Machine learning and statistical methods for predicting mortality in heart failure

Dineo Mpanya [1,5] · Turgay Celik [2,5] · Eric Klug [3] · Hopewell Ntsinjana [4]

Abstract

Heart failure is a debilitating clinical syndrome associated with increased morbidity, mortality, and frequent hospitalization, leading to increased healthcare budget utilization. Despite the exponential growth in the introduction of pharmacological agents and medical devices that improve survival, many heart failure patients, particularly those with a left ventricular ejection fraction less than 40%, still experience persistent clinical symptoms that lead to an overall decreased quality of life. Clinical risk prediction is one of the strategies that has been implemented for the selection of high-risk patients and for guiding therapy. However, most risk predictive models have not been well-integrated into the clinical setting. This is partly due to inherent limitations, such as creating risk predicting models using static clinical data that does not consider the dynamic nature of heart failure. Another limiting factor preventing clinicians from utilizing risk prediction models is the lack of insight into how predictive models are built. This review article focuses on describing how predictive models for risk-stratification of patients with heart failure are built.

Keywords Heart failure · Predict · Models · Machine learning · Deep learning

Abbreviations

ANN	Artificial neural networks
AUC	Area under the curve
CI	Confidence interval
INTER-CHF	International Congestive Heart Failure
KNN	K-nearest neighbors
LR	Linear regression
LVEF	Left ventricular ejection fraction
NB	Naïve Bayes
NYHA	New York Heart Association
PRAISE	Prospective Randomized Amlodipine Survival Evaluation
SHF	Seattle Heart Failure
SVM	Support vector machine

✉ Dineo Mpanya
    Dineo.Mpanya@wits.ac.za

1   Department of Internal Medicine, Division of Cardiology, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand and the Charlotte Maxeke Johannesburg Academic Hospital, 17 Jubilee Road, Parktown, Johannesburg, Gauteng 2193, South Africa

2   Faculty of Engineering and Built Environment, School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, South Africa

3   Netcare Sunninghill, Sunward Park Hospitals and Division of Cardiology, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand and the Charlotte Maxeke Johannesburg Academic Hospital, Johannesburg, South Africa

4   Department of Paediatrics and Child Health, School of Clinical Medicine, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

5   Institute of Data Science , University of the Witwatersrand , Johannesburg, South Africa

Introduction

Heart failure is estimated to affect at least 64.3 million individuals globally [1]. Despite the availability of therapeutic agents and medical devices that aim to prolong survival, many heart failure patients still experience progressive disease and eventually death, particularly in developing countries. The International Congestive Heart Failure (INTER-CHF) study recruited 5823 patients with heart failure and studied them over 1-year [2]. Forty-two percent of these patients were in New York Heart Association (NYHA) functional class III or IV. In the entire study cohort, the 1-year all-cause mortality was 16.5% (95% CI 15.4–17.6). However, patients residing in

Africa had a higher mortality of 34% (95% CI 30.2–37.4) within 1 year of follow-up [2]. The high mortality rate emphasizes the need for new, innovative strategies such as predictive models that aim to risk-stratify patients with heart failure, ensuring that high-risk individuals are identified early and subjected to interventions that delay disease progression.

Both statistical and recently, machine learning techniques are used to create risk predicting models. There is a significant overlap between traditional statistical methods and machine learning algorithms, and one method cannot be deemed superior to the other. While the focus with statistics is to make inferences based on the sample studied, machine learning focuses on learning patterns in datasets and subsequently making predictions based on the learned observations. This review aims to describe what predictive models are, how they are created, and their strengths and weaknesses.

Risk prediction models

Before describing what predictive models are, we will define the verb "predict" and the noun "model." The Oxford English dictionary defines "predict" as "stating that a specified event will happen in the future." A model is defined as a "simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions" [3].

Predictive models are built using both statistical methods and machine learning techniques. However, a paradigm shift has occurred, where machine learning algorithms are becoming popular because of their ability to consider complex associations between clinical parameters. The sophisticated mathematical formulas used in machine learning for automated medical image interpretation have also led to increased confidence in using machine learning algorithms in the health sector. Furthermore, a wider variety of machine learning algorithms is available for classification and regression tasks, allowing the user to handpick the best performing model.

Heart failure predictive models created with statistical methods

Rahimi et al. identified 64 multivariate statistical models used in heart failure patients published between January 1995 and March 2013 [4]. Another systematic review by Di Tanna et al. found 58 multivariate statistical models predicting outcomes in patients with heart failure, published between March 2013 and May 2018 [5]. Seventeen of the models identified by Di Tanna et al. predicted all-cause mortality, nine predicted cardiovascular death, and the rest predicted heart failure hospitalization and composite endpoints [5]. The next section will describe the techniques used to create multivariate linear and logistic regression models.

Linear regression

Linear regression (LR) is used to predict a numeral or continuous outcome. In simple linear regression, the dependent variable (outcome) is featured in the y-axis and the independent variable in the x-axis (Fig. 1). The dependent variable only changes when there is an adjustment of the values in the x-axis. A scatterplot is used to establish a relationship between the dependent and independent variables. A correlation coefficient is then calculated if a linear relationship exists between the variables in the x and y coordinates. The regression line or straight line, is then plotted through the points [6]. The straight line is given by the equation: $y = a + \beta \times x$, where y is the dependent variable; a is the y-intersect of the line; β is the slope of the line, and x is the explanatory or independent variable [6]. The predicted value (y), is found by sketching a line that originates in the x-axis, connects to the regression line, and ultimately a point in the y-axis (Fig. 1). For example, the weight, plotted in the x-axis, will predict a numerical outcome such as height (y). Multiple linear regression analysis involves adding two or more predictor variables concurrently and adjusting their respective regression coefficients [6]. This statistical method is used to determine the relative contribution of each of the predictors. The formula for multiple linear regression is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \ldots \beta_k x_k + \varepsilon$, wher y is the dependent variable, x is the independent variable, β_0 is the y-intercept, ε the residual, and β_k is the slope coefficient for each predictor variables [7].

Logistic regression

Logistic regression predicts a binary outcome by estimating the probability of an output belonging to a particular class. When building a predictive model using logistic regression, the first step is to utilize the univariable analysis to discover the unadjusted association between each variable and the predicted class [8]. Categorical and continuous variables, usually with a p value less than 0.25, are selected and included in the multivariable analysis [9].

The second step entails placing all the selected variables with a p value less than 0.25 in the multivariable logistic regression model. In the final multivariable regression model, variables with a p value more than 0.05 are subsequently eliminated. In the third step, smoothed scatter plots are created to assess the relationship between continuous variables and the logit scale outcome. Step four entails evaluating for potential interactions between the selected predictor variables. When building a model predicting mortality, the predicted variable will be plotted in the y-axis and classified into two categories: mortality ($y = 1$) and no mortality ($y = 0$). Predictor variables such as age, left ventricular ejection fraction (LVEF), NYHA, and serum sodium levels will be placed in the x-axis. The log odds or logit of mortality will be calculated

Fig. 1 Linear regression model
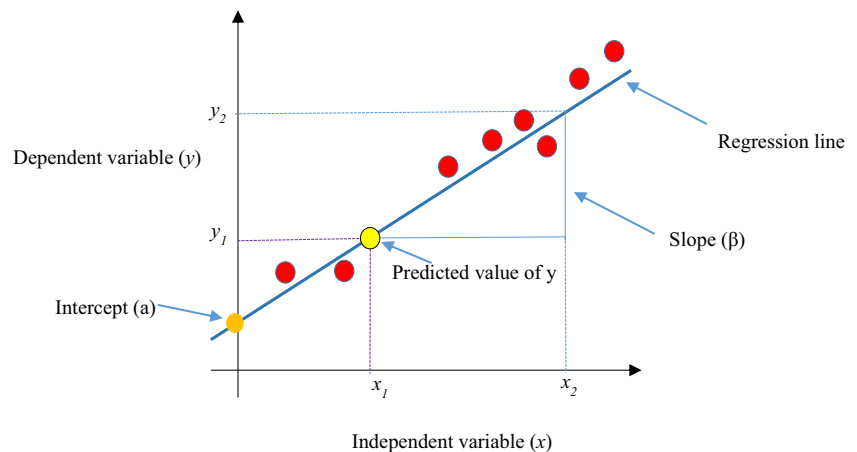
using the equation: $\log(p/1\text{-}p) = a + b \times x$, where p is the probability, a is the intercept, b is the regression coefficient of x, and x is the predictor variable [10].

Cox regression

The Seattle Heart Failure (SHF) model is one of the best performing models built using data from 1125 heart failure patients. A Cox regression model was used to select predictors. The predicted versus actual 1-year survival rates in the derivation cohort were 73.4% versus 74.3%. In the five validation cohorts, the predicted versus actual survival rates were 90.5% versus 88.5%, 86.5% versus 86.5%, 83.8% versus 83.3%, 90.9% versus 91.0%, and 89.6% versus 86.7% [11]. The data used to develop the model was obtained from a registry of carefully screened patients from the Prospective Randomized Amlodipine Survival Evaluation (PRAISE) trial, possibly limiting generalizability in the wider population. The online SHF risk calculator may be cumbersome for use in a busy practice as it requires an imputation of 24 clinical and laboratory parameters when estimating the risk of one and 5-year mortality.

Cox proportional hazard is a regression model used to measure an association between survival time and predictor variables. Patients are tracked from the time they enter the study until the occurrence of an event. An event is defined as the occurrence of a disease, relapse from therapy, discharge from hospital, or death. Multiple predictors or covariates of survival are added to the model to estimate their influence on survival, analogous to a multivariate regression model. The Cox proportional hazard model highlights how survival differs between the groups of patients studied by estimating the hazard ratio, unlike the Kaplan-Meier survival curve, which only assesses whether survival between groups differs (log-rank test with a p value < 0.05).

The Cox proportional hazard model, as the name implies, assumes that the covariates remain constant over time. This assumption has been challenged by Giolo et al., who suggested that heart failure prognostication models should be built using covariates with time varying-effects [12]. In their study, 500 patients with heart failure were followed between 2002 and 2006. Survival was analyzed using Cox proportional hazard, variations of the Cox's model and the Aalen's additive model. They used the patients' age, serum sodium levels, haemoglobin, serum creatinine, and the LVEF as covariates. A high haemoglobin and LVEF were associated with a lower risk of mortality, while the impact of age and sodium remained constant over time [12].

Heart failure predictive models created with machine learning techniques

The fundamental principle of machine learning is learning from data and predicting an output, where $y = f(x)$. The input (x) is any group of features, commonly known as variables used to predict an output (y). A machine is given a learning task under any of the following conditions: supervised, unsupervised, semi-supervised, and through reinforcement learning [13]. A computer can learn from a dataset with a labeled output or outcome in supervised learning. For example, in a dataset comprising of patients with heart failure, we may include a column header, mortality as the output. Within the column "mortality," we then capture whether the patient demised or not by adding yes or no, respectively.

Unsupervised learning takes place when a computer learns from unlabelled datasets. The data is subsequently clustered into similar categories. Ahmad et al. used unsupervised learning in a large clinical data set of more than 40000 Swedish heart failure patients to separate patients into four groups [14]. After 1 year of follow-up, there was a marked difference in outcomes per cluster, with 1-year survival rates of 69%, 77%, 92%, and 93% [14].

In contrast, when the same heart failure patients were categorized based on the LVEF, only slight differences in 1-year

survival (80%, 81%, 83%, and 84%) were noted [14]. Semi-supervised learning occurs when learning takes place using a combination of labeled and unlabelled datasets [13]. In reinforcement learning, a computer learns by trial and error by performing tasks and receiving penalty and reward scores for inaccurate and accurate responses, respectively [15].

How are predictive models built with machine learning techniques?

Prior to building a predictive model using machine learning, the data must be cleaned. When cleaning continuous variables, one can create box plots and verify the values of all the outliers. The next step is to train the model. Training involves splitting the dataset, where the majority of the dataset (e.g., 70%) is used to train the model. During the training process, the selected algorithm learns patterns in the dataset. The remaining dataset is used for testing and validating the model. During the test phase, the labeled output variable, for example, "mortality" is hidden, and the machine learning algorithm attempts predicting whether the patient demised or not.

Numerous supervised learning algorithms are available and are selected based on the task at hand, be it for classification of a binary outcome or regression analysis of continuous features. Some of the commonly used machine learning algorithms for classification or regression analysis include decision trees, random forests, support vector machines (SVM), K-nearest neighbors (KNN), and Naïve Bayes (NB).

Decision trees

A decision tree is a simple tree-shaped algorithm used to predict a categorical (binary) or numerical outcome. To illustrate how a decision tree classifies data with a categorical outcome, we will use data from patients with heart failure to build a model predicting mortality, as shown in Fig. 2. Each decision tree comprises of branches, leaf, and root nodes [16]. The root node is the most superiorly located node, and the leaf or terminal node is the final node that carries the decision. Each branch of the tree represents a decision, occurrence, or reaction. Data is then partitioned into subsets containing similar values by using mathematical equations such as the Gini index, Chi-square, information gain, and a reduction in variance [17].

Based on the known predictors of mortality in heart failure patients, the decision tree algorithm can categorize a patient with heart failure into class A (high risk) and class B (low risk) (Fig. 2). Decision trees require minimal data preparation before building a predictive model, and their performance is not affected by non-linear data distribution. The main disadvantage is overfitting, which occurs when a model demonstrates high-accuracy levels during training, makes inaccurate

predictions during testing, or has significantly lower test accuracy than the training accuracy [18].

Random forest

A random forest algorithm is an ensemble decision tree that operates by constructing multiple decision trees, creating a forest. When building a classification model using a random forest algorithm, the first step is to select random samples from a given dataset. A decision tree is then constructed for each sample. Predicted results are obtained from each decision tree. Lastly, the most voted class from the predictions made by each decision tree is selected as the final predicted class [19]. In a regression model, where the output class is numerical, the mean or average value of the predicted output is used. The random forest algorithm is ideal for handling datasets with missing values. It also performs well on large datasets and can rank features in the order of importance. The random forest algorithm's main disadvantage is that it is computationally expensive, requiring more training time than most algorithms [18].

Support vector machines

Support vector machines (SVM) perform classification and regression analysis for both linear and non-linear data. The main advantage of SVM lies in their ability to classify more than two classes within the same dataset simultaneously. Support vector machines categorize data by focusing on the observations on the edge of each class (Fig. 3). A line or hyperplane is then placed between the classes. The maximum distance between the classes (support vectors) and the hyperplane is then chosen [20]. When the data cannot be linearly separable, the data is transformed using a kernel function from low dimension to high dimensional structure, rendering the data separable.

K-nearest neighbors

The K-nearest neighbors (KNN) algorithm does not require training. It is used for predicting a binary or continuous output. Data is separated into clusters, and the number of nearest neighbors is specified by stating the value of "K", a constant [13]. If $K = 3$, three nearest neighbors are selected based on the distance measured between the new data point and nearest neighbors. If $K = 6$, the distance between the new data and the six nearest neighbors is calculated. The new data point is then placed in the class with the majority of votes. Figure 4 depicts how a KNN algorithm classifies a new data point. The biggest challenge with using K-nearest neighbors lies in deciding on the optimal number of neighbors to consider when classifying a new data point. K-nearest neighbors also do not perform well in an imbalanced dataset, since the model will
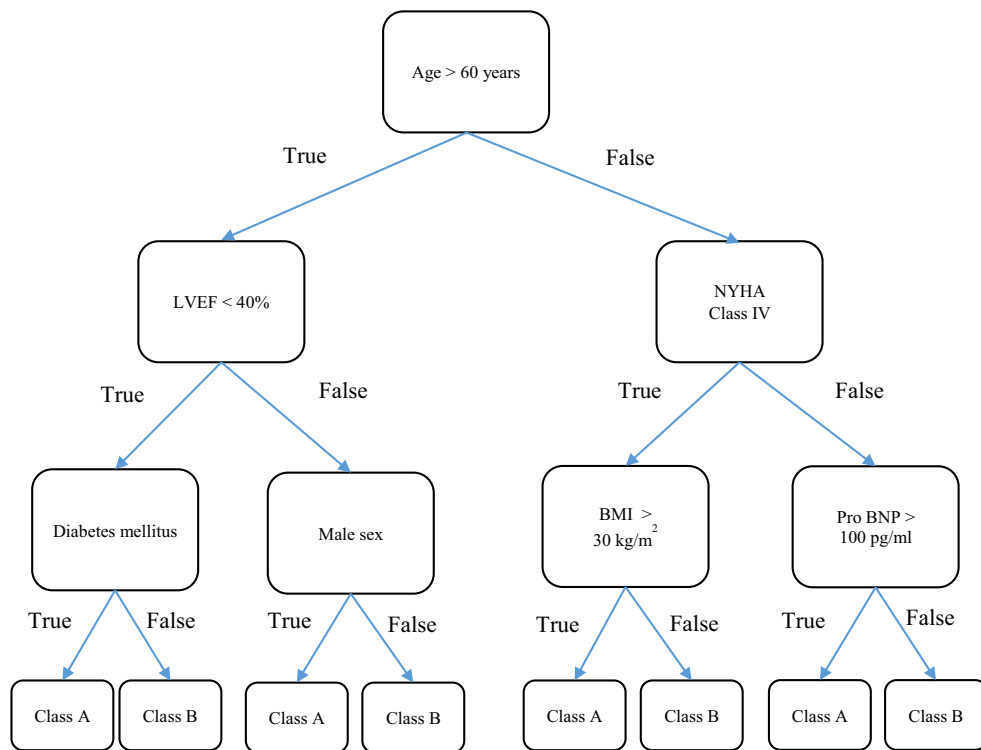
give preference to the class with the higher number of observations [18].

Naïve Bayes

Naïve Bayes is a probabilistic method borrowed from statistics, based on the Bayes' theorem, which describes the probability of an event occurring based on previous knowledge of the conditions associated with that particular event [21]. For classification tasks, a model learns the probabilities of an object for belonging to a specific class. The name "naïve" is derived from the fact that a Naïve Bayes model assumes that the occurrence of a particular feature is not dependent on the occurrence of other features [22]. Pakhomov et al. used a Naïve Bayes classifier and another classification algorithm, perceptron to train models capable of identifying patients'
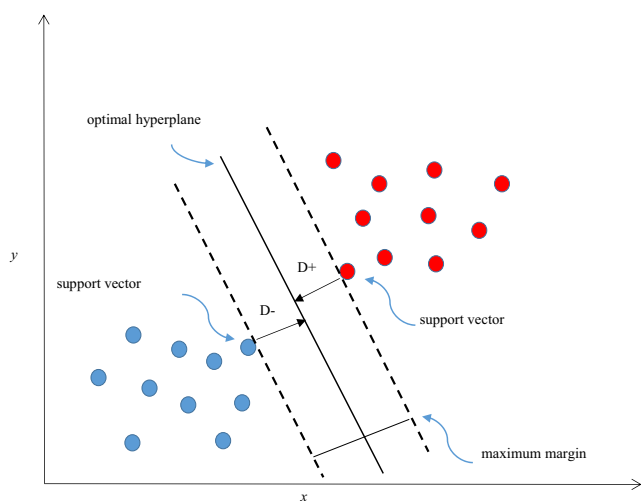
Fig. 3 Support vector machine algorithm depicting classification of linearly separable data using a decision boundary, the optimal hyperplane. D− and D+ represent the shortest distance to the closest negative and positive classes, respectively and the optimal hyperplane
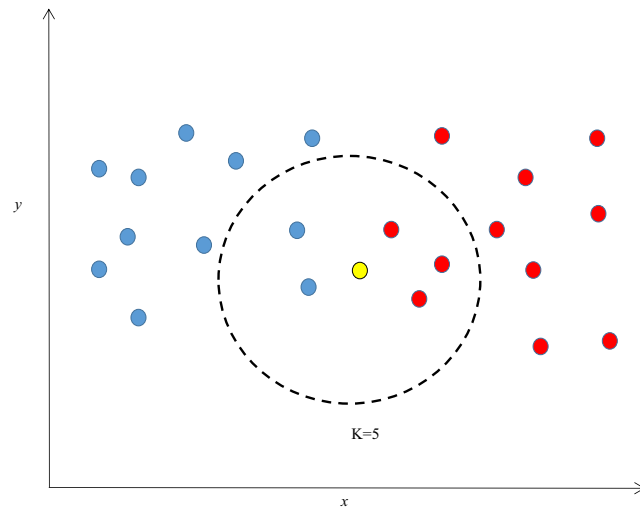

Fig. 4 K-nearest neighbor algorithm. The circle shows five data points closest to the new data point (yellow). Since the majority of the neighbors (3 out of 5) are in the red class, the new data point will be classified as belonging to the red class

clinical notes with a diagnosis of congestive cardiac failure [23]. The Naïve Bayes classifier yielded better recall on positive samples (95% vs 86%) but had low accuracy (57% vs 65%) [23]. Although the Naïve Bayes classifier is more comfortable implementing and has a lesser training period, its main limitation is the assumption that all features are mutually independent. For example, in the study by Pakhomov et al., the terms "heart" and "failure" were considered to occur independently of each other, rendering the classifier unattractive for text categorization [23].

Artificial neural networks

Artificial neural networks (ANNs) resemble neurons, the functional unit of the central nervous system. Each ANN is characterized by an input layer, a hidden layer, and an output layer [24]. The input layer represents features extracted by the model. A feature can be any continuous or categorical variable, such as the left ventricular ejection fraction, weight, or gender. The features in the input layer are used to predict an output, represented by one or more categories. The nodes displayed in the input layer communicate with each node in the hidden layer (Fig. 5). Initially, an arbitrary value between zero and one is used to grade the connection between the input and hidden layers. The weighted sum of the input layers' signal is then passed through the activation function. The sigmoid function transforms each inputs' weighted sum with a negative value to a value close to zero. Positive values are transformed into a value close to one. Most neural networks are now designed with more than one hidden layer, increasing their capacity to classify non-linear data.

Kwon et al. studied 2165 patients with acute heart failure to predict in-hospital mortality, 1-year and 3-year risk of mortality using deep neural networks, random forest, logistic regression, support vector machines, and a Bayesian network. Deep neural networks had the largest area under the curve (AUC) of 0.880, 0.782, and 0.813 for predicting in-hospital, 1-year and 3-year risk of mortality, respectively [25].
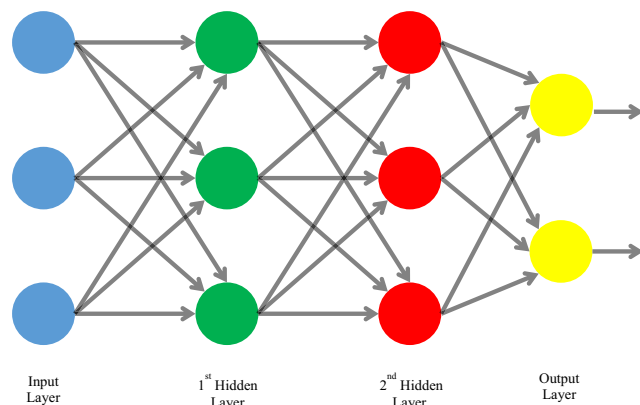
Fig. 5 An artificial neural network showing nodes in the input layer communicating with each node in the first hidden layer

Model performance

Models predicting a binary outcome

The AUC is the most common parameter used to evaluate the performance of classification models [26]. Models with good discriminatory abilities have larger AUC, generally greater than 0.70, whereas those with AUC less than 0.50 lack discriminatory abilities [26]. Model performance is also evaluated with a confusion matrix, allowing one to calculate accuracy, precision, sensitivity, and specificity. Other performance metrics include the F measure or F score, geometric mean, and logarithmic loss [27].

Models predicting a continuous outcome

The overall fit of a linear regression model is evaluated with the r-squared (r^2). Simply put, the r^2 is a measure of how much the prediction error is reduced, relative to how much potential error there is [28]. The error is estimated with indices such as the mean absolute error, mean squared error, and the root mean squared error [29].

Model flexibility

Model overfitting and underfitting are the most common problems encountered when evaluating performance. Overfitting occurs when a model shows high accuracy scores during training and low accuracy scores during validation. Model overfitting is minimized by adding more data to the training set and reducing the number of layers in the neural network. Underfitting occurs when the model fails to classify data or make predictions during the training phase. A low accuracy score and a high loss identify a model that is underfitting [30].

Limitations of statistical and machine learning predictive models

Despite numerous models predicting outcomes in patients with heart failure, only a few are easily accessible online as risk score calculators [11, 31, 32]. Although the lack of external validation of the predictive models is the driving factor for the limited availability, several inherent limitations exist for the relatively unhurried integration of predictive models into clinical practice.

Some clinicians find risk calculation cumbersome [33]. Also, most risk prediction models were designed before new heart failure drugs that further reduce mortality and the rate of heart failure hospitalizations were discovered [34, 35].

Class imbalance refers to the disproportionality between the classes of data used to train the predictive

model [36], a common problem that is not unique to medical data. When the training data with the negative outcome (e.g., dead) has significantly fewer observations than the class (e.g., alive) with the majority of observations, the classification algorithm is inclined to favor the majority class. This poses problems as the minority class, which bears the outcome of interest, will have a low accuracy score. Fortunately, it is possible to address class imbalance problems by manipulating data, algorithms, or both [36].

Conclusion

Both machine learning techniques and statistical methods should be employed when creating predictive models. Prior to clinical application, the best performing model should be externally validated in a cohort of patients not used for model derivation.

Future directions and recommendations

Unsupervised machine learning algorithms should be considered for the detection of data patterns not recognized by clinicians. Incorporating genomic data, biomarkers, imaging data, and features representing the patients' socioeconomic and psychological standing can further improve the accuracy levels of predictive models. Risk predictive models should be created with dynamic data and ideally be embedded in automated data capturing tools. The increase in training data volume will ultimately lead to a robust predictive model with higher accuracy levels.

Compliance with ethical standards

References

1. Lippi G, Sanchis-Gomar F (2020) Global epidemiology and future trends of heart failure. AME Med J 2020(5):15
2. Dokainish H, Teo K, Zhu J, Roy A, AlHabib KF, ElSayed A et al (2017) Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study. Lancet Glob Health 5(7):e665–e672
3. Concise Oxford English Dictionary, 11th ed. 2006. Oxford University Press, New York
4. Rahimi K, Bennett D, Conrad N, Williams TM, Basu J, Dwight J, Woodward M, Patel A, McMurray J, MacMahon S (2014) Risk prediction in patients with heart failure: a systematic review and analysis. JACC Heart Fail 2(5):440–446
5. Di Tanna GL, Wirtz H, Burrows KL, Globe G (2020) Evaluating risk prediction models for adults with heart failure: A systematic literature review. PLoS ONE 15(1): e0224135. https://doi.org/10.1371/journal.pone.0224135
6. Schneider A, Hommel G, Blettner M (2010) Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 107(44):776–782
7. Alexopoulos EC (2010) Introduction to multivariate regression analysis. Hippokratia 14(Suppl 1):23–28
8. Zhang Z (2016) Model building strategy for logistic regression: purposeful selection. Ann Transl Med 4(6):111–111
9. Mickey RM, Greenland S (1989) The impact of confounder selection criteria on effect estimation. Am J Epidemiol 129(1):125–137
10. Pandis N (2017) Logistic regression: part 1. Am J Orthod Dentofac Orthop 151(4):824–825
11. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M (2006) The Seattle Heart Failure Model: prediction of survival in heart failure. Circulation 113(11):1424–1433
12. Giolo SR, Krieger JE, Mansur AJ, Pereira AC (2012) Survival analysis of patients with heart failure: implications of time-varying regression effects in modeling mortality. PLoS One 7(6): e37392–e37392
13. Awad M, Khanna R (2015) Machine learning. In: Efficient learning machines: theories, concepts, and applications for engineers and system designers. Apress, Berkeley, pp 1–18
14. 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
15. Choy G, Khalilzadeh O, Michalski M, Do S, Samir AE, Pianykh OS et al (2018) Current applications and future impact of machine learning in radiology. Radiology 288(2):318–328
16. Song Y-Y, Lu Y (2015) Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 27(2):130–135
17. Kingsford C, Salzberg SL (2008) What are decision trees? Nat Biotechnol 26(9):1011–1013
18. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19(1):281
19. Fawagreh K, Gaber MM, Elyan E (2014) Random forests: from early developments to recent advancements. Syst Sci Control Eng 2(1):602-609. https://doi.org/10.1080/21642583.2014.956265
20. Shmilovici A (2005) Support vector machines. In: Maimon O, Rokach L (eds) Data mining and knowledge discovery handbook. Springer US, Boston, pp 257–276

21. Hackenberger BK (2019) Bayes or not Bayes, is this the question? Croat Med J 60(1):50–52

22. Ali L, Khan SU, Golilarz NA, Yakubu I, Qasim I, Noor A et al (2019) A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian Naive Bayes. Comput Math Methods Med 2019:6314328

23. Pakhomov SV, Buntrock J, Chute CG (2005) Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J Biomed Inform 38(2):145–153

24. Sengupta S, Basak S, Saikia P, Paul S, Tsalavoutis V, Atiah F, Ravi V, Peters A (2020) A review of deep learning with special emphasis on architectures, applications and recent trends. Knowl-Based Syst 194:105596

25. 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(7):e0219302

26. Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115(5):654–657

27. Powers D, Ailab (2011) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol 2:2229–3981

28. Hagquist C, Stenbeck M (1998) Goodness of fit in regression analysis – R2 and G2 reconsidered. Qual Quant 32(3):229–245

29. Emmert-Streib F, Dehmer M (2019) Evaluation of regression models: model assessment, model selection and generalization error. Mach Learn Knowl Extr 1:521–551

30. Nichols JA, Herbert Chan HW, Baker MAB (2019) Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev 11(1):111–118

31. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB et al (2012) Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. Eur Heart J 34(19):1404–1413

32. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, Fonarow GC, Masoudi FA, American Heart Association Get With the Guidelines-Heart Failure Program (2010) A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. Circulation 3(1):25–32

33. Eichler K, Zoller M, Tschudi P, Steurer J (2007) Barriers to apply cardiovascular prediction rules in primary care: a postal survey. BMC Fam Pract 8(1):1

34. McMurray JJV, Packer M, Desai AS, Gong J, Lefkowitz MP, Rizkala AR et al (2014) Angiotensin–neprilysin inhibition versus enalapril in heart failure. N Engl J Med 371(11):993–1004

35. McMurray JJV, Solomon SD, Inzucchi SE, Køber L, Kosiborod MN, Martinez FA, Ponikowski P, Sabatine MS, Anand IS, Bělohlávek J, Böhm M, Chiang CE, Chopra VK, de Boer RA, Desai AS, Diez M, Drozdz J, Dukát A, Ge J, Howlett JG, Katova T, Kitakaze M, Ljungman CEA, Merkely B, Nicolau JC, O'Meara E, Petrie MC, Vinh PN, Schou M, Tereshchenko S, Verma S, Held C, DeMets D, Docherty KF, Jhund PS, Bengtsson O, Sjöstrand M, Langkilde AM, DAPA-HF Trial Committees and Investigators (2019) Dapagliflozin in patients with heart failure and reduced ejection fraction. N Engl J Med 381(21):1995–2008

36. Rekha G, Tyagi AK & Reddy VK (2019) A Wide Scale Classification of Class Imbalance Problem and its Solutions: A Systematic Literature Review. J Comput Sci 15(7):886-929. https://doi.org/10.3844/jcssp.2019.886.929