Artificial Intelligence Review (2024) 57:84
https://doi.org/10.1007/s10462-024-10718-1

A lexicographic optimisation approach to promote more recent features on longitudinal decision-tree-based classifiers: applications to the English Longitudinal Study of Ageing

Caio Ribeiro[1] · **Alex A. Freitas**[1]

Accepted: 31 January 2024

Abstract
Supervised machine learning algorithms rarely cope directly with the temporal information inherent to longitudinal datasets, which have multiple measurements of the same feature across several time points and are often generated by large health studies. In this paper we report on experiments which adapt the feature-selection function of decision tree-based classifiers to consider the temporal information in longitudinal datasets, using a lexicographic optimisation approach. This approach gives higher priority to the usual objective of maximising the information gain ratio, and it favours the selection of features more recently measured as a lower priority objective. Hence, when selecting between features with equivalent information gain ratio, priority is given to more recent measurements of biomedical features in our datasets. To evaluate the proposed approach, we performed experiments with 20 longitudinal datasets created from a human ageing study. The results of these experiments show that, in addition to an improvement in predictive accuracy for random forests, the changed feature-selection function promotes models based on more recent information that is more directly related to the subject's current biomedical situation and, thus, intuitively more interpretable and actionable.

Keywords Classification · Longitudinal data · Age-related diseases

1 Introduction

Longitudinal datasets are a special case of temporal datasets (i.e., datasets that store time-related variations of feature values), where the same set of instances (e.g., patients) is followed through a number of points in time, denominated waves. Several countries have been running longitudinal populational studies of ageing, where they collect data on various

✉ Caio Ribeiro
C.E.Ribeiro@kent.ac.uk

Alex A. Freitas
A.A.Freitas@kent.ac.uk

[1] School of Computing, University of Kent, Canterbury, UK

Published online: 09 March 2024

aspects of the lives of older individuals, including physical and mental health, demographics, and socioeconomic aspects. Longitudinal datasets from human ageing studies typically span several years, with longer intervals of time between waves, and measure a large number of features (Kaiser 2013; Ribeiro et al. 2017).

Analysing longitudinal data may offer insights, for example, on cause and effect patterns, on how an event affects a feature's values, or how a pattern evolves with time. Due to the high number of predictive features (independent variables) in these studies, machine learning (ML) applications are often needed for performing holistic analyses—i.e., considering hundreds or thousands of features simultaneously.

Note that longitudinal data should not be confused with time series data (Bagnall et al. 2017), even though both have a temporal nature. In the context of supervised ML, time series data usually contain a single real-valued variable repeatedly measured across a large number of time points; whilst our target longitudinal datasets consist of a large number of both numerical and nominal variables repeatedly measured across a small number of time points.

Supervised ML techniques use training data to create a model able to make predictions about previously unseen data. Because standard supervised ML algorithms do not cope directly with the temporality of longitudinal data, they disregard time-related information that may be relevant to the problem.

One way to address this issue is to adapt existing ML algorithms to make them cope directly with the temporal information in longitudinal data, so that they can use time-related information (different measurements of a feature across several time points) to try to improve predictive performance (Niemann et al. 2015; Ribeiro and Freitas 2019). This is broadly the research direction followed in this article, which is in general an under-explored research area—since relatively few existing supervised ML methods directly cope with longitudinal data.

Thus, in this article we report the results of experiments with an algorithm adaptation to decision tree-based classification algorithms that uses the time-related information of longitudinal data to increase predictive accuracy, first described in Ribeiro and Freitas (2020). More precisely, this current article extends our recent previous work in two directions, as follows.

First, in this article we report results for two well-known decision tree-based algorithms: Random Forest (RF) (Breiman 2001) and the decision tree algorithm J48 (an implementation of the well-known C4.5 decision tree in the Weka tool) (Quinlan 1993), whilst in Ribeiro and Freitas (2020) only the Random Forest (RF) algorithm was used. We also perform more experiments with the algorithms' parameter optimisation in this article, optimising up to two parameters for each algorithm, whilst in Ribeiro and Freitas (2020) the experiments optimised only one parameter of the RF algorithm. Second, in this article we report results for 20 longitudinal classification datasets, whilst only 10 datasets where used in the experiments reported in Ribeiro and Freitas (2020).

Importantly, we created datasets from two different data sources (with two different types of predictive features) that differ on the time distance between waves (time points) and number of feature waves. By contrast, the experiments in Ribeiro and Freitas (2020) used datasets from a single data source, where all datasets had the same number of waves. The use of two classification algorithms and two data sources in this article led to further insight about the effectiveness of the proposed adaptation for decision tree-based classifiers.

The datasets were created for binary classification problems, each consisting of predicting whether participants in a longitudinal study will develop an age-related disease at the

last wave (time point) of the study, based on self-reported and biomedical health data collected over several years.

The proposed algorithm adaptation is a lexicographic bi-objective split-feature selection procedure that considers both the information gain ratio and the time-index of the candidate features when selecting the split feature of each node in the process for learning a decision tree. In essence, the lexicographic approach gives priority to selecting features with a higher information gain ratio, but when some candidate features have approximately the same highest gain ratio, the most recent feature among those is selected.

Another contribution of this article is an analysis of the top-ranked features in our best RF models. This analysis includes a discussion of the health-related features that were most relevant in our predictive models, and whether they were previously associated with the target age-related disease in the literature.

This article is organised as follows. Section 2 briefly discusses related work. Section 3 discusses the dataset creation process and the experimental setup. The algorithm adaptation of changing the split-feature selection function of decision tree-based classifiers is defined in Sect. 4, with experimental results for this approach presented in Sect. 5. In Sect. 6, we interpret the top-ranked features in the best RF classifiers. In Sect. 7 we discuss our conclusions and future work.

2 Related work

There are several possible approaches for considering temporal patterns in supervised machine learning problems (Ribeiro and Freitas 2019), such as creating temporal features in a preprocessing step (Ribeiro and Freitas 2021a), Structural Pattern Detection (Morid et al. 2020), Recurrent Neural Networks (often Long-Short Term Memory) (Aghili et al. 2018), and Deep Learning (Luo et al. 2020). In this section we focus on decision tree-based classifiers, which are popular in biomedical applications and are the focus of our proposed algorithm adaptation for longitudinal classification.

More precisely, the algorithm adaptation approach in this work consists of adapting the split-feature selection function of decision tree-based classification algorithms, by adding a secondary objective to be considered. This approach of optimising objectives in priority order is sometimes called the lexicographic approach (Freitas 2004), and it has been used in decision tree algorithms for conventional (non-longitudinal) classification before (Basgalupp et al. 2009). A similar strategy of using time-related information in the split decision was used in Deng et al. (2013), where the authors combine entropy gain and a time-related distance measure in their split criteria, for an application in time series datasets.

The lexicographic split approach is a simple adaptation that does not add significant processing to the training process and can be applied to any decision tree-based classifier, promoting the generation of classifiers that use more recent data, which is more readily available, actionable and intuitively explainable, compared to older measurements of the same features. For example, if the class variable to be predicted is the occurrence of a heart attack, intuitively a recent measurement of a patient's cholesterol level is more relevant for that prediction than a much older measurement. It is worthwhile to mention that the our approach does not reduce the interpretability of the models learned from the data, nor does it significantly increase the execution times of these algorithms, both of these being important characteristics of decision tree-based classifiers.

3 Datasets and experimental set up

3.1 The created ELSA-nurse and ELSA-core datasets

The English Longitudinal Study of Ageing (ELSA) is currently one of the most prominent populational studies of ageing (Abell et al. 2018; Banks et al. 2019). The ELSA has, in each of its waves, thousands of respondents from inhabitants of United Kingdom households, which take part in a core interview every two years (the time interval between two consecutive waves), answering questions about various aspects of their lives, including demographic, health, wellbeing and economics. Data from this core questionnaire is used to create the class labels for all our datasets, and to create the ELSA-core datasets. For this project, we used data from the core waves 1–7 (2002–2014) to create the features of the ELSA-core datasets, and data from the core wave 8 to create the class label for all our datasets.

In addition, special questionnaires are used to collect biomedical data every 2 waves (i.e., roughly every 4 years), when a professional nurse visits the respondents in their home and performs a face-to-face interview and a series of tests. The results of these nurse visits are recorded in separate files, which we used to create our ELSA-nurse datasets. We used data from four waves of the ELSA study with data collected by a nurse: waves 2, 4, 6 and 8 (2004–2016).

A total of 20 longitudinal datasets were created with the raw data files from the ELSA-core and ELSA-nurse questionnaires, each with a combination of one of two data sources (core data or data collected by a nurse) and one of 10 age-related diseases used as class (target) variables. The class variable in each dataset refers to the presence (negative class) or absence (positive class) of a diagnose for an age-related disease, for each instance (ELSA respondent), in wave 8. For all 10 diseases, the positive class (disease absence) is the majority class, with an increased degree of class imbalance for rarer diseases, such as Dementia and Parkinson's Disease. Note that, in order to have class labels for all instances (ELSA respondents), we only utilised data from respondents that participated in the ELSA's 8th wave. In cases where a respondent did not participate in any of the other waves in the dataset, the values for that wave's features were set as missing for that respondent.

The 10 ELSA-nurse datasets share the same set of predictive features, as do the 10 ELSA-core datasets, even though they have different class variables (representing different age-related diseases), as explained in more detail later. The ELSA class labels represent diagnosis for Angina, Arthritis, Cataract, Dementia, Diabetes, High blood pressure, Heart attack, Osteoporosis, Parkinson's Disease, and Stroke. Note that we have not used any class labels from previous waves (before wave 8) as predictive features, meaning the models have no information of earlier diagnosis, as this would change the prediction problem (our aim is to make predictions based solely on biomedical variables and self-reported health data, not based on previous diagnosis results).

Most predictive features in our datasets have multiple measures, one for each wave. For instance, cholesterol is a feature measured in multiple waves, taking a value for each wave. As an exception, some demographic features (such as sex) take just one value, which is set with the most recent feature wave as time-index by default.

It is important to highlight that the ELSA participants themselves are reporting the diagnosis of the target diseases in their core interviews, and there is no clinical data available corroborating their answers. Thus, even though we take the data available as

ground-truth, it is likely that some patients were undiagnosed or did not report their diagnosis, and that some patients wrongly reported their positive diagnosis.

Table 1 shows the names of the class variables in terms of age-related diseases, the names of the original ELSA variables used to create the class variables in this work, and the class imbalance ratios for the ELSA-nurse and ELSA-core class variables. The class imbalance ratio is calculated by dividing the number of majority class instances by the number of minority class instances.

In addition to selecting the features in the questionnaires that were relevant to our classification problem, and performing standard data cleaning tasks such as coding missing values and merging similar variables to reduce the dimensionality of the dataset, we also imputed all missing values in the dataset in the data preprocessing stage. For this, we utilised a data-driven missing value replacement approach proposed in Ribeiro and Freitas (2021b), which completely replaces the existing missing values in a longitudinal dataset, using various imputation methods, including some designed specifically for longitudinal data. Importantly, this missing value imputation was performed using the training set only, without using the test set. After the data preparation, the final ELSA-nurse and ELSA-core datasets had 141 and 171 predictive features, and 7097 and 8405 instances (ELSA participants), respectively.

3.2 Experimental setup

For our experiments, we created classification models using the J48 Decision Tree (Quinlan 1993) and Random Forest (RF) (Breiman 2001) algorithms. The algorithm adaptation discussed in this article can be implemented in any decision tree-based classification algorithm; here we use one decision tree method (an implementation of the widely used C4.5) and one ensemble method (Random Forests) as well-known representatives of two decision tree-based approaches for classification. Note that RFs handle well datasets with a high ratio of features to instances, which are prone to overfitting (Scornet et al. 2015). This is desirable as longitudinal datasets often have a large number of features, as variables are measured at multiple time points and each measurement becomes a predictive feature.

To cope with the class imbalance in our datasets, decision tree training sets were undersampled to a ratio of 1:1 (randomly removing majority class instances), and RF training sets were balanced using the Balanced Random Forest (BRF) method (Chen et al. 2004). BRF applies a majority class undersampling for each bootstrap sample taken at each tree of the forest, so the subset of instances used to generate each decision tree has a balanced ratio (1:1) of instances of the two classes. The 1:1 ratio is a default approach adopted by several studies (López et al. 2013; Weiss and Provost 2003). Intuitively, this ratio encourages the algorithms to try to predict well both classes, preventing the algorithms from overwhelmingly predicting the majority class. Note that, for both algorithms, majority-class undersampling is applied to the training set only; i.e. the test remains with the original imbalanced class distribution, which best reflects the true class distribution in the real-world.

The decision trees and RFs were trained and tested using the Weka toolkit[1] (Eiben et al. 2016). We performed two types of experiments. In the first type, we used the default parameter settings of the decision tree algorithm and the default settings for the "standard parameters" of the decision tree and random forest algorithms. However, the

[1] Version 3.9, available at: https://www.cs.waikato.ac.nz/ml/weka/

Table 1 ELSA-nurse and ELSA-core class variables and their class imbalance ratios

Class variable	ELSA-core variables used to create the class label	ELSA-nurse Class Imbalance Ratio	ELSA-core Class Imbalance Ratio
Heart Attack	Hedacmi—Whether confirms heart attack diagnosis Hediami—Heart attack diagnosis newly reported	16.70	19.06
Angina	Hedacan—Whether confirms angina diagnosis Hedasan—Whether still has angina Hediaan—Angina diagnosis newly reported Hediman—Angina diagnosis newly reported (merged)	26.51	29.49
Stroke	Hedacst—Whether confirms stroke diagnosis Hediast—Stroke diagnosis newly reported	15.86	18.35
Diabetes	Hedacdi—Whether confirms diabetes or high blood sugar diagnosis Heacd—Whether ever been told has diabetes by doctor Hediadi—Diabetes or high blood sugar diagnosis newly reported	6.50	7.80
High Blood Pressure	Hedacbp—Whether confirms high blood pressure diagnosis Hedasbp—Whether still has high blood pressure Hediabp—High blood pressure diagnosis newly reported Hedimbp—High blood pressure diagnosis newly reported (merged)	1.49	2.58
Dementia	Hedbdde—Whether confirms dementia diagnosis Hedbsde—Whether still has dementia Hedibde—Dementia diagnosis newly reported	56.96	52.20
Cataract	Heopcca—Whether confirms cataract diagnosis Heopsca—Whether still has cataract Heoptca—Cataract diagnosis newly reported	2.06	3.38
Arthritis	Hedbdar—Whether confirms arthritis diagnosis Hedbsar—Whether still has arthritis Hedibar—Arthritis diagnosis newly reported	1.35	2.52
Osteoporosis	Hedbdos—Whether confirms osteoporosis diagnosis Hedbsos—Whether still has osteoporosis Hedibos—Osteoporosis diagnosis newly reported	9.85	11.84

Table 1 (continued)

Class variable	ELSA-core variables used to create the class label	ELSA-nurse Class Imbalance Ratio	ELSA-core Class Imbalance Ratio
Parkinsons	Hedbdpd—Whether confirms Parkinsons Disease diagnosis Hedbspd—Whether still has Parkinsons Disease Hedibpd—Parkinsons Disease diagnosis newly reported	160.30	112.07

lexicographic versions of these two algorithms have a tie-threshold parameter, defined later in Sect. 4.2. So, in this first type of experiment, we optimise only this non-standard tie-threshold parameter, for the lexicographic decision tree and random forest algorithms, as described in Sect. 4.2. Since the standard (non-lexicographic) decision tree and random forest algorithms do not have this tie-threshold parameter, in this first type of experiment these standard algorithms do not have any parameter optimised.

In the second type of experiment, both the lexicographic and non-lexicographic versions of each algorithm (decision tree and random forest) have two parameters optimised in controlled experiments, where each version of an algorithm is given the same computational budget for parameter optimisation. To fix this computational budget, each version of an algorithm optimises two parameters considering in total 10 combinations of candidate values for those parameters (i.e. 10 algorithm configurations). The 10 combinations of candidate parameter values are evaluated using a well-known grid search, via an internal cross-validation on the training set, i.e., the performance on that internal cross-validation is used to select the parameter values with the best performance on the current training set. Then, the chosen algorithm configuration is applied to the full training set to learn a model that is evaluated on the test set. This process is repeated for all training-test set pairs in the external cross-validation (i.e. this process performs a "nested cross-validation" procedure). Note that test data is not used for optimising parameters; test data is used only for measuring generalisation performance, as usual.

The default parameter settings on Weka for the chosen classifiers are: for decision trees, $C = 0.25$ (confidence factor used for pruning the trees), and $M = 2$ (minimum number of instances that can constitute a leaf node); for RFs $ntrees = 100$ (number of trees), $minleafsamples = 1$ (minimum number of instances in a leaf node) and $mtry = \lfloor log_2(d) \rfloor + 1$ (number of features randomly sampled to be used as candidate features at each tree node), where the total number of features is d, and $\lfloor x \rfloor$ is the "floor" of x, i.e., the biggest integer which is smaller than or equal to x. In this second type of experiment, the optimised parameters and their candidate values for each version of an algorithm were as follows:

- Standard (non-lexicographic) decision tree: candidate C values: [0.1, 0.15, 0.2, 0.25, 0.3], candidate M values: [2, 5] (total computational budget: 10 models);
- Lexicographic decision tree: candidate C values: [0.15, 0.25], candidate tie-threshold values: [0.01, 0.02, 0.03, 0.04, 0.05] (total computational budget: 10 models);
- Standard (non-lexicographic) random forest: candidate mtry values: $\lfloor log_2(d) \rfloor + 1$, $2 * (\lfloor log_2(d) \rfloor + 1)$, $\left\lfloor 0.5 * \sqrt{d} \right\rfloor$, $\left\lfloor \sqrt{d} \right\rfloor$, $\left\lfloor 2 * \sqrt{d} \right\rfloor$], candidate minleafsamples values: [1, 3] (total computational budget: 10 models);
- Lexicographic random forest: candidate mtry values: $[\lfloor log_2(d) \rfloor + 1, \left\lfloor \sqrt{d} \right\rfloor]$, candidate tie-threshold values: [0.01, 0.02, 0.03, 0.04, 0.05] (total computational budget: 10 models);

The classifiers were evaluated using 4 metrics. These measures are formally defined as follows, in terms of the numbers of True Positives (TP), False Positives (FP), False Negatives (FN):

- Sensitivity (or Recall): a local metric of the true positive rate (given by Eq. 1, where # denotes "the number of"). For problems where false negatives are the least desir-

able outcome, such as clinical diagnosis applications, the ML algorithm needs to maximise mainly Sensitivity.
- Specificity: a local metric that represents the true negative rate (given by Eq. 2). It is a complementary measure to Sensitivity.
- Accuracy: the fraction of correct predictions made by the model over all predictions (given by Eq. 3). This is a widely used global performance metric, however in highly imbalanced datasets the majority class has a much bigger impact on Accuracy, which can mask bad results for the minority class in a model.
- GMean: The geometric mean between Sensitivity and Specificity (given by Eq. 4). This is another global performance metric, but it gives the exact same weight to both classes regardless of the class distribution in the data.

$$Sensitivity = \frac{\#TP}{\#TP + \#FN} \tag{1}$$

$$Specificity = \frac{\#TN}{\#FP + \#TN} \tag{2}$$

$$Accuracy = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN} \tag{3}$$

$$GMean = \sqrt{Sensitivity * Specificity} \tag{4}$$

These metrics were chosen partially based on Malley et al. (2011, Chap. 4), which claim that, for imbalanced biomedical data, models should be evaluated using "local" metrics that consider their ability to predict each class separately (like Sensitivity and Specificity) and at least one "global" metric of performance over both classes. The Accuracy is a widely used measure of global importance, and we chose to add the GMean as a second global measure because it assigns equal importance to the prediction of both minority-class and majority-class instances, whilst the Accuracy assigns greater importance to the majority class.

The experiments used the well-known 10-fold cross-validation procedure, and report the average of these four performance metrics over the 10 test sets of the cross-validation.

4 A lexicographic bi-objective function for selecting node-split features in decision tree-based classifiers

The lexicographic split feature-selection adaptation for decision tree-based classifiers, originally introduced in Ribeiro and Freitas (2020), consists of considering not only the features' information gain ratios but also their time points (wave ids) when choosing the split feature inside a decision tree's node, making the decision bi-objective. In this article we report the results comparing the standard and modified versions of Random Forests and J48 decision tree classifiers. Although this lexicographic approach has been described in Ribeiro and Freitas (2020), we also describe it here in order to make this current article self-contained.

4.1 The rationale for a bi-objective split function

The modified split function works as follows. When choosing the feature to be used in a node's split, the decision trees in the adapted algorithms will consider maximising the gain ratio as the primary objective and maximising the time-index of the features (ELSA wave numbers) as the secondary objective. The rationale for this bi-objective feature evaluation is that this adds a desirable bias favouring more recent information. This is based on the heuristic that more recent values of biomedical features tend to be more useful for predicting future occurrences of diseases than older values of the same features.

We believe this to be the case for longitudinal biomedical datasets in general. Intuitively, the further in the past a feature value was measured, the less it is related to the class label. However, we always prioritise gain ratio over the time-index, it remains the most important criterion for improving predictive accuracy; whilst preferring more recent feature values as a tie-breaking criterion is a heuristic for improving accuracy.

One limitation of the proposed heuristic with a bias favouring more recent features is that, as any other heuristic for feature selection, it is not guaranteed to lead to better results for all datasets. Note that, for some diseases such as dementia, early indications in biomedical variables might be helpful for prediction (Javeed et al. 2023), so removing older data outright would not be advisable. Therefore, we keep all available data in the dataset, still giving earlier data a chance to be selected by the split function (based on their information gain ratio). The rationale for the bi-objective split is simply choosing more recent data in cases where candidate features have equivalent information gain ratios.

Another argument for using the proposed lexicographic feature-selection criterion is that it leads to classification models that are less dependent on older data. This is desirable because longitudinal datasets created for classification problems, especially in the ageing studies used as data sources in our experiments, tend to have more missing values in the earlier waves. As many instances are added to the study as it has new waves added, and instances from participants who left the study will not be present in the target wave (so they don't have a class value, and must be discarded for the classification datasets), the tendency is that the closer to the target wave, the less likely a feature is to have a missing value due to attrition.

4.2 Description of the lexicographic split function

In the standard J48 decision tree algorithm, the split-feature selection considers every feature of the dataset in each node of the tree, ordering them based on their Information gain ratio $g(f_{i,j})$ (feature i measured at time j) for that node, selecting the feature with the greater gain value for splitting the data.

For the standard split-feature selection used by Random Trees in the RF algorithm, instead of using all available features, the algorithm first randomly samples a set of candidate features S from the dataset ($|S| = mtry$, with $mtry$ being a user-defined parameter for how many features are sampled). Then, it selects, among the features in S, the one with the greatest information gain ratio.

For the lexicographic split-feature selection approach, we define a threshold th as an additional parameter, and consider two features equivalent when the difference between their gain ratios is lower than this threshold. All eligible features (i.e., all features in J48 decision trees and the randomised pool of features sampled for the current node in Random

Trees of RFs) that were considered equivalent to the initial best feature are compared based on their time-indexes (wave id), and the most recent feature is selected. This process is described in Algorithm 1 (Ribeiro and Freitas 2020). Note that, although we are considering the gain ratio function $g(f_{i,j})$ as the primary metric for selecting the split feature, it could be replaced by other metrics such as the information gain.

Algorithm 1 The lexicographic split-feature selection function, applied at eachnode of a decision tree. It receives a set of eligible features *S* and a user-specifiedtie-threshold *th*, and returns the selected *splitfeature*, based on gain ratio and thefeature's time-index

```
1: function LexicographicSplitFeatureSelection(S, th)
2:     S.DescendingOrder(gainratio)
3:     splitfeature ← S[0]
4:     CandidateFeatures.add(splitfeature)
5:     pos ← 1
6:     while |g(splitfeature) − g(S[pos])| < th AND pos < S.length do
7:         CandidateFeatures.add(S[pos])
8:         pos + +
9:     end while
10:    CandidateFeatures.DescendingOrder(time-index)
11:    splitfeature ← CandidateFeatures[0]
12:    return splitfeature
13: end function
```

The disadvantage of the lexicographic approach is the additional parameter to be selected by the user, the tie-definition threshold *th*. As an alternative to a user-defined parameter, we propose automating the choice of tie-threshold value by performing an internal 5-fold cross-validation using only the training set instances. In the experiments where only this tie-threshold parameter is optimised (i.e. the other decision tree/random forest parameters are left with their default values), this internal cross-validation creates classifiers using 11 possible threshold values (from 0.0 to 0.05, with 0.005 increments), and chooses the value that yields the model with the best average Geometric Mean of Sensitivity and Specificity in the internal cross-validation (accessing the training set only). We use these 11 candidate tie-threshold values as they worked well in Ribeiro and Freitas (2020). Note that, in another type of experiment used in this current work, only 5 (rather than 11) candidate tie-threshold values are considered because the algorithm has to spend part of its "parameter optimisation budget" optimising another parameter, as described in Sect. 3.2.

Note also that a *th* value of 0.0 does not mean that the lexicographic feature-selection criterion would not be applied (i.e., two features would never be considered tied). The information gain ratio values of different candidate features are often very close, with differences small enough that a subtraction operation in Java (the programming language used in our code) would return a 0 value. For nodes in lower depths of a decision tree, where the number of instances in the dataset is very low, exact ties happen often and are detected even with a 0.0 tie-threshold.

4.3 An example of the lexographic feature selection process

As an example of how the lexicographic feature-split approach works, consider a set *S* (the set of eligible features to be selected on a given node split) consisting of a feature

$f_{1,1}$ (feature 1 measured at time point 1) with a gain ratio of $g(f_{1,1}) = 0.7$, and a feature $f_{2,2}$ with a gain ratio of $g(f_{2,2}) = 0.67$. In the standard decision tree algorithm, $f_{1,1}$ would be selected for the split as it has the greater gain value. In the lexicographic approach, that depends on the value of th. If $th = 0.05$, we have $|g(f_{1,1}) - g(f_{2,2})| < th$, so the features' gain ratios are considered equivalent and $f_{2,2}$ is selected instead, because it was measured at time point 2 instead of 1 (giving it a greater time-index value). However, if $th = 0.01$, we have $|g(f_{1,1}) - g(f_{2,2})| > th$, so the features' gain ratios are not considered equivalent, and the selection proceeds normally, selecting $f_{1,1}$ based on its higher gain ratio. In case of a tie for both the gain ratio value and the time-index criterion, a random selection is performed (the algorithm's default tie break).

5 Experimental results

In this Section we report on experiments comparing using our proposed lexicographic bi-objective split-feature selection approach (named Lexic in the result Tables) to using the baseline classification algorithms, with no changes (named NoLexic in the result Tables). As mentioned, we performed experiments using both Random Forests (Sects. 5.1 and 5.2) and J48 decision tree (Sects. 5.3 and 5.4) classifiers, both used in two scenarios: (a) using the classifiers with their default parameter settings in general, optimising only the tie-threshold of the Lexic classifier as an exception; and (b) optimising two parameters of the Lexic and NoLexic classifiers with a grid search via an internal cross-validation on the training set. These experiments aim to investigate the impact of our proposal on ensemble classifiers and single decision-tree classifiers. For details of which parameters are optimised for each version (Lexic and NoLexic) of each algorithm (decision tree and random forest), see Sects. 3.2 and 4.2. All result tables show, for each performance metric, the average and the standard error over the 10-fold cross-validation for each dataset, then the average ranks obtained by each algorithm for each type of data source (ELSA-nurse and ELSA-core) and overall (for the 20 datasets) in the last 3 rows, with the best (smallest) average rank in boldface.

5.1 Random forest results optimising only lexicographic threshold

The results of our experiments comparing the standard (non-lexicograpnic) and the lexicographic Random Forests with default parameter settings (except that the lexicographic Random Forest optimises the tie-threshold via internal cross-validation) are presented in Tables 2 and 3, respectively.

These results show an overall trend of the Lexic approach learning models that have better or equivalent predictive accuracy to the NoLexic approach. Considering the ELSA-core datasets, with 7 feature waves (that are 2 years apart from each other), Lexic wins for all 4 performance metrics with substantial differences in the average ranks. The average rank difference is particularly large for the Geometric Mean of Sensitivity and Specificity (GMean), where the lexicographic and non-lexicographic versions of Random Forest achieved the average ranks of 1.1 and 1.9, respectively.

Regarding the ELSA-nurse datasets, with 4 feature waves (that are 4 years apart from each other, instead of 2 years apart), the results are slightly positive – Lexic wins for Sensitivity and Accuracy, but loses for Specificity and ties for GMean.

Table 2 Sensitivity and Specificity results for Lexic and NoLexic random forests, optimising only the parameter tie-threshold for Lexic random forest

Datasets	Sensitivity		Specificity	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.693 ± 0.035	0.684 ± 0.028	0.69 ± 0.007	0.702 ± 0.008
EN_Arthritis	0.669 ± 0.009	0.671 ± 0.007	0.589 ± 0.07	0.586 ± 0.008
EN_Cataract	0.63 ± 0.012	0.62 ± 0.013	0.72 ± 0.008	0.723 ± 0.008
EN_Dementia	0.74 ± 0.035	0.729 ± 0.027	0.723 ± 0.005	0.709 ± 0.006
EN_Diabetes	0.843 ± 0.007	0.841 ± 0.009	0.863 ± 0.006	0.866 ± 0.008
EN_HBP	0.647 ± 0.008	0.651 ± 0.012	0.749 ± 0.005	0.749 ± 0.006
EN_HeartAttack	0.703 ± 0.021	0.7 ± 0.017	0.741 ± 0.009	0.738 ± 0.009
EN_Osteoporosis	0.654 ± 0.016	0.649 ± 0.017	0.716 ± 0.007	0.696 ± 0.006
EN_Parkinsons	0.634 ± 0.005	0.628 ± 0.075	0.636 ± 0.006	0.712 ± 0.007
EN_Stroke	0.677 ± 0.024	0.67 ± 0.023	0.713 ± 0.007	0.724 ± 0.008
EC_Angina	0.709 ± 0.02	0.711 ± 0.026	0.737 ± 0.003	0.723 ± 0.004
EC_Arthritis	0.752 ± 0.006	0.749 ± 0.005	0.719 ± 0.008	0.717 ± 0.008
EC_Cataract	0.625 ± 0.009	0.609 ± 0.01	0.715 ± 0.007	0.717 ± 0.004
EC_Dementia	0.768 ± 0.022	0.764 ± 0.032	0.776 ± 0.005	0.77 ± 0.003
EC_Diabetes	0.672 ± 0.007	0.674 ± 0.009	0.764 ± 0.004	0.747 ± 0.005
EC_HBP	0.633 ± 0.01	0.641 ± 0.007	0.665 ± 0.005	0.662 ± 0.007
EC_HeartAttack	0.683 ± 0.033	0.678 ± 0.03	0.689 ± 0.004	0.692 ± 0.004
EC_Osteoporosis	0.701 ± 0.018	0.7 ± 0.017	0.677 ± 0.004	0.676 ± 0.003
EC_Parkinsons	0.701 ± 0.072	0.697 ± 0.062	0.733 ± 0.005	0.693 ± 0.004
EC_Stroke	0.697 ± 0.002	0.694 ± 0.019	0.729 ± 0.005	0.721 ± 0.005
AvgRank ELSA-nurse	**1.20**	1.80	1.55	**1.45**
AvgRank ELSA-core	**1.30**	1.70	**1.20**	1.80
AvgRank Overall	**1.25**	1.75	**1.38**	1.63

We compared the ranks of the Lexic and NoLexic approaches for each performance metric using the Wilcoxon signed-rank test (Wilcoxon 1992). For this analysis, we ran the Wilcoxon test over the results for all 20 datasets, i.e. considering the ranks of the Lexic and NoLexic versions of RF for each data set separately. In this overall results analysis, none of the p-values were significant. When considering only the results for the 10 ELSA-nurse datasets, Lexic was significantly better than NoLexic for Sensitivity (p-value: 0.0248) and Accuracy (p-value: 0.0170). In the 10 ELSA-core datasets there was a significant difference in Specificity (p-value: 0.0364) and GMean (p-value: 0.0142), both in favour of Lexic.

We also measured the effect the proposed lexicographic feature-selection approach had on the resulting Random Forest models, i.e., how different the models generated with the Lexic approach were from the the NoLexic models. For this, we counted in each RF model the proportion of nodes where a tie happened (nodes where more than one candidate feature had equivalent information gain ratios, according to the tie-threshold parameter) and the proportion of nodes where a replacement happened (nodes where the tie led to a different, more recent feature being selected by the Lexic approach).

In the ELSA-nurse models we had an average of 50.3% of nodes where a tie occurred (at least one candidate feature had an equivalent information gain ratio to the first-ranked

Table 3 Accuracy and Geometric Mean of Sensitivity and Specificity (GMean) for Lexic and NoLexic random forests: optimising only the tie-threshold parameter for Lexic random forests

Datasets	Accuracy		GMean	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.693 ± 0.007	0.684 ± 0.008	0.6920.019	0.693 ± 0.015
EN_Arthritis	0.635 ± 0.004	0.635 ± 0.003	0.628 ± 0.004	0.627 ± 0.003
EN_Cataract	0.660 ± 0.005	0.654 ± 0.004	0.674 ± 0.005	0.67 ± 0.005
EN_Dementia	0.740 ± 0.005	0.729 ± 0.005	0.732 ± 0.017	0.719 ± 0.011
EN_Diabetes	0.845 ± 0.005	0.845 ± 0.005	0.853 ± 0.006	0.855 ± 0.007
EN_HBP	0.688 ± 0.004	0.690 ± 0.005	0.696 ± 0.004	0.698 ± 0.005
EN_HeartAttack	0.705 ± 0.008	0.702 ± 0.009	0.722 ± 0.011	0.719 ± 0.011
EN_Osteoporosis	0.660 ± 0.007	0.654 ± 0.005	0.684 ± 0.008	0.672 ± 0.008
EN_Parkinsons	0.634 ± 0.005	0.629 ± 0.006	0.635 ± 0.021	0.669 ± 0.039
EN_Stroke	0.679 ± 0.007	0.674 ± 0.008	0.695 ± 0.014	0.697 ± 0.013
EC_Angina	0.710 ± 0.003	0.711 ± 0.004	0.723 ± 0.009	0.717 ± 0.013
EC_Arthritis	0.739 ± 0.005	0.736 ± 0.004	0.735 ± 0.005	0.733 ± 0.004
EC_Cataract	0.651 ± 0.005	0.641 ± 0.005	0.668 ± 0.005	0.661 ± 0.006
EC_Dementia	0.768 ± 0.005	0.764 ± 0.003	0.772 ± 0.011	0.767 ± 0.015
EC_Diabetes	0.684 ± 0.004	0.684 ± 0.005	0.717 ± 0.005	0.71 ± 0.05
EC_HBP	0.645 ± 0.006	0.649 ± 0.006	0.649 ± 0.007	0.651 ± 0.006
EC_HeartAttack	0.684 ± 0.003	0.679 ± 0.004	0.686 ± 0.016	0.685 ± 0.015
EC_Osteoporosis	0.699 ± 0.004	0.698 ± 0.003	0.689 ± 0.009	0.668 ± 0.009
EC_Parkinsons	0.702 ± 0.005	0.697 ± 0.004	0.717 ± 0.038	0.695 ± 0.032
EC_Stroke	0.699 ± 0.005	0.695 ± 0.005	0.713 ± 0.011	0.707 ± 0.01
AvgRank ELSA-nurse	**1.20**	1.80	**1.50**	**1.50**
AvgRank ELSA-core	**1.25**	1.75	**1.10**	1.90
AvgRank Overall	**1.23**	1.78	**1.30**	1.70

feature), making them eligible for changing the split feature based on the secondary objective, the time-index. About half of these nodes switched the chosen feature for a more recent feature, resulting in final models that were 26.6% different from the NoLexic models (learned using a standard feature-selection function). In the ELSA-core datasets we had 42.9% average nodes with ties, leading to a 23.1% average difference in the models.

5.2 Random forest results optimising two parameters

The results of our experiments comparing the standard (non-lexicographic) and lexicographic Random Forests, each with two parameters optimised via internal cross-validation, are presented in Tables 4 and 5. Recall that the parameters optimised by the NoLexic Random Forest were mtry and minleafsamples, whilst the parameters optimised by Lexic Random Forest were mtry and the tie-threshold (see Subsection 3.2), with both Random Forest versions using the same computational budget for parameter optimisation.

Overall the Lexic Random Forest outperformed the NoLexic Random Forest in these experiments. Notably, although both methods had similar Sensitivity results (average ranks of 1.45 for Lexic and 1.55 for NoLexic), most models (17/20) had higher

Table 4 Sensitivity and Specificity results for Lexic and NoLexic random forests: optimising two parameters for both Lexic and NoLexic random forests

Datasets	Sensitivity		Specificity	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.706 ± 0.033	0.699 ± 0.038	0.683 ± 0.006	0.681 ± 0.008
EN_Arthritis	0.598 ± 0.007	0.593 ± 0.007	0.672 ± 0.009	0.662 ± 0.009
EN_Cataract	0.72 ± 0.012	0.722 ± 0.012	0.624 ± 0.009	0.617 ± 0.007
EN_Dementia	0.733 ± 0.037	0.737 ± 0.04	0.735 ± 0.005	0.734 ± 0.007
EN_Diabetes	0.858 ± 0.009	0.859 ± 0.013	0.847 ± 0.006	0.844 ± 0.006
EN_HBP	0.751 ± 0.007	0.759 ± 0.01	0.648 ± 0.006	0.644 ± 0.008
EN_HeartAttack	0.718 ± 0.021	0.724 ± 0.014	0.698 ± 0.007	0.692 ± 0.008
EN_Osteoporosis	0.707 ± 0.015	0.71 ± 0.017	0.652 ± 0.007	0.649 ± 0.006
EN_Parkinsons	0.71 ± 0.081	0.68 ± 0.072	0.631 ± 0.009	0.616 ± 0.011
EN_Stroke	0.691 ± 0.026	0.684 ± 0.021	0.673 ± 0.007	0.653 ± 0.018
EC_Angina	0.738 ± 0.034	0.736 ± 0.025	0.713 ± 0.004	0.708 ± 0.004
EC_Arthritis	0.712 ± 0.006	0.718 ± 0.006	0.748 ± 0.007	0.747 ± 0.005
EC_Cataract	0.728 ± 0.008	0.718 ± 0.009	0.625 ± 0.004	0.621 ± 0.004
EC_Dementia	0.827 ± 0.032	0.772 ± 0.034	0.767 ± 0.004	0.761 ± 0.005
EC_Diabetes	0.754 ± 0.007	0.746 ± 0.008	0.672 ± 0.005	0.673 ± 0.004
EC_HBP	0.661 ± 0.007	0.677 ± 0.007	0.638 ± 0.004	0.638 ± 0.005
EC_HeartAttack	0.699 ± 0.036	0.683 ± 0.029	0.683 ± 0.003	0.685 ± 0.003
EC_Osteoporosis	0.696 ± 0.022	0.666 ± 0.017	0.699 ± 0.005	0.697 ± 0.003
EC_Parkinsons	0.637 ± 0.081	0.713 ± 0.055	0.689 ± 0.005	0.685 ± 0.007
EC_Stroke	0.72 ± 0.02	0.693 ± 0.021	0.689 ± 0.006	0.695 ± 0.006
AvgRank ELSA-nurse	1.6	**1.4**	**1**	2
AvgRank ELSA-core	**1.3**	1.7	**1.3**	1.7
AvgRank Overall	**1.45**	1.55	**1.15**	1.85

Specificity values in the Lexic models. Regarding the global performance metrics, Lexic performed better in most cases, only tying for the average Accuracy rank in ELSA-core datasets. The Wilcoxon signed-ranked test for this set of experiments had three significant results when comparing all 20 datasets: Specificity (p-value: 0.001), Accuracy (p-value: 0.00453) and GMean (p-value: 0.01743), all in favour of the Lexic approach. The non-significant result for Sensitivity was p-value: 0.119. Regarding the impact of the Lexic approach in the resulting models, there was no significant change from the results reported in the previous subsection, with the resulting RFs having about 25% nodes with a more recent feature being selected after a tie.

The results of this current subsection and the previous subsection corroborate the core principle of the lexicographic feature-selection approach, that adding a bias in favour of more recent features could increase predictive accuracy. However, even though there was clearly a significant change in the learned Random Forest models, across all datasets, the resulting reflection on predictive accuracy is not expected to be large. This is because the split features eligible for replacement are, by design, equivalent from each other in terms of information gain. However, in these experiments the increased

Table 5 Accuracy and Geometric Mean of Sensitivity and Specificity (GMean) for Lexic and NoLexic random forests: optimising two parameters for both Lexic and NoLexic random forests

Datasets	Accuracy		GMean	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.684 ± 0.007	0.681 ± 0.008	0.693 ± 0.018	0.688 ± 0.021
EN_Arthritis	0.64 ± 0.006	0.633 ± 0.004	0.633 ± 0.006	0.627 ± 0.004
EN_Cataract	0.656 ± 0.006	0.652 ± 0.004	0.67 ± 0.005	0.667 ± 0.005
EN_Dementia	0.735 ± 0.005	0.734 ± 0.006	0.732 ± 0.018	0.732 ± 0.019
EN_Diabetes	0.848 ± 0.006	0.846 ± 0.006	0.852 ± 0.006	0.851 ± 0.008
EN_HBP	0.69 ± 0.004	0.69 ± 0.006	0.697 ± 0.004	0.699 ± 0.006
EN_HeartAttack	0.699 ± 0.007	0.693 ± 0.008	0.707 ± 0.011	0.707 ± 0.007
EN_Osteoporosis	0.657 ± 0.006	0.655 ± 0.005	0.679 ± 0.008	0.678 ± 0.008
EN_Parkinsons	0.631 ± 0.008	0.616 ± 0.011	0.654 ± 0.04	0.634 ± 0.037
EN_Stroke	0.673 ± 0.007	0.655 ± 0.016	0.681 ± 0.014	0.667 ± 0.014
EC_Angina	0.714 ± 0.003	0.709 ± 0.003	0.724 ± 0.015	0.721 ± 0.011
EC_Arthritis	0.734 ± 0.004	0.735 ± 0.003	0.73 ± 0.004	0.732 ± 0.003
EC_Cataract	0.656 ± 0.004	0.65 ± 0.003	0.675 ± 0.005	0.668 ± 0.004
EC_Dementia	0.768 ± 0.004	0.761 ± 0.005	0.795 ± 0.016	0.765 ± 0.015
EC_Diabetes	0.683 ± 0.004	0.683 ± 0.004	0.712 ± 0.004	0.709 ± 0.005
EC_HBP	0.647 ± 0.004	0.653 ± 0.005	0.649 ± 0.004	0.657 ± 0.005
EC_HeartAttack	0.684 ± 0.003	0.685 ± 0.003	0.689 ± 0.018	0.682 ± 0.014
EC_Osteoporosis	0.699 ± 0.005	0.695 ± 0.004	0.697 ± 0.012	0.681 ± 0.009
EC_Parkinsons	0.689 ± 0.005	0.685 ± 0.007	0.649 ± 0.046	0.694 ± 0.027
EC_Stroke	0.691 ± 0.007	0.695 ± 0.006	0.704 ± 0.012	0.693 ± 0.011
AvgRank ELSA-nurse	**1.1**	1.9	**1.2**	1.8
AvgRank ELSA-core	**1.5**	**1.5**	**1.3**	1.7
AvgRank Overall	**1.3**	1.7	**1.25**	1.75

predictive accuracy associated with the lexicographic approach was enough to lead to statistically significant results in most cases.

5.3 Decision Tree Results Optimising Only Lexicographic Threshold

The results of our experiments comparing the standard (non-lexicographic) and the lexicographic J48 decision trees with default parameter settings (except that the lexicographic J48 decision tree optimises the tie-threshold via internal cross-validation) are presented in Tables 6 and 7.

In these experiments the predictive performances of both the Lexic and NoLexic versions of J48 were almost equivalent, with slightly smaller (better) average rank for Lexic in Sensitivity and Accuracy and ties for Specificity and GMean, when considering the overall average ranks over the 20 datasets. Considering each type of data source separately, for ELSA-nurse datasets we have Lexic with smaller (better) average ranks for Sensitivity, Accuracy and GMean, and a tie for Specificity. Surprisingly, for ELSA-core datasets NoLexic wins by a small margin for Sensitivity, Accuracy and GMean, and also ties for Specificity.

Table 6 Sensitivity and Specificity results for Lexic and NoLexic decision trees, optimising only the parameter tie-threshold for Lexic decision tree

Dataset	Sensitivity		Specificity	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.617 ± 0.041	0.608 ± 0.039	0.574 ± 0.009	0.589 ± 0.001
EN_Arthritis	0.561 ± 0.009	0.57 ± 0.007	0.571 ± 0.007	0.56 ± 0.006
EN_Cataract	0.596 ± 0.014	0.598 ± 0.016	0.585 ± 0.008	0.578 ± 0.011
EN_Dementia	0.683 ± 0.04	0.679 ± 0.039	0.649 ± 0.01	0.622 ± 0.008
EN_Diabetes	0.808 ± 0.009	0.804 ± 0.01	0.797 ± 0.007	0.805 ± 0.007
EN_HBP	0.618 ± 0.008	0.587 ± 0.012	0.616 ± 0.005	0.603 ± 0.008
EN_HeartAttack	0.641 ± 0.025	0.646 ± 0.025	0.621 ± 0.007	0.638 ± 0.013
EN_Osteoporosis	0.614 ± 0.029	0.594 ± 0.015	0.612 ± 0.009	0.63 ± 0.004
EN_Parkinsons	0.646 ± 0.074	0.633 ± 0.058	0.636 ± 0.018	0.5 ± 0.025
EN_Stroke	0.629 ± 0.022	0.629 ± 0.03	0.57 ± 0.008	0.575 ± 0.009
EC_Angina	0.675 ± 0.027	0.674 ± 0.024	0.632 ± 0.011	0.635 ± 0.007
EC_Arthritis	0.724 ± 0.005	0.706 ± 0.007	0.665 ± 0.005	0.661 ± 0.008
EC_Cataract	0.646 ± 0.013	0.661 ± 0.011	0.652 ± 0.006	0.642 ± 0.007
EC_Dementia	0.715 ± 0.031	0.747 ± 0.022	0.789 ± 0.008	0.795 ± 0.01
EC_Diabetes	0.652 ± 0.016	0.673 ± 0.01	0.686 ± 0.005	0.689 ± 0.005
EC_HBP	0.621 ± 0.011	0.614 ± 0.012	0.598 ± 0.008	0.592 ± 0.006
EC_HeartAttack	0.636 ± 0.024	0.631 ± 0.028	0.659 ± 0.008	0.634 ± 0.008
EC_Osteoporosis	0.68 ± 0.03	0.684 ± 0.024	0.661 ± 0.006	0.638 ± 0.01
EC_Parkinsons	0.612 ± 0.068	0.657 ± 0.056	0.64 ± 0.018	0.676 ± 0.019
EC_Stroke	0.644 ± 0.017	0.649 ± 0.026	0.6 ± 0.011	0.666 ± 0.005
AvgRank ELSA-nurse	**1.35**	1.65	**1.5**	**1.5**
AvgRank ELSA-core	1.6	**1.4**	**1.5**	**1.5**
AvgRank Overall	**1.48**	1.53	**1.5**	**1.5**

None of the Wilcoxon signed rank tests were significant for this set of experiments, but notably the largest differences in average ranks were found in the ELSA-nurse results in favour of the Lexic approach (for the Sensitivity and Accuracy metrics).

Regarding the impact of the lexicographic approach on the resulting decision trees, ties happened much more often as every single feature is considered at each node, and replacements also happened more often: for about of 30% of the nodes. We observed that the tree size (number of nodes and leaf nodes) was not significantly changed by using the lexicographic approach (i.e., no clear pattern, some trees were slightly larger or smaller when applying the changed split function).

5.4 Decision Tree Results Optimising Two Parameters

The results of our experiments comparing the standard (non-lexicographic) and lexicographic J48 decision trees, each with two parameters optimised via internal cross-validation, are shown in Tables 8 and 9. Recall that the parameters optimised by the NoLexic J48 decision trees were the pruning confidence factor C and minimum number of instances in a leaf node, M; whilst the parameters optimised by Lexic J48 decision trees were C and the

Table 7 Accuracy and Geometric Mean of Sensitivity and Specificity (GMean) for Lexic and NoLexic decision trees: optimising only the tie-threshold parameter for Lexic decision trees

Dataset	Accuracy		GMean	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.615 ± 0.009	0.608 ± 0.01	0.595 ± 0.02	0.599 ± 0.02
EN_Arthritis	0.565 ± 0.006	0.566 ± 0.004	0.566 ± 0.006	0.565 ± 0.004
EN_Cataract	0.592 ± 0.004	0.592 ± 0.007	0.59 ± 0.005	0.588 ± 0.006
EN_Dementia	0.683 ± 0.009	0.678 ± 0.007	0.666 ± 0.018	0.65 ± 0.021
EN_Diabetes	0.806 ± 0.006	0.804 ± 0.006	0.802 ± 0.007	0.805 ± 0.006
EN_HBP	0.617 ± 0.004	0.594 ± 0.006	0.617 ± 0.004	0.595 ± 0.007
EN_HeartAttack	0.640 ± 0.007	0.646 ± 0.012	0.631 ± 0.013	0.642 ± 0.012
EN_Osteoporosis	0.613 ± 0.006	0.598 ± 0.004	0.613 ± 0.011	0.612 ± 0.008
EN_Parkinsons	0.646 ± 0.018	0.632 ± 0.024	0.641 ± 0.039	0.563 ± 0.025
EN_Stroke	0.626 ± 0.007	0.625 ± 0.008	0.599 ± 0.012	0.601 ± 0.014
EC_Angina	0.673 ± 0.01	0.672 ± 0.006	0.653 ± 0.012	0.654 ± 0.012
EC_Arthritis	0.700 ± 0.002	0.688 ± 0.006	0.694 ± 0.002	0.683 ± 0.006
EC_Cataract	0.647 ± 0.006	0.655 ± 0.004	0.649 ± 0.007	0.651 ± 0.005
EC_Dementia	0.717 ± 0.008	0.748 ± 0.01	0.751 ± 0.015	0.771 ± 0.012
EC_Diabetes	0.656 ± 0.005	0.675 ± 0.005	0.669 ± 0.008	0.681 ± 0.006
EC_HBP	0.612 ± 0.002	0.606 ± 0.004	0.609 ± 0.002	0.603 ± 0.005
EC_HeartAttack	0.637 ± 0.007	0.631 ± 0.007	0.647 ± 0.01	0.632 ± 0.014
EC_Osteoporosis	0.678 ± 0.006	0.680 ± 0.01	0.67 ± 0.015	0.661 ± 0.015
EC_Parkinsons	0.612 ± 0.018	0.658 ± 0.019	0.626 ± 0.044	0.667 ± 0.028
EC_Stroke	0.642 ± 0.001	0.650 ± 0.005	0.622 ± 0.008	0.658 ± 0.014
AvgRank ELSA-nurse	**1.25**	1.75	**1.4**	1.6
AvgRank ELSA-core	1.60	**1.40**	1.6	**1.4**
AvgRank Overall	**1.43**	1.58	**1.5**	**1.5**

tie-threshold (Sect. 3.2), with both J48 versions using the same computational budget for parameter optimisation.

In these experiments, there was a clear trend against the Lexic approach. The NoLexic method had lower (better) average ranks in all cases, and only for Specificity the p-value of the Wilcoxon signed-rank test comparing all 20 results was not significant (p-value: 0.117). The Significant p-values for Sensitivity, Accuracy and GMean were 0.0049, 0.002 and 0.0025, respectively. We believe this is due to the decision tree's volatility and sensibility to parameter tuning, which will be discussed in more detail in the Conclusions section. From these results we can conclude that a single decision-tree classifier does not benefit from the lexicographic feature-selection approach, in regards to predictive accuracy, as the RF does.

In this set of experiments, regarding the perceived impact of the lexicographic split in the resulting models, there was no significant change from the results reported in the previous subsection. That is, for about 30% of the decision tree nodes a more recent feature was selected after candidate features tied with similar information gain ratios.

In summary, the lexicographic feature-selection approach has a significant impact on how recent the selected features of a decision tree-based model are, resulting in decision

Table 8 Sensitivity and Specificity results for Lexic and NoLexic decision trees: optimising two parameters for both Lexic and NoLexic decision trees

Dataset	Sensitivity		Specificity	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.622 ± 0.039	0.615 ± 0.044	0.599 ± 0.01	0.623 ± 0.013
EN_Arthritis	0.582 ± 0.008	0.583 ± 0.009	0.575 ± 0.006	0.572 ± 0.007
EN_Cataract	0.581 ± 0.007	0.62 ± 0.023	0.622 ± 0.011	0.608 ± 0.011
EN_Dementia	0.66 ± 0.039	0.66 ± 0.039	0.686 ± 0.008	0.701 ± 0.013
EN_Diabetes	0.792 ± 0.011	0.827 ± 0.014	0.813 ± 0.007	0.82 ± 0.009
EN_HBP	0.62 ± 0.013	0.644 ± 0.014	0.626 ± 0.009	0.622 ± 0.01
EN_HeartAttack	0.627 ± 0.025	0.661 ± 0.034	0.646 ± 0.013	0.633 ± 0.01
EN_Osteoporosis	0.605 ± 0.015	0.634 ± 0.023	0.609 ± 0.004	0.629 ± 0.012
EN_Parkinsons	0.685 ± 0.056	0.685 ± 0.056	0.608 ± 0.019	0.608 ± 0.019
EN_Stroke	0.647 ± 0.03	0.618 ± 0.036	0.629 ± 0.009	0.643 ± 0.011
EC_Angina	0.64 ± 0.024	0.612 ± 0.038	0.671 ± 0.006	0.677 ± 0.012
EC_Arthritis	0.686 ± 0.01	0.685 ± 0.008	0.722 ± 0.009	0.732 ± 0.007
EC_Cataract	0.644 ± 0.013	0.75 ± 0.013	0.642 ± 0.008	0.619 ± 0.008
EC_Dementia	0.801 ± 0.027	0.821 ± 0.04	0.697 ± 0.011	0.7 ± 0.014
EC_Diabetes	0.687 ± 0.01	0.703 ± 0.014	0.663 ± 0.005	0.675 ± 0.007
EC_HBP	0.622 ± 0.009	0.635 ± 0.009	0.627 ± 0.004	0.636 ± 0.007
EC_HeartAttack	0.606 ± 0.028	0.657 ± 0.028	0.629 ± 0.008	0.618 ± 0.011
EC_Osteoporosis	0.61 ± 0.024	0.654 ± 0.017	0.683 ± 0.01	0.698 ± 0.014
EC_Parkinsons	0.603 ± 0.058	0.603 ± 0.058	0.683 ± 0.025	0.68 ± 0.024
EC_Stroke	0.567 ± 0.026	0.628 ± 0.014	0.662 ± 0.005	0.661 ± 0.009
AvgRank ELSA-nurse	1.7	**1.3**	1.55	**1.45**
AvgRank ELSA-core	1.75	**1.25**	1.6	**1.4**
AvgRank Overall	1.725	**1.275**	1.575	**1.425**

trees and Random Forests where on average 25-30% of the nodes have split features based on more recent data.

For RF classifiers, the positive impact of the lexicographic split was clear, as it increased predictive performance in most cases, with several statistically significant results. Hence, we recommend using the lexicographic approach with tree-ensemble classifiers such as RF.

For decision trees, however, the Lexic approach optimising only the tie-threshold produced decision trees with predictive accuracy comparable to the trees produced by the NoLexic approach without parameter optimisation. However, when both the Lexic and NoLexic J48 decision trees were allowed to optimise two parameters with the same computational budget, the NoLexic approach outperformed the Lexic approach, with statistical significance in most cases. However, the use of the lexicographic feature-selection criterion in decision tree classifiers applied to longitudinal could still be beneficial for other reasons (beyond accuracy), e.g. to learn decision tree models with more recent features, which could be more easily acceptable by users (considering that, for most diseases, a diagnosis based on recent medical tests seems more meaningful than a diagnosis based on older tests).

Table 9 Accuracy and the Geometric Mean of Sensitivity and Specificity (GMean) for Lexic and NoLexic decision trees: optimising two parameters for both Lexic and NoLexic decision trees

Datasets	Accuracy		GMean	
	Lexic	NoLexic	Lexic	NoLexic
EN_Angina	0.6 ± 0.01	0.622 ± 0.012	0.608 ± 0.02	0.614 ± 0.022
EN_Arthritis	0.578 ± 0.004	0.577 ± 0.007	0.578 ± 0.004	0.577 ± 0.007
EN_Cataract	0.609 ± 0.008	0.612 ± 0.005	0.601 ± 0.006	0.612 ± 0.008
EN_Dementia	0.686 ± 0.007	0.7 ± 0.012	0.67 ± 0.021	0.677 ± 0.02
EN_Diabetes	0.811 ± 0.006	0.822 ± 0.008	0.802 ± 0.006	0.823 ± 0.008
EN_HBP	0.624 ± 0.007	0.631 ± 0.006	0.622 ± 0.007	0.632 ± 0.006
EN_HeartAttack	0.645 ± 0.012	0.634 ± 0.008	0.634 ± 0.012	0.644 ± 0.014
EN_Osteoporosis	0.608 ± 0.004	0.629 ± 0.011	0.606 ± 0.008	0.63 ± 0.012
EN_Parkinsons	0.609 ± 0.019	0.609 ± 0.019	0.64 ± 0.028	0.64 ± 0.028
EN_Stroke	0.63 ± 0.008	0.641 ± 0.01	0.636 ± 0.014	0.627 ± 0.017
EC_Angina	0.67 ± 0.006	0.675 ± 0.011	0.654 ± 0.012	0.639 ± 0.017
EC_Arthritis	0.708 ± 0.005	0.714 ± 0.005	0.703 ± 0.005	0.708 ± 0.005
EC_Cataract	0.643 ± 0.004	0.658 ± 0.005	0.642 ± 0.004	0.681 ± 0.005
EC_Dementia	0.699 ± 0.011	0.702 ± 0.013	0.746 ± 0.012	0.755 ± 0.017
EC_Diabetes	0.667 ± 0.005	0.679 ± 0.005	0.675 ± 0.006	0.689 ± 0.005
EC_HBP	0.625 ± 0.005	0.636 ± 0.006	0.624 ± 0.005	0.635 ± 0.006
EC_HeartAttack	0.627 ± 0.007	0.62 ± 0.01	0.615 ± 0.014	0.635 ± 0.012
EC_Osteoporosis	0.677 ± 0.01	0.694 ± 0.013	0.645 ± 0.015	0.675 ± 0.011
EC_Parkinsons	0.683 ± 0.024	0.68 ± 0.024	0.631 ± 0.025	0.629 ± 0.025
EC_Stroke	0.657 ± 0.005	0.659 ± 0.009	0.611 ± 0.014	0.644 ± 0.008
AvgRank ELSA-nurse	1.75	**1.25**	1.75	**1.25**
AvgRank ELSA-core	1.8	**1.2**	1.8	**1.2**
AvgRank Overall	1.775	**1.225**	1.775	**1.225**

6 Interpreting the most accurate classification models

In this Section we will discuss what insight we can get from analysing our classification models, by analysing the best RF models learned from our datasets. Before proceeding, it is important to explain why we chose to interpret RF models, rather than decision trees, since decision trees are a directly interpretable type of model representation in general, due to their graphical nature (Quinlan 1993; Freitas 2014). Our motivation for only analysing the best RF models is twofold. First, the decision trees in our models tended to be too large for interpretation; and second, the RF models substantially outperformed the decision trees in terms of predictive performance. We believe it is preferable in this case to interpret the best models in terms of predictive performance, rather than arguably more interpretable models with sub-optimal performance.

For RF models, directly interpreting each random tree in the forest is not feasible, due to the large number of trees. However, we can calculate feature importance measures such as the average value of the information gain ratio (entropy metric used in our RF models) across the nodes where the feature was selected. Hence, we can indirectly

discuss how the RF models make predictions by reporting the most important features for classification across all trees in the forest.

For our analysis in this Section we report the most important features in the best RF models for our datasets, where these "best models" were selected as follows. First, we selected the datasets that, over all four experimental setups (Lexic vs NoLexic, optimising only Lexic tie-threshold vs optimising two parameters), had at least one result above a minimum threshold of 0.7 average GMean. Then, we selected the best model obtained for each of these datasets. We focus here on the GMean because it assigns equal importance to maximise the predictive accuracy for both the majority and minority classes, unlike the Accuracy measure, which unduly assigns greater importance to the majority class.

This led to the selection of 9 models, all using random forests: 3 models from ELSA-nurse datasets and 6 from ELSA-core datasets. In 8 of these 9 selected datasets the Lexic version of RF outperformed the NoLexic version; the only exception was the model for the ELSA-nurse Diabetes dataset in Table 3, where NoLexic had a 0.855 average GMean and Lexic had 0.853, a very small difference of 0.002. For consistency, we chose to interpret only models learned by the Lexic version of RFs.

In the second step of this analysis, for each of the 9 selected datasets, we trained a new Lexic RF model using the entire dataset (i.e., no training and test set division), to ensure that the models interpreted in this analysis would consider all available data, maximising their quality. As done earlier, we coped with the imbalanced classes by training the models using the Balanced Random Forest undersampling approach, meaning each random tree in the forest is learned from a bootstrap sample which is undersampled to a 1:1 ratio of instances of both classes. The datasets had their missing values replaced in a preprocessing phase as mentioned earlier. The parameter values for the RFs optimising two parameters were those selected most often during the 10-fold cross-validation, and were the default values for the other models. However, we increased the number of trees from the default 100 to 1000, to have a more reliable set of most important features.

Table 10 shows the 5 top-ranked features (i.e. most important features) in each of the selected RF models from ELSA-nurse datasets. The feature ranking is based on the average impurity decrease (AID, the arithmetic mean of information gain ratio), calculated over all nodes where the feature was selected, in all trees in the RF, for each dataset. This measure represents the predictive power associated with the feature in the trees. In Tables 10 and 11, the "_w" suffix at the end of a feature's name indicates the wave number (time point when the feature was measured).

For the ELSA-nurse Dementia model, we have three different measurements (at different time points) of the *hastro* variable, related to heart disease. There are several studies that connect heart disease and risk of dementia, and some heart conditions such as coronary heart disease are widely accepted to be a risk factor for dementia (Wolters et al. 2018). The other two variables are related to respiratory infections and mobility, which are more often connected to consequences of dementia than risk factors (Eisenmann et al. 2020).

For the Diabetes RF model, the age (*indager_w8*) and *sex* features were the highest ranked. Naturally, all age-related diseases are correlated with the age feature, and diabetes is more prevalent among men (Gale and Gillespie 2001). The other top features are 3 blood sample features, namely *cfib*, *clotb* and *hgb*. Diabetes is known to increase the chance of heart diseases (Dal Canto et al. 2019), so it is possible the RF models detected patterns among ELSA respondents with heart or blood pressure problems.

For the ELSA-nurse Heart Attack model, five different variables were selected, only one of them directly related to heart disease (clotting disorder). Among the others, height and the blood triglyceride level are indirectly associated with risk of

Table 10 The 5 features with the greatest average impurity decrease (AID) values in the best RF models for ELSA-nurse datasets

Selected model (average GMean)	Feature	Description	AID
Dementia Lexic RF optimising only tie-threshold (0.732)	hastro_w6	Whether been admitted to hospital with a heart complaint in the past month	1
	chestin_w2	Lung function: Whether had any respiratory infection in last 3 weeks	0.92
	mmstre_w2	Outcome of semi-tandem stand	0.92
	hastro_w4	Whether been admitted to hospital with a heart complaint in the past month	0.72
	hastro_w2	Whether been admitted to hospital with a heart complaint in the past month	0.67
Diabetes Lexic RF optimising only tie-threshold (0.853)	sex	Sex of the participant	0.6
	indager_w8	Age at wave 8	0.51
	cfib_w8	Blood Fibrinogen level (g/l)	0.49
	clotb_w8	Blood sample: whether has clotting disorder	0.48
	hgb_w8	Blood haemoglobin level (g/dl)	0.46
Heart Attack Lexic RF optimising only tie-threshold (0.722)	clotb_w4	Blood sample: whether has clotting disorder	0.76
	eyesurg_w6	Whether have a detached retina or had eye or ear surgery in the past 3 months	0.45
	htval_w2	Height (cm)	0.45
	trig_w2	Blood triglyceride level (mmol/l)	0.44
	mmlsre_w2	Leg raise (eyes shut): Outcome	0.43

Table 11 The 5 features with the greatest average impurity decrease (AID) values in the best RF models for 6 ELSA-core datasets

Selected model (average GMean)	Feature	Description	AID
Angina Lexic RF optimising two parameters (0.723)	hefrac_w4	Whether has fractured hip	0.96
	hecanb_w5	Cancer: whether received treatment in last 2 years	0.81
	hecanb_w1	Cancer: whether received treatment in last 2 years	0.81
	heji_w3	Whether had joint replacement	0.77
	helng_w2	Whether taking medication for lung condition	0.73
Arthritis Lexic RF optimising only tie-threshold (0.735)	cesd_w7	Depression questionnaire score	0.56
	indager_w7	Age of the participant at a given wave	0.56
	dicdnm_w7	Cause of death of mother of respondent	0.56
	cesd_w5	Depression questionnaire score	0.56
	heiadlX-of-9_w6	Reported IADL difficulties (count)	0.54
Dementia Lexic RF optimising two parameters (0.795)	helng_w3	Whether taking medication for lung condition	1
	hefrac_w5	Whether has fractured hip	0.92
	heji_w1	Whether had joint replacement	0.88
	helng_w5	Whether taking medication for lung condition	0.79
	hepsyX-of-9_w5	Psychiatric problem report counts	0.69
Diabetes Lexic RF optimising only tie-threshold(0.717)	hefrac_w1	Whether has fractured hip	0.64
	helng_w4	Whether taking medication for lung condition	0.63
	heji_w3	Whether had joint replacement	0.63
	heam_w2	Whether taking medication for asthma	0.63
	indager_w7	Age of the participant at a given wave	0.6
Parkinson's Disease Lexic RF optimising only tie-threshold (0.717)	heji_w5	Whether had joint replacement	1
	heyrc_w4	Experienced psychiatric problems in last 2 years	0.95
	helng_w4	Whether taking medication for lung condition	0.81
	hepawX-of-7_w5	Pain reported (count)	0.76
	hecanb_w4	Cancer: whether received treatment in last 2 years	0.74

Table 11 (continued)

Selected model (average GMean)	Feature	Description	AID
Stroke Lexic RF optimising only tie-threshold (0.713)	hefrac_w5	Whether has fractured hip	0.88
	hefrac_w1	Whether has fractured hip	0.74
	hefrac_w4	Whether has fractured hip	0.68
	heyrc_w1	Experienced psychiatric problems in last 2 years	0.66
	heill_w5	Whether has self-reported long-standing illness	0.61

cardiovascular disease (Samaras 2013; Reiner 2017). The other two variables selected are more related to general health history and mobility.

Table 11 shows the 5 best-ranked features in each of the best RF models learned for the 6 selected ELSA-core datasets.

In the ELSA-core Angina model, only one top feature, the *helng*, is clearly connected to the class variable, as chest pain is a common side-effect of respiratory diseases (Indrakumari et al. 2020). Regarding the other four variables, a history of recent cancer treatment, joint replacements and fractured hips are likely associated with older age and frailty, which naturally increase the risk of cardiovascular disease.

In the ELSA-core Arthritis model, two measurements (at different time points) of the feature that measures a self-reported depression score were selected among the top predictors. Depression is considered a risk factor for several age-related diseases, including Arthritis (Vallerand et al. 2019). Among the other selected variables we have the age of the respondent, and the IADL (Instrumental Activities of Daily Living) score, both predictors of overall health. Finally, the model selected the cause of death of the mother as a top feature which is likely related to the fact that some hereditary factors can increase the risk of Arthritis (Ren et al. 2020).

For the ELSA-core Dementia model one of the top features was *hepsyX-of-9*, an indication of mental health self-reported by the patient that is clearly related to the class variable. The other top-ranked features are about the medical history and current health status of the respondent, namely *helng*, *heji* and *hefrac*. As mentioned, such features are likely used as general measurements of overall health and frailty.

Regarding the ELSA-core Diabetes model, diabetes is connected to several other health complications, such as asthma (Perez and Piedimonte 2014) (explaining the *helng* and *heam* variables), and hip fractures (Vilaca et al. 2020) (explaining the *hefrac* and *heji* variables). Finally, the age of the participant was chosen among the top predictors.

Regarding the ELSA-core Parkinson's Disease model, among the top features selected by this model we have again some variables related to overall health medical history (*heji*, *helng* and *hecanb*). The other two variables selected among the top features are more directly related to this disease: *heyrc* regards mental health, and *hepaw* is related to reported pain. The latter has been correlated with mental health as patients may report pain less often if they are suffering from cognitive decline (McAuliffe et al. 2012).

Finally, the ELSA-core Stroke model selected the variable *hefrac* three times (at different time points) among its top 5 features. Hip fractures have been a good predictor of general frailty over several models and were possibly used to separate younger and healthier ELSA respondents from those in greater risk of developing the target age-related disease. The *heill* variable is also indirectly associated with poor health and frailty. Regarding the *heyrc* variable, psychological distress has been associated with an increased risk of Stroke (Surtees et al. 2008), so this might be a good predictor that is more directly connected to the class variable.

In summary, the ELSA-core variables are mostly focused on self-reported health and wellbeing assessments, as well as medical history. Thus, the connections between the selected top features and the target variables are expected to be less direct in these models. General representations of physical frailty such as hip fractures were often used as predictors by our models, but they also often selected variables that are known risk factors of the age-related diseases they were predicting.

7 Conclusions

In this article we reported extended computational results for our recently proposed adaptation of decision tree-based classifiers for longitudinal datasets (Ribeiro and Freitas 2020). This adaptation is a lexicographic bi-objective feature-selection approach, which uses time-related information available in longitudinal data when selecting features during the training of decision tree-based classifiers, which are a popular type of classifier in biomedical applications – particularly ensembles of decision trees, like random forests. The results in this article extended the results in Ribeiro and Freitas (2020) by including two types of decision tree-based classifiers (random forests and J48), reporting on experiments optimising up to two parameters of each algorithm, using 20 longitudinal classification datasets from two data sources (with different types of features and numbers of feature waves); whilst the results in Ribeiro and Freitas (2020) included only one type of decision tree-based classifier (random forests, optimising only the lexicographic tie-threshold parameter) and 10 longitudinal classification datasets from a single data source (with a single number of waves). Hence, the extended experiments led to a more robust evaluation of the effectiveness of the proposed lexicographic approach for longitudinal classification.

The rationale for the lexicographic adaptation is that more recent measurements of a feature are intuitively more valuable for increasing predictive accuracy, particularly when predicting the occurrence of age-related diseases (like in this work). Features measured closer to the target variable's measurement tend to more actionable, as they often represent information that remains currently relevant, as opposed to older measurements of the feature. More recent features are also less likely to have missing data due to attrition, so adding a bias in their favour reduces the chances of selecting features with missing or estimated values when training the classifier.

The proposed approach can be summarised as follows. The lexicographic split adds the time-indexes of the candidate features as a secondary objective, to be used as a tie-breaking criterion between features with equivalent information gain ratios (or other primary selection criterion). In order to determine when candidate features can be considered equivalent regarding the primary criterion, we use a tie-threshold parameter, so that features with information gain ratio differences smaller than this threshold are considered equivalent. The algorithm then uses the time-indexes of the tied features as the selection criterion instead, with the most recent feature being selected. In order to avoid having this additional parameter being manually and subjectively selected by the user, we implemented a data-driven automated threshold selection, which is a more reliable way to set the tie-threshold value using the training data.

We performed experiments using 20 real-world datasets prepared for this study, created from the English Longitudinal Study of Ageing (ELSA), a prominent longitudinal study from the United Kingdom. Our experiments compared the standard (non-lexicographic) and adapted (lexicographic) versions of decision trees and random forests, as a way to gauge the impact of the adaptation on a single decision tree and on ensemble decision tree-based classifiers. We performed two types of experiments regarding the algorithms' parameter optimisation. In the first type, we optimised only the tie-threshold parameter of the lexicographic decision tree and random forest algorithms. In the second type of experiment, we optimised two parameters of both the lexicographic and non-lexicographic versions of both decision tree and random forest algorithms in controlled experiments, giving the same computational budget to each version of each algorithm.

The proposed lexicographic approach improved the predictive accuracy of random forest classifiers, when compared to the standard split criterion based only on the information gain ratio, in the majority of the experiments. The results were statistically significant in most cases when both the lexicographic and the non-lexicographic random forests had two parameters optimised; and the results were significant in some cases when we optimised only the tie-threshold parameter of lexicographic random forest.

For decision tree classifiers, the lexicographic and non-lexicographic classifiers had similar performances (no statistically significant differences) when only the tie-threshold of the lexicographic decision tree was optimised, but the non-lexicographic classifiers had statistically significantly better performance in most cases when both the lexicographic and the non-lexicographic decision trees had two parameters optimised. We believe this latter unexpected result was due to decision tree classifiers being more sensitive to parameter tuning, which may have given the non-lexicographic approach an advantage. This is because, as we gave the same computational budget for parameter optimisation to both the lexicographic and the non-lexicographic decision tree classifiers, the lexicographic classifier spent part of that budget optimising the tie-threshold parameter (which is not used by the non-lexicographic classifier); whilst the non-lexicographic classifier was able to spend its entire budget on optimising parameters that turned out to be more important, leading to higher predictive performance.

We also investigated how often the lexicographic approach led to the decision tree or random forest algorithm to select a different feature. On average 25% of nodes in RFs and 30% of nodes in DTs selected different features because of the lexicographic split function. Therefore, the added bias in favour of more recent features resulted in considerably different classifiers for our longitudinal datasets, shifting the average time-index of selected features further towards the most recent wave (time point), i.e., the wave of the class label. It is important to highlight that longitudinal datasets tend to grow over time, so the impact of the lexicographic approach also tends to increase as new waves are added to longitudinal datasets.

We also interpreted the top features in our best random forest models, as an additional contribution. We were able to find several existing connections between the top-ranking features and peer-reviewed medical research.

As a final consideration, our investigations have shown that there is more exploration to be done in identifying and making use of the temporal information that longitudinal data brings. As more longitudinal studies progress, and more longitudinal data becomes available, it becomes more and more important to develop machine learning algorithms specialised for longitudinal classification, particularly considering that this is still an under-explored area in machine learning. Therefore, we hope that our results can encourage the development of other methods for coping with longitudinal data, including potentially the application of the lexicographic approach to other types of machine learning algorithms.

Regarding code availability, we have created a public GitHub project website (github.com/caioedurib/lexic_split_Weka) where we made the source code of our implementation of the Lexicographic Split Function available in a Java script, as well as instructions for using it with the Weka data mining tool.

Author contributions This research project was jointly conceived by both authors. In addition, both authors designed the structure of the datasets and the overall machine learning methodology used in the experiments. All the program code implementation and all computational experiments were performed by CR. Both jointly analysed the results. The manuscript was written mainly by CR, but both authors contributed to writing and revising the manuscript.

Declarations

Competing interest The authors declare no competing interests.

References

Abell J, Amin-Smith N, Banks J, Batty GD, Breeden J, Buffel T, Cadar D, Crawford R, Demakakos P, de Oliveira C, Hussey D, Lassale C, Matthews K, Nazroo J, Norton M, Oldfield Z, Oskala A, Prattley J, Steptoe A, Zaninotto P (2018) The dynamics of ageing: evidence from the English Longitudinal Study of Ageing 2002-2016 (Wave 8). Institute for Fiscal Studies, London. https://doi.org/10.1920/re.ifs.2019.0000. https://www.ifs.org.uk/publications/13510

Aghili M, Tabarestani S, Adjouadi M, Adeli E (2018) Predictive modeling of longitudinal data for Alzheimer's disease diagnosis using rnns. In: International workshop on PRedictive Intelligence In MEdicine, pp 112–119. Springer

Bagnall A, Lines J, Bostrom A, Large J, Keogh E (2017) The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining Knowl Discov 31(3):606–660

Banks J, Batty G, Coughlin K, Deepchand K, Marmot M, Nazroo J, Oldfield Z, Steel N, Steptoe MA, Wood, Zaninotto P (2019) English longitudinal study of ageing: waves 0–8, 1998–2017 [data collection]

Basgalupp MP, Barros RC, de Carvalho AC, Freitas AA, Ruiz DD (2009) Legal-tree: a lexicographic multi-objective genetic algorithm for decision tree induction. In: Proceedings of the 2009 ACM symposium on applied computing. ACM, pp 1085–1090

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Chen C, Liaw A, Breiman L et al (2004) Using random forest to learn imbalanced data. Univ Calif Berkeley 110(1–12):24

Dal Canto E, Ceriello A, Rydén L, Ferrini M, Hansen TB, Schnell O, Standl E, Beulens JW (2019) Diabetes as a cardiovascular risk factor: an overview of global trends of macro and micro vascular complications. Eur J Prev Cardiol 26(2_suppl):25–32

Deng H, Runger G, Tuv E, Vladimir M (2013) A time series forest for classification and feature extraction. Inf Sci 239:142–153

Eiben F, Hall MA, Witten IH (2016) The weka workbench (online appendix). In: Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, San Francisco

Eisenmann Y, Golla H, Schmidt H, Voltz R, Perrar KM (2020) Palliative care in advanced dementia. Front Psychiatry 11:699

Freitas AA (2004) A critical review of multi-objective optimization in data mining: a position paper. ACM SIGKDD Explor Newslett 6(2):77–86

Freitas AA (2014) Comprehensible classification models: a position paper. ACM SIGKDD Explor Newslett 15(1):1–10

Gale EA, Gillespie KM (2001) Diabetes and gender. Diabetologia 44(1):3–15

Indrakumari R, Poongodi T, Jena SR (2020) Heart disease prediction using exploratory data analysis. Procedia Comput Sci 173:130–139

Javeed A, Dallora AL, Berglund JS, Idrisoglu A, Ali L, Rauf HT, Anderberg P (2023) Early prediction of dementia using feature extraction battery (feb) and optimized support vector machine (svm) for classification. Biomedicines 11(2):439

Kaiser A (2013) A review of longitudinal datasets on ageing. J Popul Ageing 6(1–2):5–27

López V, Fernández A, García S, Palade V, Herrera F (2013) An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. Inf Sci 250:113–141

Luo J, Ye M, Xiao C, Ma F (2020) Hitanet: hierarchical time-aware attention networks for risk prediction on electronic health records. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 647–656

Malley JD, Malley KG, Pajevic S (2011) Statistical learning for biomedical data. Cambridge University Press, Cambridge

McAuliffe L, Brown D, Fetherstonhaugh D (2012) Pain and dementia: an overview of the literature. Int J Older People Nurs 7(3):219–226

Morid MA, Sheng ORL, Del Fiol G, Facelli JC, Bray BE, Abdelrahman S (2020) Temporal pattern detection to predict adverse events in critical care: case study with acute kidney injury. JMIR Med Inform 8(3):14

Niemann U, Hielscher T, Spiliopoulou M, Völzke H, Kühn J-P (2015) Can we classify the participants of a longitudinal epidemiological study from their previous evolution? In: 2015 IEEE 28th international symposium on computer-based medical systems (CBMS). IEEE, pp 121–126

Perez MK, Piedimonte G (2014) Metabolic asthma: is there a link between obesity, diabetes, and asthma? Immunol Allergy Clin 34(4):777–784

Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc., San Francisco

Reiner Ž (2017) Hypertriglyceridaemia and risk of coronary artery disease. Nat Rev Cardiol 14(7):401–411

Ren J, Masi AT, Aldag JC, Asche CV (2020) Hereditary, socio-behavioural, and immuno-hormonal predictors of incident rheumatoid arthritis and therapy response influences on survival versus matched control subjects using a generalised structural equation model. Clin Exp Rheumatol 38(4):640–648

Ribeiro C, Freitas AA (2019) A mini-survey of supervised machine learning approaches for coping with ageing-related longitudinal datasets. In: 3rd workshop on AI for aging, rehabilitation and independent assisted living (ARIAL), held as part of IJCAI-2019

Ribeiro C, Freitas AA (2020) A new random forest method for longitudinal data classification using a lexicographic bi-objective approach. In: 2020 IEEE symposium series on computational intelligence (SSCI). IEEE, pp 806–813

Ribeiro C, Freitas AA (2021a) Constructed temporal features for longitudinal classification of human ageing data. In: 2021 IEEE international conference on healthcare informatics. IEEE, pp 106–112

Ribeiro C, Freitas AA (2021) A data-driven missing value imputation approach for longitudinal datasets. Artif Intell Rev 54:6277–6307

Ribeiro C, Brito LHS, Nobre CN, Freitas AA, Zárate LE (2017) A revision and analysis of the comprehensiveness of the main longitudinal studies of human aging for data mining research. Wiley Interdiscip Rev: Data Mining Knowl Discov 7(3):e1202

Samaras TT (2013) Shorter height is related to lower cardiovascular disease risk-a narrative review. Indian Heart J 65(1):66–71

Scornet E, Biau G, Vert J-P et al (2015) Consistency of random forests. Ann Stat 43(4):1716–1741

Surtees P, Wainwright N, Luben R, Wareham N, Bingham S, Khaw K-T (2008) Psychological distress, major depressive disorder, and risk of stroke. Neurology 70(10):788–794

Vallerand IA, Patten SB, Barnabe C (2019) Depression and the risk of rheumatoid arthritis. Curr Opin Rheumatol 31(3):279

Vilaca T, Schini M, Harnan S, Sutton A, Poku E, Allen IE, Cummings SR, Eastell R (2020) The risk of hip and non-vertebral fractures in type 1 and type 2 diabetes: a systematic review and meta-analysis update. Bone 137:115457

Weiss GM, Provost F (2003) Learning when training data are costly: the effect of class distribution on tree induction. J Artif Intell Res 19:315–354

Wilcoxon F (1992) Individual comparisons by ranking methods. In: Breakthroughs in statistics. Springer, pp 196–202

Wolters FJ, Segufa RA, Darweesh SK, Bos D, Ikram MA, Sabayan B, Hofman A, Sedaghat S (2018) Coronary heart disease, heart failure, and the risk of dementia: a systematic review and meta-analysis. Alzheimer's Dement 14(11):1493–1504