A Comparison of Methods for Disease Progression Prediction Through a GoDARTS Study

Agnes Martine Nielsen (✉ agni@gmail.com)
Danmarks Tekniske Universitet Institut for Matematik og Computer Science https://orcid.org/0000-0002-1347-3937

Rikke Linnemann Nielsen
Danmarks Tekniske Universitet

Louise Donnelly
University of Dundee

Kaixin Zhou
University of Dundee

Anders Dahl
Danmarks Tekniske Universitet Institut for Matematik og Computer Science

Ramneek Gupta
Danmarks Tekniske Universitet

Bjarne Ersbøll
Danmarks Tekniske Universitet Institut for Matematik og Computer Science

Ewan Pearson
University of Dundee

Line Clemmensen
Danmarks Tekniske Universitet Institut for Matematik og Computer Science

Research article

Keywords: machine learning, comparison study, survival model, Cox proportional hazard model, disease progression

Posted Date: March 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-16189/v1

RESEARCH

A Comparison of Methods for Disease Progression Prediction Through a GoDARTS Study

Agnes M Nielsen[1*], Rikke L Nielsen[2,3], Louise Donnelly[4], Kaixin Zhou[4], Anders B Dahl[1], Ramneek Gupta[2], Bjarne K Ersbøll[1], Ewan Pearson[4] and Line Clemmensen[1]

[*]Correspondence: agni@dtu.dk
[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Kgs. Lyngby, Denmark
Full list of author information is available at the end of the article

Abstract

Background: In recent years, a variety of new machine learning methods are being employed in prediction of disease progression, e.g. random forest or neural networks, but how do they compare to and are they direct substitutes for the more traditional statistical methods like the Cox proportional hazards model? In this paper, we compare three of the most commonly used approaches to model prediction of disease progression.

We consider a type 2 diabetes case from a cohort-based population in Tayside, UK. In this study, the time until a patient goes onto insulin treatment is of interest; in particular discriminating between slow and fast progression. This means that we are both interested in the results as a raw time-to-insulin prediction but also in a dichotomized outcome making the prediction a classification.

Methods: Three different methods for prediction are considered: A Cox proportional hazards model, random forest for survival data and a neural network on the dichotomized outcome. The performance is evaluated using survival performance measures (concordance indices and the integrated Brier score) and using the accuracy, sensitivity, specificity, and Matthews correlation coefficient for the corresponding classification problems.

Results: We found no improvement when using the conditional inference forest over the Cox model. The neural network out performed the conditional inference forest in the classification problem.

We discuss the limitations of the three approaches and where they each excel in terms of prediction performance, interpretation, and how they handle data imbalance.

Keywords: machine learning; comparison study; survival model; Cox proportional hazard model; disease progression

1 Background

In the medical literature, there are typically three approaches to modeling a time-to-event study where it is of interest to predict the time until a patient receives treatment, e.g. goes on insulin. The first approach is to use parametric or semi-parametric models of which the most common is the Cox proportional hazards model, which directly models the time-to-event for each subject [1]. Limiting ourselves to diabetes literature, as our case concerns diabetes, this approach has for example been used to predict incident diabetes in [2], the risk of diabetes in [3], incidents of cardiovascular events, diabetes and death in [4], risk of type 2 diabetes in [5], and progression to diabetes in [6].

The second approach is to use newer developments like decision trees or random forest survival models [7–9]. These model the survival time in a non-parametric way by splitting the data into smaller subsets. A survival tree has been used to analyse disease progression in type 1 diabetes in [10]. Random forests for survival data have for example been used to predict the risk of diabetes complications [11] and the risk of diabetic retinopathy [12].

The third approach is to predict a given time step ahead, e.g. 1, 2, or 5 years, by use of simpler linear models or non-linear machine learning methods. Linear regression models have for example been used to study 1-year HbA1c reduction [13] and predict asymmetric dimethylarginine levels in patients with type 2 diabetes mellitus [14]. Additionally, logistic regression has been used to predict (classify) diabetes and cardiovascular disease at the time of the study [15], drug-treated diabetes diagnosed during 5-year follow-up [16], 5-year diabetes risk after gestational diabetes mellitus in [17], and prediction of highest quartiles of asymmetric dimethylarginine levels in patients with type

Table 1: Summary of biochemical variables with longitudinal measurements.

Variable	MPP	Baseline	Total	Total w. ≥ 3 m.
Alanine transaminase	1 (0-2)	36.39 (33.60)	4261	1376
Aspartate aminotransferase	0 (0-0)	28.74 (26.94)	156	31
Cholesterol	2 (1-3)	5.36 (1.34)	5546	1785
Creatinine	2 (1-4)	79.63 (23.42)	5702	2994
Glycated haemoglobin (HbA1c)	2 (1-3)	8.37 (2.03)	5842	2113
High-density lipoprotein	2 (1-3)	1.19 (0.33)	5317	1645
Low-density lipoprotein	0 (0-1)	2.86 (1.03)	3136	159
Triglycerides	1 (0-1)	3.08 (3.14)	3693	297

The table shows the median and quantiles (first and third) or the number of measurements per person (MPP), the mean and standard deviation of the baseline measurement (Baseline), number of people with a baseline measurement (Total), and number of people with more than three measurements within the year around diagnosis (total w. ≥ 3 m.).

2 diabetes mellitus [14]. Linear regression and logistic regression models come with the additional advantage of hypothesis tests for the covariate coefficients. Non-linear machine learning methods on the other hand often give rise to better predictions due to handling non-linearities as well as co-linearities typically at the cost of missing out on hypothesis testing for individual covariates. This makes the models harder to interpret. Machine learning methods have seen an increased use in the recent years as for example in [18] where several methods including support vector machines and random forests have been used to predict diabetes complications at 3, 5, and 7 years from the first visit. They find that random forests perform the best but choose logistic regression for application in the clinic due to it being easier to interpret. Recently, there have been efforts to interpret machine learning methods [19, 20] and tests for inclusion of variables in a random forest [21, 22].

We compare these three approaches to elucidate their pros and cons for the analysis of medical time-to-event data, with a particular focus on a diabetes case. We believe this comparison is a useful basis for other medical areas with similar cohort studies. First we choose one representative method from each of these approaches and consider how well they handle data imbalance, interpretation of the model, prediction performance, and the effects of dichotomization. We then apply the methods to a subset of the Genetics of Diabetes Audit and Research (GoDARTS) data set [23]. This data set investigates the time until a type 2 diabetes patient goes onto insulin from the day of diagnosis. It consists of around 7000 patients and has both clinical and biochemical variables. The first method is Cox proportional hazards model which has previously been applied to this data set [24–26]; the second is a random forest approach for survival data; and the last is a neural network which has also been applied in this data set in [27]. We consider two versions of the data set: Using the biochemical values closest to the diagnosis and using features extracted from one year around diagnosis. This is done to investigate whether it is a good feature extraction strategy and to have two versions of the data set for the method comparisons. Finally, we give recommendations based on this study.

The rest of the paper is organised as follows. In Section 2, we present the data set as well as the methods used in the analysis. The results of the analysis of the data set are presented in Section 3, a discussion of the methods is given in Section 4, and conclusions based on the study are given in Section 5.

2 Methods

2.1 Data

We study the medical records of 6871 patients with type 2 diabetes extracted from the GoDARTS database [23–26].

The data consist of biochemical markers which were obtained during regular patient visits and have been recorded at different times and with varying frequency for each patient (Table 1), clinical variables including anthropometric, life-style and drug prescription variables (Table 2), and the time

Table 2: Summary of clinical variables.

Variable	Mean (SD) or totals	Total
BMI (kg/m^2)	31.56 (6.23)	5139
Gender		6324
Female	2803	
Male	3521	
Smoking		6324
No	1593	
Yes	4731	
Social class	3 (2-4)*	6229
Age at diagnosis (days)	22574 (3947.75)	6324
Year of diagnosis	2002 (1999-2005)*	6324
Treatment at diagnosis		6324
None or missing	4924	
Mono	1354	
Dual	46	
Weight (kg)	88.22 (19.25)	5139
Diastolic blood pressure	82.19 (11.02)	5720
Systolic blood pressure	143.3 (20.15)	5720
Glutamic acid decaboxylase (U/ml)		4860
<11	4655	
>11	205	

*Shows the median and the quantiles.

The table shows the mean and standard deviation of the variables or for categorical variables the number of observations in each category, and the total number of observations.

from diagnosis to first insulin treatment or they left the study as well as whether insulin treatment was given. The diagnosis is confirmed when HbA1c> 6.5% or first drug is received and the date of first insulin is defined ad when they first received insulin or when two measurements of HbA1c> 8.5% were taken within two months. In this data set, there is around 58% censoring meaning patients who did not receive insulin treatment while participating in the study. We only consider patients with type 2 diabetes. In order to minimize the number of type 1 diabetes patients, only patients diagnosed after 35 years of age are included. Besides this only patients diagnosed in 2010 or earlier are included. This leaves 6324 patients.

2.2 Feature Extraction

We define two data sets extracted from the biochemical markers listed in Table 1. The first is the *baseline data set* which consists of one measurement per variable j and person i. It is the measurement closest to the day of confirmed diagnosis and no more than 182.5 days before or after (Table 1).

The second data set, the *time model data set*, is created as follows. For each variable j and observation i we consider all K^{ij} measurements within 182.5 days of diagnosis and extract three features describing the time-dependent behavior. A linear trend or each observation i is estimated by solving [28].

$$\underset{a_0^{ij},a_1^{ij}}{\text{minimize}} \sum_{k=2}^{K^{ij}} (t_k^{ij} - t_{k-1}^{ij})(x_k^{ij} - a_0^{ij} - a_1^{ij} t_k^{ij})^2 \tag{1}$$

where t_k^{ij} is the k'th time point for variable j, x_k^{ij} is the k'th measurement value of variable j, and a_0^j and a_1^{ij} are the two parameters for variable j. Let $z_k^{ij} = x_k^{ij} - a_0^{ij} - a_1^{ij} t_k^{ij}$ and we fit a first order auto regressive model by solving [29].

$$\underset{\tau^{ij}}{\text{minimize}} \sum_{k=2}^{K^{ij}} \left(z_k^{ij} - \exp\left(\frac{-(t_k^{ij} - t_{k-1}^{ij})}{\tau^{ij}} \right) z_{k-1}^{ij} \right)^2 \tag{2}$$

where τ^{ij} is the auto-regressive parameter. This gives three extracted features for each biochemical variable j: $\hat{a}_0^{ij}$ describing the general level, $\hat{a}_1^{ij}$ describing a linear trend, and $\hat{\tau}^{ij}$ capturing the auto regressive aspect. These variables are extracted for all biochemical variables that have at least 3 measurements (Table 1).

2.3 Prediction

A Cox proportional hazard model [1] and a random forest model for survival (also known as the conditional inference forest) [30–33] are fitted to the survival data $\{X_i, T_i, \delta_i\}_{i=1}^n$ where X_i contains the explanatory variables, T_i is the time-to-event and δ_i is the event status, i.e. whether the event happens within the study period ($\delta_i = 1$) or not ($\delta_i = 0$), for observation i.

Besides this, a neural network is fitted to the dichotomized data $\{X_i, Y_i\}_{i=1}^n$ where $Y_i = I(T_i < t_c \wedge \delta_i = 1)$, t_c is the cut off, and I is the indicator function, i.e. Y_i indicates whether the event happens before a set time point t_c.

2.3.1 Cox Proportional Hazards Model

The Cox proportional hazards model is given by [1] as

$$\lambda(t|X_i) = \lambda_0(t)\exp(X_i\beta) \tag{3}$$

which models the hazard at time t (event rate at time t given survival until at least time t) for observation i with the explanatory variables X_i. Here β contains the model parameters and λ_0 is the unknown base hazard. From this model, it is not possible to give the predicted survival times because the base hazard is unknown. The predictions can therefore be given as either the linear predictors in the test data set ($X_{\text{new}}\hat{\beta}$) where $\hat{\beta}$ is the estimate of the β and X_{new} is the test data; or as an estimate$\hat{S}(t)$ of the survival function $S(t) = P(T > t)$ where T is a continuous random variable. The model is fitted using the *rms* package in R [34, 35].

2.3.2 Random Forest for Survival Data

The random forest model fits a number of trees to form a forest [30]. A tree is created by recursively splitting the data into smaller subsets based on some criteria (see an example in Figure 1). The starting point which contains all the data is called the root node. When the data is split it forms two child nodes. These can also be split into child nodes meaning that the subset of data is split again. This can continue until only one observation is in each child node or until a stopping criteria is reached. The final nodes are called terminal nodes. Trees are unbiased predictors but have high variance. Other advantages are that they can handle mixed variables and missing data, and that they can perform variable selection.

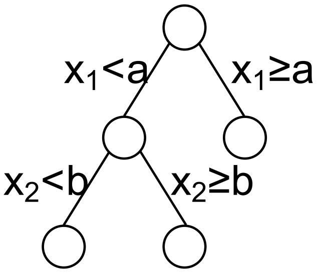

Figure 1: **Tree example.** Example of a tree where the data is first split on variable x_1 and then on x_2.

A forest is created by bootstrapping the data, and building a single tree for each bootstrap sample of the data set in a process called bagging (Algorithm 1) [36]. This is done to decorrelate the trees which lowers the variance of the predictions. The forest then inherits many of the advantages of trees but reduces the variance. In random forest by [30] the trees are further decorrelated by selecting a variable at each split in a tree from a random subset of the variables.

Algorithm 1 Pseudo-algorithm for bagging of trees.

```
1 From 1 to number of trees
    (a) Take a bootstrap sample
    (b) Build tree on sample
2 Aggregate results from all trees
```

The conditional inference forest (CIF) by [7] which builds each tree testing a global hypothesis is chosen over the random survival forest by [9] which adheres strictly to the random forest conditions laid out by [30], because it performs at least as well as the latter [37].

A conditional inference tree is built by recursively repeating two steps [8] (Algorithm 2).

Algorithm 2 Pseudo-algorithm for conditional inference tree.

```
Let each node be defined by a non-negative case weight vector $\mathbf{w} \in \mathbb{R}_+^n$ such that observations which
are elements of the node have non-negative weights and the weights are otherwise zero.
Repeat the following steps:
1 For case weights $\mathbf{w}$ test the global null hypothesis of independence between any of the variables
  and the response. If the hypothesis cannot be rejected stop and otherwise select the variable with
  the strongest association to the time to event.
2 Split the sample space of the chosen variable into two disjoint sets with each assigned to one node.
  The case weight vectors of the two new child nodes are calculated by multiplying each case weight
  with 1 if the observation is in the node's corresponding set and 0 otherwise.
```

The conditional inference forest consists of a number of conditional inference trees. The prediction can either be given as an estimate $\hat{T}_i$ of the survival time T_i or as an estimate of the survival function, $\hat{S}(t)$. The survival function $S(t)$ is estimated by a Kaplan-Meier estimate on the aggregation of terminal nodes which a new observation falls in [7]. The estimate of the survival time is calculated by the observation weighted average of the trees. The model can handle missing data by using surrogate splits meaning if a variable is missing it splits on another variable which leads to the same subsets [38]. It is fitted using the R package *party* [39].

2.3.3 Neural Network

A single-hidden-layer neural network (NN) is used for the dichotomized response [40]. This model consists of three layers of artificial neurons (also known as units or nodes), input, hidden, and output (Figure 2). Each layer consists of a number of neurons that receives input from all neurons in the previous layer and sends the output to all neurons in the next layer. For binary classification the output layer only has one neuron. Neurons in the hidden and output layers process their inputs by multiplying by a weight, summing them, adding a constant and then taking a fixed activation function. The result is,

$$\hat{y} = \phi_0 \left(\alpha + \sum_h w_h \phi_h \left(\alpha_h + \sum_i w_{ih} X_i \right) \right) \tag{4}$$

where ϕ_0 is the output layer activation function, ϕ_h is the hidden layer activation function, w_h and w_{ih} are weights, α and α_h are constants, h runs over the hidden units, and i runs over the observations. This function is fitted for example using the logistic cost function as a loss function (in the two class case)

$$-\sum_{i=1}^{n} (Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i) + \lambda \sum_{ih} w_{ih}^2 \tag{5}$$

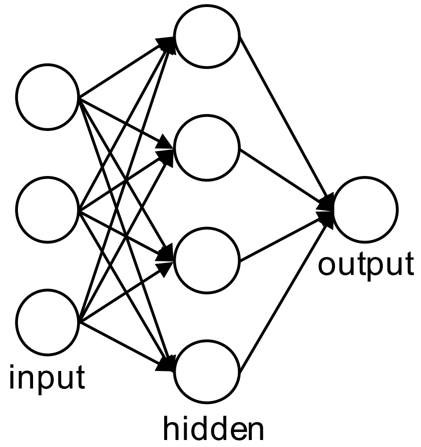

Figure 2: **Example of a neural network.**

where $p_i = \exp(\hat{y}_i)/(\exp(\hat{y}_i) + \exp(1 - \hat{y}_i))$ and λ is a decay parameter. Y_i was defined previously as whether the event happened before a set time point.

We optimise the number of neurons in the hidden layer and the decay parameter in an inner cross-validation loop according to the area under the receiver operating curve [41]. Further, for this model all continuous variables are standardised by mean and standard deviation within the cross-validation. This is fitted using the R package *nnet* [42].

2.4 Missing Data Strategies

The Cox proportional hazards model cannot handle missing data. In the baseline data set, only the biochemical variables cholesterol, creatinine, HbA1c, and high-density lipoprotein are included with the clinical variables as the others have more than 50% missing values and all biochemical variables except high-density lipoprotein are log-transformed due to skewness. Similarly, in the time model data, we only include extracted features with less than 50% missing values. These are the extracted features for cholesterol, creatinine, HbA1c, and high-density lipoprotein.

We have then removed all observations with missing values leaving only 1063 observations. This is done for simplicity. It is generally not advisable as it drastically reduces the number of observations and in cases where the missing data are not missing completely at random, the results are generally biased [43, 44]. However, the alternatives of imputing the missing data using single imputation often causes us to be overly certain about the results and more advanced multiple imputation methods are generally better but also have their own problems [45].

The neural network model used here cannot handle missing data either. In this case, we use the imputation strategy of replacing missing values with "-999". This strategy can be used if there is a belief that the values are not missing at random because it makes the missing values very different from any observed values. However, this method also has its problems, e.g. that it changes the distribution of the data.

As previously mentioned, the random forest model can handle missing data using surrogate splits and is therefore fitted on the data as is. However, we note that the biochemical variables cholesterol, creatinine and HbA1c are log-transformed in the baseline data set as they were for the Cox model.

2.5 Data Imbalance

Class imbalance is a well studied area in classification [46]. A variety of sampling methods can be used to address the imbalance, e.g. downsampling, upsampling, and synthetic oversampling techniques [47, 48].

We have used downsampling as it is the simplest method. It is used both to balance the classes created by dichotomization (denoted as *Y1*, *Y3* and *Y5* if downsampled based on the dichotomization over year 1, 3 or 5, respectively) but also the censoring (denoted *cens.*), i.e. downsampling the censored observations such that we have an equal number of censored and uncensored observations.

This is done because recent methods for dealing with the imbalance by resampling shows that it improved the performance [49]. If no downsampling is done it is denoted *none*.

2.6 Model Evaluation

2.6.1 Cross-Validation

The performance is evaluated in a 5 fold cross validation which has been repeated twice [50]. Cross-validation is stratified according to either the censoring or the dichotomized class on which the downsampling has also been done. For the neural network, stratification on the class is also done for the non-downsampled results. All downsampling is done on only the training set and within the cross-validation.

2.6.2 Performance Measures

The performance measures for the survival models have been chosen because they are either routinely reported or account for censoring while they can be calculated for models other than the Cox model [51].

The discriminative powers of the survival models are evaluated using two different concordance indices (C-indices). The C-index by [52] which is calculated as the number of concordant pairs of observations, i.e. pairs where the observation which has the lowest survival time is also predicted to be lowest while the event happened for that observation, over the number of comparable pairs, i.e. pairs where the event happened for the observation with the lowest survival time.

$$C_H = \frac{\sum_{i=1}^n \sum_{j=1}^n I(T_i > T_j) I(\eta_j > \eta_i)\delta_j}{\sum_{i=1}^n \sum_{j=1}^n I(T_i > T_j)\delta_j} \tag{6}$$

where $\eta_i \in \mathbb{R}$ is a one dimensional score computed for each observation, i.e. the predicted survival time or the linear predictors. T_i is the time-to-event δ_i is the event status. The C-index by [53] is given by

$$C_U = \frac{\sum_{i=1}^n \sum_{j=1}^n I(T_i > T_j) I(\eta_j > \eta_i)\delta_j \hat{G}(T_j)^{-2}}{\sum_{i=1}^n \sum_{j=1}^n I(T_i > T_j)\delta_j \hat{G}(T_j)^{-2}} \tag{7}$$

where $\hat{G(\cdot)}$ is the Kaplan-Meier estimator of the censoring distribution. This C-index takes censoring into account. Both of the measures lie between 0 and 1 where 0.5 corresponds to random guessing and 1 is the best.

Besides this we report the integrated Brier score [54],

$$\text{IBS} = \int_0^{\max(t)} \frac{1}{n} \sum_{i=1}^n (I(T_i > t) - \hat{S}(t))^2 dt. \tag{8}$$

for which 0 represents a perfect fit and smaller values are better.

The classification performance is evaluated by the accuracy,

$$\text{accuracy} = \frac{\sum_{i=1}^n I(\hat{Y}_i = Y_i)}{n} \tag{9}$$

which measures the proportion of correctly classified observations. It is also evaluated by the sensitivity which measures the proportion of positives correctly identified,

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{10}$$

where $\text{TP} = \sum_{i=1}^n I(\hat{Y}_i = 1 \wedge Y_i = 1)$ are the true positives and $\text{FN} = \sum_{i=1}^n I(\hat{Y}_i = 0 \wedge Y_i = 1)$ are the false negatives. The performance is evaluated by specificity which measures the proportion of negatives correctly identified,

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{11}$$

Table 3: Performance of survival methods.

Method	Data	Sampling	C_H	C_U	IBS
Cox	B	None	0.69 (0.02)	0.64 (0.06)	0.18 (0.01)
Cox	T	None	0.71 (0.02)	0.65 (0.04)	0.18 (0.01)
CIF	B	None	0.68 (0.01)	0.53 (0.10)	0.19 (0.01)
CIF	T	None	0.68 (0.01)	0.54 (0.07)	0.19 (0.01)
Cox	B	cens.	0.69 (0.03)	0.64 (0.05)	0.18 (0.01)
Cox	T	cens.	0.71 (0.04)	0.66 (0.03)	0.19 (0.02)
CIF	B	cens.	0.69 (0.01)	0.58 (0.07)	0.19 (0.01)
CIF	T	cens.	0.69 (0.01)	0.61 (0.05)	0.19 (0.01)
CIF	B	Y1	0.66 (0.01)	0.60 (0.04)	0.26 (0.01)
CIF	T	Y1	0.66 (0.01)	0.60 (0.04)	0.26 (0.01)
CIF	B	Y3	0.68 (0.01)	0.60 (0.04)	0.24 (0.01)
CIF	T	Y3	0.68 (0.02)	0.64 (0.05)	0.25 (0.01)
CIF	B	Y5	0.69 (0.01)	0.64 (0.02)	0.22 (0.01)
CIF	T	Y5	0.69 (0.01)	0.65 (0.03)	0.22 (0.01)

Mean and standard deviation. Two methods are compared Cox proportional hazards model (Cox) and the conditional inference forest (CIF) on two data sets baseline (B) and time model (T) (see Section 2.2). The sampling refers to whether the data was downsampled on censoring (cens.), year 1 (Y1), 3 (Y3) or 5 (Y5), or not at all (None) (see Section 2.5).

where $\mathrm{TN} = \sum_{i=1}^{n} I(\hat{Y}_i = 0 \wedge Y_i = 0)$ are the true negatives and $\mathrm{FP} = \sum_{i=1}^{n} I(\hat{Y}_i = 1 \wedge Y_i = 0)$ are the false positives. Finally, the performance is evaluated by Matthews correlation coefficient (MCC) [55] because it is a balanced measure that can be used even for highly imbalanced data,

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{12}$$

which gives a value between -1 and 1 where 0 means not better than random and 1 means perfect prediction.

3 Results

The performance of the survival models are improved by including the extracted time model features in terms of C-indices and IBS (Table 3). There is no improvement by using the random forest model even though more data is utilized and it can model interactions.

Downsampling based on censoring gave a small improvement in C_U which takes censoring into account but not C_H and IBS. Downsampling on censoring did not affect the classification performance but this is due to the models already just predicting the majority class (Table 4). When downsampling based on one of the classes (year 1, 3, or 5), the performance in terms of C_H is largely unchanged and in terms of IBS worse, possibly due to the reduced number of observations. In terms of C_U it however improves compared to not having resampled. The classification accuracy of the survival models decreases when the majority class (high survival time) is downsampled. This is because it no longer just predicts a high survival time which can be seen by the specificity increasing. MCC also increases when the majority class is downsampled.

The performance of the neural network is also improved by downsampling. The accuracy increases and the sensitivity and specificity become more balanced. MCC does however only increase for year 5 due to high specificity for the other years before downsampling. For all three years, the neural network performs better than the conditional inference forest.

4 Discussion

The Cox proportional hazards model is the most commonly used model for survival data [56]. The model is easily interpretable due to its simplicity and because it allows for hypothesis testing

Table 4: Classification results.

Method	Data	Sampling	Class	Accuracy	Sensitivity	Specificity	MCC
CIF	B	None	1	96.2 (0.43)	0.00 (0.00)	100 (0.00)	0.00 (0.00)
CIF	T	None	1	96.2 (0.43)	0.00 (0.00)	100 (0.00)	0.00 (0.00)
NN	B	None	1	95.7 (1.2)	50.2 (20.4)	99.1 (0.3)	0.60 (0.17)
CIF	B	cens.	1	96.2 (0.55)	0.00 (0.00)	100 (0.00)	0.00 (0.00)
CIF	T	cens.	1	96.2 (0.55)	0.00 (0.00)	100 (0.00)	0.00 (0.00)
CIF	B	Y1	1	67.5 (2.7)	72.7 (4.7)	67.3 (2.8)	0.16 (0.02)
CIF	T	Y1	1	70.2 (2.2)	70.5 (6.2)	70.2 (2.4)	0.17 (0.02)
NN	B	Y1	1	81.1 (5.6)	77.0 (17.4)	81.4 (5.4)	0.36 (0.13)
CIF	B	None	3	89.3 (0.99)	0.00 (0.00)	100 (0.00)	0.00 (0.00)
CIF	T	None	3	89.3 (0.99)	0.00 (0.00)	100 (0.00)	0.00 (0.00)
NN	B	None	3	83.6 (3.9)	37.8 (21.3)	97.0 (1.5)	0.42 (0.26)
CIF	B	cens.	3	89.3 (1.1)	0.08 (0.26)	100 (0.00)	0.01 (0.03)
CIF	T	cens.	3	89.3 (1.1)	0.00 (0.00)	100 (0.00)	0.00 (0.00)
CIF	B	Y3	3	66.1 (2.3)	73.0 (3.4)	65.3 (2.7)	0.24 (0.02)
CIF	T	Y3	3	65.8 (2.3)	72.7 (3.8)	65.0 (2.7)	0.24 (0.03)
NN	B	Y3	3	68.4 (10.7)	62.5 (5.8)	70.2 (13.8)	0.30 (0.14)
CIF	B	None	5	80.6 (1.4)	3.74 (1.7)	99.4 (0.47)	0.11 (0.03)
CIF	T	None	5	80.5 (1.5)	1.15 (0.79)	99.9 (0.13)	0.07 (0.04)
NN	B	None	5	68.6 (7.6)	57.4 (7.8)	78.5 (8.5)	0.37 (0.16)
CIF	B	cens.	5	81.0 (1.4)	8.29 (1.8)	98.8 (0.05)	0.18 (0.04)
CIF	T	cens.	5	80.8 (1.5)	3.14 (1.19)	99.8 (0.16)	0.13 (0.04)
CIF	B	Y5	5	65.8 (1.6)	71.7 (3.8)	64.4 (2.4)	0.29 (0.02)
CIF	T	Y5	5	64.9 (1.6)	72.6 (4.2)	63.1 (2.5)	0.27 (0.03)
NN	B	Y5	5	71.0 (2.9)	67.3 (1.9)	74.3 (6.5)	0.42 (0.06)

Classification results in percent (except for MCC which is on a scale from -1 to 1). Mean and standard deviations. Two methods are compared the conditional inference forest (CIF) and a neural network (NN) on two data sets baseline (B) and time model (T) (see Section 2.2). The sampling refers to whether the data was downsampled on censoring (cens.), year 1 (Y1), 3 (Y3) or 5 (Y5), or not at all (None). Class is the year on which the dichotomized response is cut-off.

for relevance of included variables [1]. However, due to being a proportional hazards model with unknown base hazard it cannot give predictions of the actual time-to-event. It further has the issue that it cannot handle missing data.

The latter two issues are addressed by the conditional inference forest model. It handles missing data by using surrogate splits [38]. Because the methods have different strategies for missing values, the performance is compared on different subsets of the data. This means that the results are not controlled such that difference in the results are due to one difference in the approach. However, this is the reality of clinical data so a comparison of methods has to consider missing data differently for different methods. The conditional inference forest can further natively give predictions of survival times. These advantages come at the expense of a more complicated model. It is not easily interpretable, though an experimental variable importance measure is available for the conditional inference forest [39] and a variable importance measure is available for the random survival forest [57]. In our data set the Cox model performs slightly better than the conditional inference forest. The reason for this could be that the assumptions of the Cox model might be satisfied or that the flexibility of the random forest might not needed.

That the conditional inference forest model can give a prediction of the time-to-event which gives the possibility to compare its performance to methods for the dichotomized response. Generally, dichotomizing is not advisable [58]. However, one might have non-statistical reasons that this form of the response is of interest, e.g. easier translation of knowledge to clinicians. If one chooses this approach, then survival models are no longer a natural choice of model as the problem becomes a binary classification. However, one advantage of using a conditional inference forest for survival data is that it can be used to classify for any cut off.

We consider a neural network for classification. This implementation cannot handle missing data. But the main limitation of this approach is that it cannot utilize the full information of the time to insulin because this has been encoded as a binary class label. It does have the advantage of optimizing the model to classification and not the survival time.

If the dichotomization is done such that one class is much larger than the other then there is a need for resampling in order to make the classes even. When downsampling is chosen we lose some data. We therefore lose more information on top of the information lost due to dichotomization. However, dealing with the imbalance of the data is necessary to produce generalizable results. In our data, the neural network outperforms the conditional inference forest for the survival data in all three years when the resampling is done. This means if the dichotomized response is of interest we are better off fitting a model directly to this response. If there is no resampling, both methods perform poorly. The same downsampling is required in order to get comparable performance from dichotomizing the survival time predictions from the conditional inference forest meaning that using the model for any cut off did produce desirable results (Table 4).

5 Conclusions

In this final section, we give our recommendations and conclusions based on this study.

Imbalance Based on this study, downsampling on censoring in survival prediction has a small but positive effect and might be relevant to investigate. Downsampling in classification or otherwise dealing with the imbalance is however important to produce generalizable results.

Interpretation The Cox model performed well in our case. This model is preferable if the interpretation of the model parameters is of interest.

Prediction We found that specifically trained models performed the best for that setting, resulting in the conditional inference forest being outperformed in both settings.

Dichotomization If interested in the dichotomized response then it is preferable to build a model on this response directly rather than dichotomizing post training.

Abbreviations

GoDARTS: Genetics of Diabetes Audit and Research; CIF: Conditional inference forest; NN: Neural network; C-index: concordance index; IBS: Integrated Brier Score; MCC: Matthews correlation coefficient.

Ethics approval and consent to participate

The GoDARTS study was approved by the Tayside Committee on Medical Research Ethics, and informed consent was obtained from all patients (REC reference 053/04).

Consent for publication

Data are not publicly available, but may be available on request.

Availability of data and materials

Ewan Pearson is the data guarantor.

Competing interests

The authors declare that they have no competing interests.

Funding

RLN was supported by the Sino-Danish Center for Education and Research. RLN and AMN were supported by the Poul V Andersen Foundation. ERP holds a Wellcome Trust New Investigator award (102820/Z/13/Z).

Authors' contributions

AMN, LC, RLN, RG contributed to the analysis. AMN, RLN, LD and KZ prepared the data set. First draft of manuscript by AMN and LC. Input into writing the manuscript by AMN, LC, RLN, RG, ABD, RG, BKE and EP. All authors read and reviewed the manuscript.

Acknowledgements

The authors are grateful to all the participants who took part in this study; the general practitioners; the Scottish School of Primary Care for its help in recruiting the participants; and the whole team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. Part of the material has been presented at the Pacific symposium of Biocomputing 2019 (PSB19) where the abstract was published in presentation of the workshop "Text Mining and Machine Learning for Precision Medicine"; Pacific Symposium on Biocomputing 24:449-454(2019), URL: `http://psb.stanford.edu/psb-online/proceedings/psb19/wkshop-text.pdf`. We thank Peter Wad for helping prepare the dataset.

Author details

[1]Department of Applied Mathematics and Computer Science, Technical University of Denmark, Richard Petersens Plads, Kgs. Lyngby, Denmark. [2]Department of Health Technology, Technical University of Denmark, Øtsteds Plads, Kgs. Lyngby, Denmark. [3]Sino-Danish Center for Education and Research, University of Chinese Academy of Sciences, 380 Huaibeizhuang, Beijing, China. [4]School of Medicine, University of Dundee, Nethergate, Dundee, Scotland, UK.

References

1. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187–220. Available from: http://www.jstor.org/stable/2985181.
2. Sattar N, Scherbakova O, Ford I, O'Reilly DSJ, Stanley A, Forrest E, et al. Elevated alanine aminotransferase predicts new-onset type 2 diabetes independently of classical risk factors, metabolic syndrome, and C-reactive protein in the west of Scotland coronary prevention study. Diabetes. 2004;53(11):2855–2860.
3. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. Bmj. 2009;338:b880.
4. Liljestrand J, Havulinna A, Paju S, Männistö S, Salomaa V, Pussinen P. Missing teeth predict incident cardiovascular events, diabetes, and death. Journal of dental research. 2015;94(8):1055–1062.
5. Hippisley-Cox J, Coupland C. Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. bmj. 2017;359:j5019.
6. Steck AK, Dong F, Frohnert BI, Waugh K, Hoffman M, Norris JM, et al. Predicting progression to diabetes in islet autoantibody positive children. Journal of autoimmunity. 2018;90:59–63.
7. Hothorn T, Lausen B, Benner A, Radespiel-Tröger M. Bagging survival trees. Statistics in medicine. 2004;23(1):77–91.
8. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics. 2006;15(3):651–674.
9. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS, et al. Random survival forests. The annals of applied statistics. 2008;2(3):841–860.
10. Xu P, Krischer JP, Type 1 Diabetes TrialNet Study Group. Prognostic classification factors associated with development of multiple autoantibodies, dysglycemia, and type 1 diabetes - a recursive partitioning analysis. Diabetes Care. 2016;p. dc152292.
11. Lagani V, Chiarugi F, Thomson S, Fursse J, Lakasing E, Jones RW, et al. Development and validation of risk assessment models for diabetes-related complications based on the DCCT/EDIC data. Journal of Diabetes and its Complications. 2015;29(4):479–487.
12. Semeraro F, Parrinello G, Cancarini A, Pasquini L, Zarra E, Cimino A, et al. Predicting the risk of diabetic retinopathy in type 2 diabetic patients. Journal of Diabetes and its Complications. 2011;25(5):292–297.
13. Farmer AJ, Rodgers LR, Lonergan M, Shields B, Weedon MN, Donnelly L, et al. Adherence to Oral Glucose–Lowering Therapies and Associations With 1-Year HbA1c: A Retrospective Cohort Analysis in a Large Primary Care Database. Diabetes care. 2015;p. dc151194.
14. Ganz T, Wainstein J, Gilad S, Limor R, Boaz M, Stern N. Serum asymmetric dimethylarginine and arginine levels predict microvascular and macrovascular complications in type 2 diabetes mellitus. Diabetes/metabolism research and reviews. 2017;33(2):e2836.
15. Janiszewski PM, Janssen I, Ross R. Does waist circumference predict diabetes and cardiovascular disease beyond commonly evaluated cardiometabolic risk factors? Diabetes care. 2007;30(12):3105–3109.
16. Lindström J, Tuomilehto J. The diabetes risk score: a practical tool to predict type 2 diabetes risk. Diabetes care. 2003;26(3):725–731.
17. Claesson R, Ignell C, Shaat N, Berntorp K. HbA1c as a predictor of diabetes after gestational diabetes mellitus. Primary care diabetes. 2017;11(1):46–51.
18. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine learning methods to predict diabetes complications. Journal of diabetes science and technology. 2018;12(2):295–302.
19. Welling SH, Refsgaard HH, Brockhoff PB, Clemmensen LH. Forest floor visualizations of random forests. arXiv preprint arXiv:160509196. 2016;.
20. Ribeiro MT, Singh S, Guestrin C. Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016. p. 1135–1144.
21. Mentch L, Hooker G. Formal Hypothesis Tests for Additive Structure in Random Forests. Journal of Computational and Graphical Statistics. 2017;26(3):589–597.
22. Wager S, Athey S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association. 2018;0(0):1–15.
23. Hebert HL, Shepherd B, Milburn K, Veluchamy A, Meng W, Carr F, et al. Cohort Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). International Journal of Epidemiology. 2017 09;47(2):380–381j. Available from: https://doi.org/10.1093/ije/dyx140.
24. Doney AS, Fischer B, Leese G, Morris AD, Palmer CN. Cardiovascular risk in type 2 diabetes is associated with variation at the PPARG locus: a Go-DARTS study. Arteriosclerosis, thrombosis, and vascular biology. 2004;24(12):2403–2407.
25. Doney AS, Lee S, Leese GP, Morris AD, Palmer CN. Increased Cardiovascular Morbidity and Mortality in Type 2 Diabetes Is Associated With the Glutathione S Transferase Theta–Null Genotype: A Go-DARTS Study. Circulation. 2005;111(22):2927–2934.
26. Zhou K, Donnelly LA, Morris AD, Franks PW, Jennison C, Palmer CN, et al. Clinical and genetic determinants of progression of type 2 diabetes: a DIRECT study. Diabetes care. 2014;37(3):718–724.
27. Nielsen RL, Donnelly L, Nielsen AM, Zhou K, Dawed A, Tsirigos K, et al. Prediction of time to insulin requirement in patients with type 2 diabetes using artificial intelligence: A GoDARTS study. In preparation. 2020;.
28. Eckner A. A note on trend and seasonality estimation for unevenly-spaced time series. unpublished; 2012.
29. Mudelsee M. Climate time series analysis. Springer; 2014.
30. Breiman L. Random forests. Machine learning. 2001;45(1):5–32.
31. Hothorn T, Buehlmann P, Dudoit S, Molinaro A, Van Der Laan M. Survival Ensembles. Biostatistics. 2006;7(3):355–373.
32. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics. 2007;8(25). Available from: http://www.biomedcentral.com/1471-2105/8/25.
33. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional Variable Importance for Random Forests. BMC Bioinformatics. 2008;9(307). Available from: http://www.biomedcentral.com/1471-2105/9/307.
34. Harrell Jr FE. rms: Regression Modeling Strategies; 2018. R package version 5.1-2. Available from: https://CRAN.R-project.org/package=rms.
35. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2017. Available from: https://www.R-project.org/.
36. Breiman L. Bagging predictors. Machine learning. 1996;24(2):123–140.
37. Nasejje JB, Mwambi H, Dheda K, Lesosky M. A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. BMC medical research methodology. 2017;17(1):115.
38. Hothorn T, Hornik K, Strobl C, Zeileis A. Party: A laboratory for recursive partytioning; 2010.

39. Hothorn T, Hornik K, Strobl C, Zeileis A. party: A Laboratory for Recursive Partytioning; 2018. R package version 1.3-1.
40. Venables WN, Ripley BD. Modern Applied Statistics with S. 4th ed. New York: Springer; 2002. ISBN 0-387-95457-0. Available from: http://www.stats.ox.ac.uk/pub/MASS4.
41. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1):29–36.
42. Ripley BD, Venables WN. nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models; 2016. R package version 7.3-12. Available from: http://www.stats.ox.ac.uk/pub/MASS4/.
43. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592.
44. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values. Journal of clinical epidemiology. 2006;59(10):1087–1091.
45. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009;338:b2393.
46. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent data analysis. 2002;6(5):429–449.
47. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321–357.
48. Menardi G, Torelli N. Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery. 2014;28(1):92–122.
49. Afrin K, Illangovan G, Srivatsa SS, Bukkapatnam ST. Balanced Random Survival Forests for Extremely Unbalanced, Right Censored Data. arXiv preprint arXiv:180309177. 2018;.
50. Kohavi R, et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai. vol. 14. Montreal, Canada; 1995. p. 1137–1145.
51. Rahman MS, Ambler G, Choodari-Oskooei B, Omar RZ. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC medical research methodology. 2017;17(1):60.
52. Harrell Jr FE, Califf RM, Pryor DB, Lee KL, Rosati RA, et al. Evaluating the yield of medical tests. Jama. 1982;247(18):2543–2546.
53. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei L. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in medicine. 2011;30(10):1105–1117.
54. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Statistics in medicine. 1999;18(17-18):2529–2545.
55. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1975;405(2):442–451.
56. Harrell FE. Cox proportional hazards regression model. In: Regression modeling strategies. Springer; 2015. p. 475–519.
57. Ishwaran H, et al. Variable importance in binary regression trees and forests. Electronic Journal of Statistics. 2007;1:519–537.
58. Fedorov V, Mannino F, Zhang R. Consequences of dichotomization. Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry. 2009;8(1):50–61.

Figures

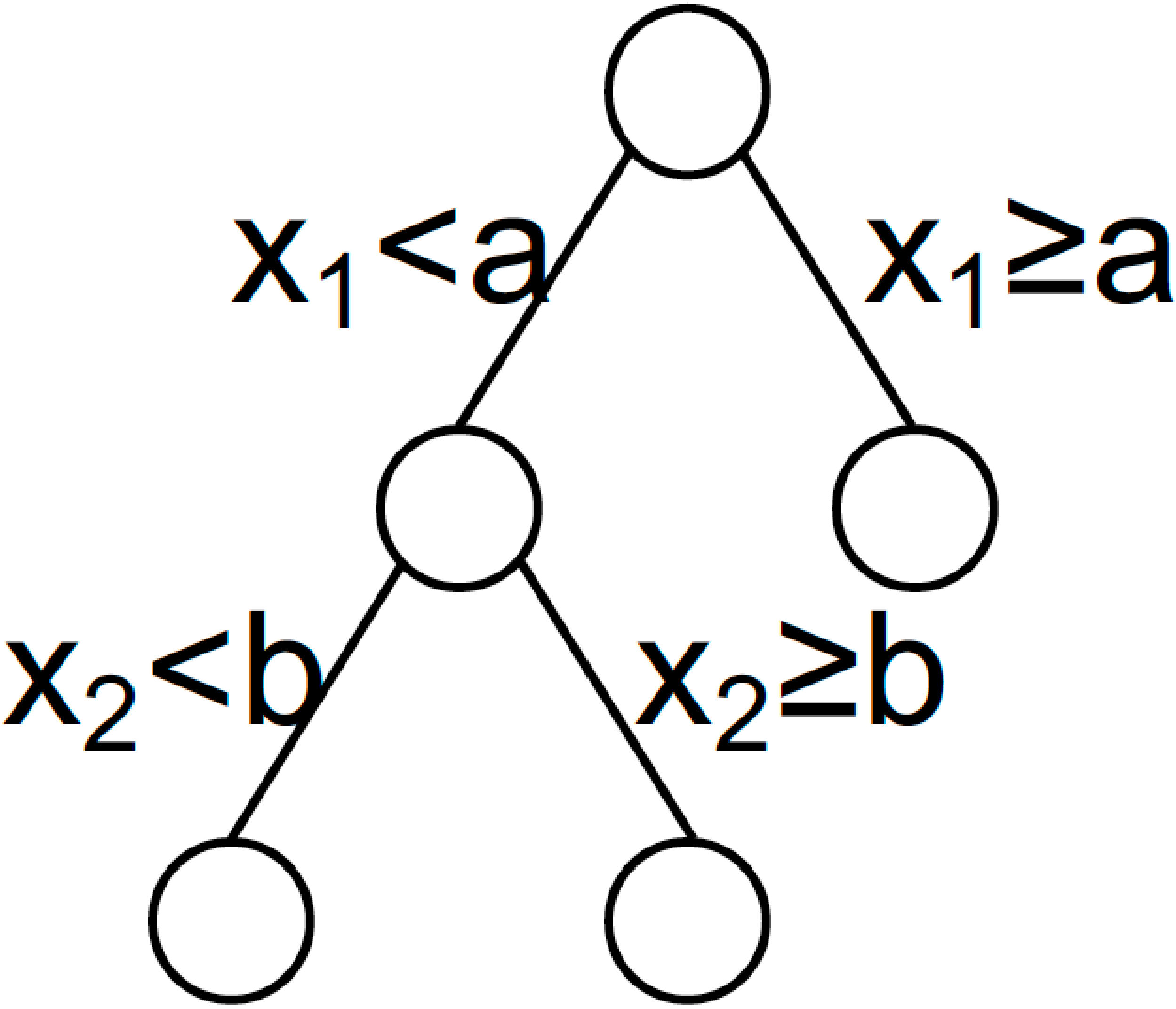

Figure 1

Tree example. Example of a tree where the data is first split on variable x1 and then on x2.

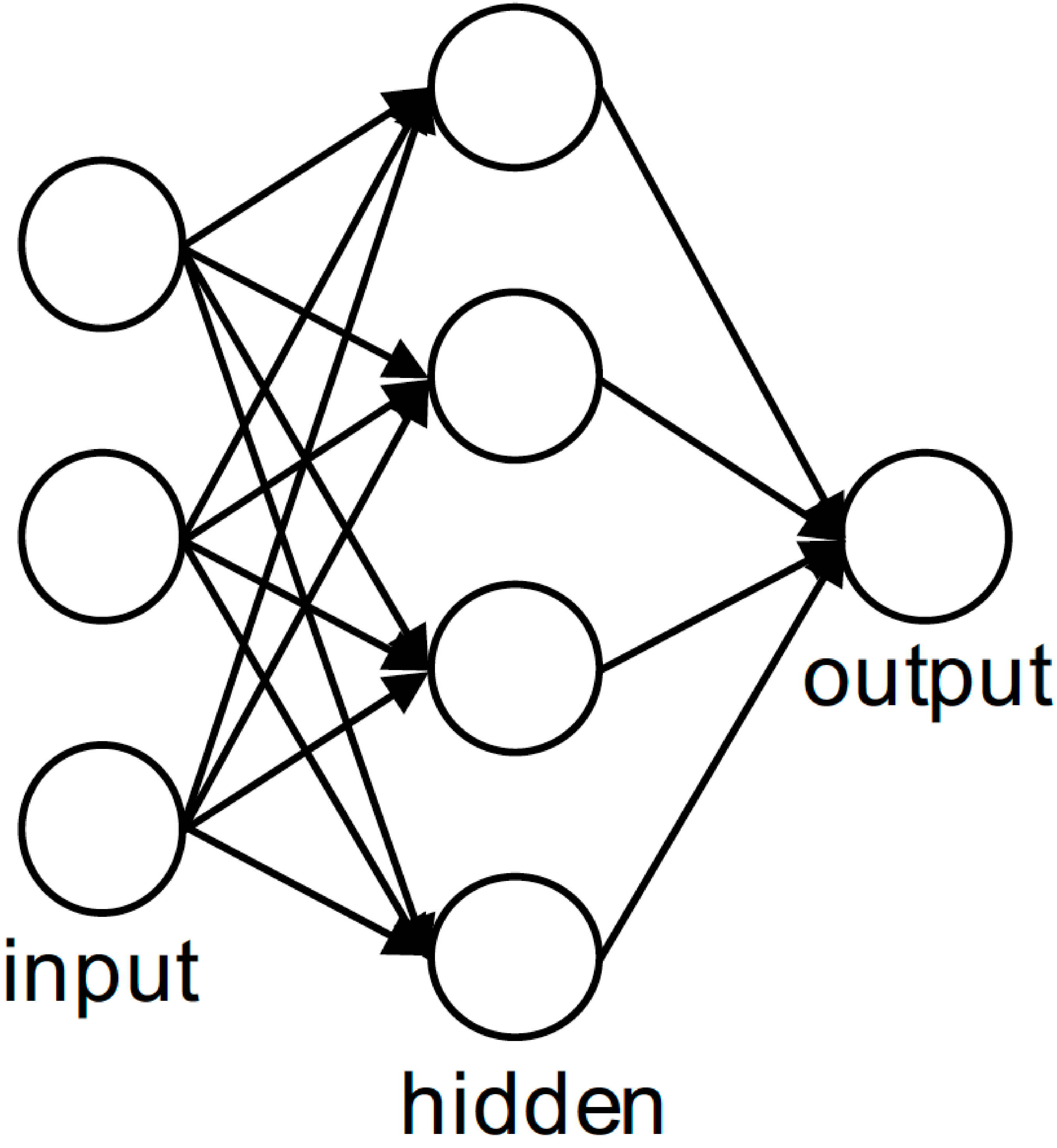

Figure 2

Figure 2: Example of a neural network.