, *MLP* multilayer perceptron, *NN* neural networks, *NPV* negative prognostic value, *PH* proportional hazard, *PPV* positive prognostic value, *ppVo2* predicted peak oxygen uptake, *RF* random forest, *SENS* sensitivity, *SPE* specificity, *SVM* support vector machine

assessment technique to leverage multiple types of information from the electronic health records and predicted a medication plan for each new patient on a cohort of HF patients with area under the curve (AUC) of 0.74 [OS116]. Blecker et al. (2016) employed ML techniques and improved real-time identification of hospitalized patients with HF using both structured and unstructured electronic health records data, demonstrating high efficiency of ML analytics [OS112]. Although the accuracy varies, existing studies demonstrated that it is feasible to use ML to facilitate individualized interventions for hospitalized patients with HF.

Real-time identification of HF syndrome among hospitalized individuals is of great importance, as it likely to result in improvement of patient care and outcomes. Use of ML techniques for the identification of HF patients from electronic medical records and identification of HF patients with similar characteristics may lead to delivery of more tailored clinical care.

Decision support from clinical notes

Another meaningful consideration for the implementation of ML techniques is the extraction of important clinical data from diverse sources of narrative text. Our search found 3 studies regarding this aim (Supplementary Table 5). All studies were classified as high quality in the quality assessment (Supplementary Table 9).

Kim et al. (2013) improved HF information extraction through developing a natural language processing-based application to extract congestive HF treatment performance measures from echocardiographic reports (i.e., the source domain) with high recall and precision (92.4% and 95.3%, respectively) [OS117]. Meystre et al. (2017) demonstrated that the rich and detailed clinical information extracted from narrative notes may help improve management and outpatient treatment of HF patients [OS118]. Zhang et al. (2018) used random forest-based model to identify New York Heart Association (NYHA) class from clinical notes, with F-measure 93.78% [OS119].

The extracted clinical and medical information is critical to the understanding of a patient's clinical and medication status for better healthcare safety and quality. Furthermore, these algorithms can identify patients who do not receive appropriate HF medications and thus may help reduce the number of undertreated patients (Table 1).

Prediction of outcomes in left ventricular assist device (LVAD) patients

Our search retrieved 7 studies that focused on the prediction of outcomes in LVAD patients (Supplementary Table 6). All studies were classified as high quality in the quality assessment (Supplementary Table 9).

Loghmanpour et al. (2015) developed a Bayesian network-based risk stratification model to predict the short-term and long-term LVAD mortality with approximately 95% accuracy in predicting mortality at 30 days post-implant [OS120]. Mason et al. (2010) employed neural networks and waveform analysis methods for the non-invasive prediction of the pulsatile LVAD (HeartMate XVE (Thoratec Corporation, Pleasanton, CA)) pump failure within 30 days post-implantation [OS123]. Wang et al. (2012) found that the decision tree method can quantitatively provide improved prognosis of RV support through encoding the non-linear, synergic interactions among pre-operative variables, with an AUC of 0.87 [OS125]. The method can be used as an effective prognostic tool for triage of LVAD therapy. Lüneburg et al. (2019) used a U-net convolutional neural network for driveline tube segmentation and showed that the deep learning techniques can efficiently recognize LVAD on driveline exit site images [OS126]. Michaels and Cowger provide a review of the HF risk assessment as a referral guide for advanced HF therapies [24].

LVAD therapy is a life-saving treatment option as a destination therapy for end-stage HF patients who are ineligible for heart transplantation. However, the identification of high-risk patients who are prone to LVAD complications or adverse outcomes is crucial for patient selection who will benefit from this therapy (Table 1).

Prediction of cardiac resynchronization therapy response

Our search retrieved 5 studies regarding the role of ML techniques in CRT response prediction to overcome the challenge of significant nonresponse rates of current guidelines (Supplementary Table 7). All studies were classified as intermediate-high (intermediate: 1 study, high: 4 studies) quality in the quality assessment (Supplementary Table 9).

Kalscheur et al. (2018) employed random forest method to predict cardiac resynchronization therapy outcomes and showed that the ML method can utilize the information of bundle branch block morphology and QRS duration to derive the risk of the composite end point of all-cause mortality or HF hospitalization [OS128]. Feeny et al. (2019) analyzed CRT patients using ML techniques and showed that the performance can be improved incrementally by adding up to nine variables demonstrating that ML models have the potential to improve the shared decision-making in CRT [OS131].

Due to the high percentage of non-responders to CRT therapy [25], the reported performance of ML algorithms in the prediction of patients who will benefit from this treatment option is of great clinical importance. The implementation of ML algorithms in clinical practice is expected reduce the number of CRT patients who will not benefit by this high cost

treatment option who is related with higher rates of peri- and post-procedural complications.

Prediction of other HF-related outcomes

Our search also retrieved 8 studies regarding the role of ML techniques in alternative outcomes (i.e., prediction of treatment adherence [OS137, OS138], prediction of adherence use of remote HF monitoring systems [OS133], association of HF symptoms with depression [OS134], prediction of LV filling pressures [OS132], chronic HF management [OS135], prediction of missing data in wireless health projects [OS136], pathways delineation of death in patients with LVAD [OS139] (Supplementary Table 8). All studies were classified as intermediate-high quality (intermediate: 2 studies, high: 6 studies) in the quality assessment (Supplementary Table 9).

Specifically, Son et al. (2010) observed superior performance of support vector machine to predict medication adherence of patients with HF [OS138]. Karanasiou et al. (2016) found that ML methods can predict the medication/nutrition/physical activity adherence of patients with HF with an accuracy ranging from 0.82 to 0.91 [OS137]. Evangelista et al. (2017) predicted HF patient's adherence use of remote health monitoring systems with ML with an accuracy that ranged from 87.5 to 94.5% [OS133]. Graven et al. (2018) revealed the relationship between HF and depression with random forest algorithms [OS134]. Dini et al. (2010) developed an echo-Doppler decision model to predict left ventricular filling pressure in patients with HF [26]. Specifically, patients were correctly allocated according to pulmonary capillary wedge pressure with a sensitivity of 87% and specificity of 90% [OS132]. Seese et al. (2019) used a hierarchical clustering ML approach to create a descriptive model for delineating the pathways to death in patients with a LVAD, suggesting that there are two predominant types of adverse events which lead to mortality associated with multiorgan dysfunction (group 1: bleeding and infection and group 2: renal and respiratory complications) [OS139]. Another application of ML techniques has aimed to improve follow-up monitoring and management of chronic HF patients, following hospitalization [OS135].

Finally, a significant problem in the implementation of wireless health projects is the presence of missing data due to system misuse, non-use, and failure. Suh et al. (2011) adopted ML techniques to predict both non-binomial and binomial data missing data in wireless health projects with accuracies ranged between 85.7 and 98.5% [OS136].

Discussion

The main finding of our systematic review is that ML techniques may play a unique role in the contemporary management of HF patients. This includes classification of HF

patients into categories who will benefit from specific treatment strategies, discrimination of HF patients from no HF subjects or differential diagnosis of HF from other conditions with similar clinical presentation and prediction of outcomes in different patient populations, such as those with LVAD and CRT.

An important advantage of ML techniques compared to conventional prognostic algorithms is that ML techniques do not assume linear relationships between variables and outcomes, thus resulting in better performance in identifying individualized outcome predictions [27]. Recent data show that ML algorithms outperform logistic regression models in the prediction of HF outcomes [28–30]. Specifically, the better accuracy of ML algorithms compared to conventional tools has been demonstrated for the prediction of mortality in the setting of acute HF [30], mortality and hospitalization for HFpEF [29], and hospital readmissions [31]. Nonetheless, there is still room for improvement of ML techniques in predicting outcomes in these patients. For example, in a recent study, ML algorithms showed limited improvement in the prediction of all-cause mortality and HF hospitalization compared to traditional logistic regression analysis when using binary variables, while after including continuous variables, ML approaches generally performed better than logistic regression modeling [28].

Early diagnosis of the HF syndrome is the cornerstone for the early initiation of appropriate treatment and improving patients' prognosis. Therefore, existence of an objective, non-invasive, and low-cost tool for the diagnosis of HF is of great importance. Our search showed that ML techniques have a good discrimination performance in identifying HF patients by using different easily obtainable variables including demographics, clinical examination findings, echocardiographic parameters, electrocardiographic indices, etc. [OS24, OS26]. ML techniques can provide real-time identification of in-hospital patients with HF and extraction of important clinical as well as medication related information from unstructured data (i.e., clinical notes) that result in the improvement of HF management and treatment [OS113, OS114, OS117–119]. This is extremely important because hospitalized patients with HF often receive insufficient education and suboptimal transition of care planning, early post-discharge follow-up, or secondary prevention management, leading to high readmission rates, which in turn are associated with an unacceptably high rates of morbidity and mortality [OS153].

Classification of HF patients into subtypes with different prognosis and treatment needs is clinically important. Recent guidelines classify HF patients into HFpEF, HF with mid-range ejection fraction (HFmrEF) and HFrEF mainly using EF values [3]. However, this classification has some disadvantages especially due to the definition of HFpEF and HFmrEF patients. ML-based models can sufficiently classify HF patients (including the gray zone) using different clinical

variables [OS18–21] The clinical implications for patient-specific classification of HF patients cannot be overemphasized. For example, in light of the favorable results of sacubitril-valsartan in women, but not men with HFpEF [OS154], it has been argued that a different cut-off value for EF should be used in women vs. men. In the future, ML techniques may be able to apply sex-specific classification criteria for HF patients, which will facilitate clinical decisions regarding implementation of appropriate therapy. Another example refers to phenomapping of patients with HFpEF to different phenotypic groups, with different prognosis and response to pharmacologic interventions, such as spironolactone [OS155]. Given that no pharmacologic therapy has been shown to improve clinical outcomes in HFpEF [OS156], identification of a subset of patients with HFpEF who might benefit from certain medications becomes of utmost importance.

Our search showed that ML techniques have been applied successfully in the identification of high-risk patients and in the early initiation of appropriate treatment with the aim of reducing HF related mortality and hospitalizations. Different risk scores have been proposed for the identification of high-risk patients [OS140–142]. Specifically, Ahmad et al. [19] implement ML techniques to classify HF patients into four groups using the eight strongest derived predictors (age, creatinine, hemoglobin, weight, heart rate, systolic blood pressure, mean arterial pressure, and income) of mortality. This type of classification proved to be superior to current classification methods of HF patients, in terms of prognostication and response to medications, and may replace patient classification in different clinical settings.

Prediction of patients who may respond to CRT therapy is of great importance [OS143, OS144]; however, approximately 30% of CRT recipients do not respond to this treatment [OS145]. ML techniques have been successfully implemented in creating score models with improved measure estimates regarding the prediction of CRT responders, compared to conventional techniques [OS146–148]. As a result, risk scores produced by employing ML techniques can become the cornerstone for appropriate CRT candidate selection. In addition, ML techniques have been implemented with success in predicting outcomes of LVAD patients, implying that ML techniques may play an even important role in the decision-making regarding LVAD candidates in the future.

Furthermore, our review also found that ML techniques have been applied in other aspects of the management of HF patients, e.g., ML techniques can be applied in the identification of patients who may adhere to the prescribed medications or may need additional measured for treatment adherence [OS137, OS138]. Another significant role of ML techniques in the management of HF patients is the identification of HF patients who are at high risk for other comorbidities (i.e., depression) [OS134], or in remote HF monitoring systems resulting in improvement of HF clinical outcomes [OS133,

OS149–151]. Effective ML techniques have been implemented to protect implantable devices from cybersecurity attacks [OS152]. Since ML algorithms have been implemented in identifying risk factors for predicting treatment-related discontinuations in various clinical settings [32, 33], identification of HF patients who are at increased risk of treatment discontinuation because of drug related adverse effects may be another important area for ML algorithm implementation.

Finally, while a series of critical issues (i.e., the role of physicians and patients in the decision-making process, reliability, transparency, accountability, liability, handling of personal data, different kinds of bias, continuous monitoring of AI adverse events/system failure, cybersecurity, and system upgrading) have led to skepticism with respect to the implementation and adoption of AI algorithms in clinical practice, the ML impact on health economics, is expected to be beneficial to both patients and health insurance providers, justified by an earlier and more accurate diagnosis, reduction of unnecessary expensive diagnostic exams, and selection of optimal candidates for expensive treatment options. Consequently, the implementation of ML algorithms in clinical practice is a complex process and an integrated regulatory framework for the research, development and adoption of ML in medicine, is needed.

Study limitations

The following limitations should be considered: a quantitative synthesis was inappropriate because of the heterogeneity between the included studies regarding the reported outcomes and measured estimates. Therefore, the reported outcomes in each included study are prone to different kinds of biases mainly depended on the ML method that was used. Moreover, the outcomes of a number of studies should be interpreted with caution because of the small number of patients. Finally, our results should be interpreted in light of the fact the tool for quality assessment of the included studies is relatively new and has not been validated in multiple studies.

Conclusions

ML techniques play an important role in different aspects of the management of HF patients and show inspiring promise in the efficient construction of methodologies aiming to improve HF diagnosis, management, and prediction of outcomes in different clinical settings, with generally an improved performance compared to conventional techniques.

While a regulatory framework for the implementation of ML in clinical practice is needed, intelligent analysis of health data with ML techniques still acts as auxiliary decisional role and at the moment cannot replace clinical cardiologists.

Availability of data and material Not applicable

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Code availability Not applicable

References

1. Yancy CW, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Colvin MM et al (2017) 2017 ACC/AHA/HFSA Focused Update of the 2013 ACCF/AHA Guideline for the Management of Heart Failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. Circulation 136:e137–ee61

2. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS et al (2016) 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 Heart J 37:2129–2200

3. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS et al (2016) 2016 ESC guidelines for the diagnosis and treatment of acute and chronic heart failure. Rev Esp Cardiol (Engl Ed) 69:1167

4. Ponikowski P, Anker SD, AlHabib KF, Cowie MR, Force TL, Hu S et al (2014) Heart failure: preventing disease and death worldwide. ESC Heart Failure 1:4–25

5. Writing Group M, Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ et al (2016) Heart disease and stroke statistics— 2016 update: a report from the American Heart Association. Circulation 133:e38–e360

6. Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, Delling FN, Djousse L, Elkind MSV, Ferguson JF, Fornage M, Jordan LC, Khan SS, Kissela BM, Knutson KL, Kwan TW, Lackland DT, Lewis TT, Lichtman JH, Longenecker CT, Loop MS, Lutsey PL, Martin SS, Matsushita K, Moran AE, Mussolino ME, O'Flaherty M, Pandey A, Perak AM, Rosamond WD, Roth GA, Sampson UKA, Satou GM, Schroeder EB, Shah SH, Spartano NL, Stokes A, Tirschwell DL, Tsao CW, Turakhia MP, VanWagner L, Wilkins JT, Wong SS, Virani SS, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee (2019) Heart disease and stroke statistics—2019 update: a report from the American Heart Association. Circulation 139:e56–e528

7. Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, Ayuso-Mateos JL, Haro JM, Chatterji S, Georgousopoulou EN, Pitsavos C, Panagiotakos DB (2018) Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. BMC Med Res Methodol 18:179

8. Wang S, Summers RM (2012) Machine learning and radiology. Med Image Anal 16:933–951

9. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H (2018) eDoctor: machine learning and the future of medicine. J Intern Med 284:603–619

10. Sevakula RK, Au-Yeung WM, Singh JP, Heist EK, Isselbacher EM, Armoundas AA (2020) State-of-the-art machine learning techniques aiming to improve patient outcomes pertaining to the cardiovascular system. J Am Heart Assoc 9:e013924

11. Kerut EK, To F, Summers KL, Sheahan C, Sheahan M (2019) Statistical and machine learning methodology for abdominal aortic aneurysm prediction from ultrasound screenings. Echocardiography 36:1989–1996

12. Le S, Hoffman J, Barton C, Fitzgerald JC, Allen A, Pellegrini E et al (2019) Pediatric severe sepsis prediction using machine learning. Front Pediatr 7:413

13. Erickson BJ (2017) Machine learning: discovering the future of medical imaging. J Digit Imaging 30:391

14. Alizadehsani R, Roshanzamir M, Abdar M, Beykikhoshk A, Khosravi A, Panahiazar M, Koohestani A, Khozeimeh F, Nahavandi S, Sarrafzadegan N (2019) A database for using machine learning and data mining techniques for coronary artery disease diagnosis. Scientific Data 6:227

15. Serag A, Ion-Margineanu A, Qureshi H, McMillan R, Saint Martin MJ, Diamond J, O'Reilly P, Hamilton P (2019) Translational AI and deep learning in diagnostic pathology. Front Med 6:185

16. Wu C, Zhao X, Welsh M, Costello K, Cao K, Abou Tayoun A et al (2019) Using machine learning to identify true somatic variants from next-generation sequencing. Clin Chem 66(1):239–246

17. Quitadamo LR, Cavrini F, Sbernini L, Riillo F, Bianchi L, Seri S, Saggio G (2017) Support vector machines to detect physiological patterns for EEG and EMG-based human-computer interaction: a review. J Neural Eng 14:011001

18. Mo X, Chen X, Li H, Li J, Zeng F, Chen Y, He F, Zhang S, Li H, Pan L, Zeng P, Xie Y, Li H, Huang M, He Y, Liang H, Zeng H (2019) Early and accurate prediction of clinical response to methotrexate treatment in juvenile idiopathic arthritis using machine learning. Front Pharmacol 10:1155

19. Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B et al (2018) 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 7(8):e008081. https://doi.org/10.1161/JAHA.117.008081

20. Soboczenski F, Trikalinos TA, Kuiper J, Bias RG, Wallace BC, Marshall IJ (2019) Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. BMC Medical Informatics and Decision Making 19:96

21. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6:e1000097

22. Qiao N (2019) A systematic review on machine learning in sellar region diseases: quality and reporting items. Endocr Connect 8(7):952–960

23. Webb GI, Zheng Z (2004) Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. IEEE Trans Knowl Data Eng 16:980–991

24. Michaels A, Cowger J (2019) Patient selection for destination LVAD therapy: predicting success in the short and long term. Current Heart Failure Rep 16:140–149

25. Versteeg H, Schiffer AA, Widdershoven JW, Meine MM, Doevendans PA, Pedersen SS (2009) Response to cardiac resynchronization therapy: is it time to expand the criteria? Pacing and Clinical Electrophysiology: PACE 32:1247–1256

26. Dini FL, Ballo P, Badano L, Barbier P, Chella P, Conti U, de Tommasi SM, Galderisi M, Ghio S, Magagnini E, Pieroni A, Rossi A, Rusconi C, Temporelli PL (2010) Validation of an echo-Doppler decision model to predict left ventricular filling pressure in patients with heart failure independently of ejection fraction. Eur J Echocardiogr 11:703–710

27. Gibson WJ, Nafee T, Travis R, Yee M, Kerneis M, Ohman M, Gibson CM (2020) Machine learning versus traditional risk stratification methods in acute coronary syndrome: a pooled randomized clinical trial analysis. J Thromb Thrombolysis 49:1–9

28. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S (2020) Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. JAMA Netw Open 3:e1918962

29. Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM (2020) Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. JACC Heart Failure 8:12–21

30. Kwon JM, Kim KH, Jeon KH, Lee SE, Lee HY, Cho HJ, Choi JO, Jeon ES, Kim MS, Kim JJ, Hwang KK, Chae SC, Baek SH, Kang SM, Choi DJ, Yoo BS, Kim KH, Park HY, Cho MC, Oh BH (2019) Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLoS One 14:e0219302

31. Turgeman L, May JH (2016) A mixed-ensemble model for hospital readmission. Artif Intell Med 72:72–82

32. Westborg I, Rosso A (2018) Risk factors for discontinuation of treatment for neovascular age-related macular degeneration. Ophthalmic Epidemiol 25:176–182

33. Pradier MF, McCoy TH Jr, Hughes M, Perlis RH, Doshi-Velez F (2020) Predicting treatment dropout after antidepressant initiation. Transl Psychiatry 10:60