Constructed Temporal Features for Longitudinal Classification of Human Ageing Data

Caio Ribeiro
School of Computing
University of Kent
Canterbury, UK 30332–0250
Email: C.E.Ribeiro@kent.ac.uk

Alex Freitas
School of Computing
University of Kent
Canterbury, UK 30332–0250
Email: A.A.Freitas@kent.ac.uk

Abstract—**Standard classification algorithms ignore the time-related information contained in longitudinal data, as they do not consider the time indexes of the features' different measurements. Accounting for temporal patterns may improve the algorithms' performance, when applied to longitudinal data. Representing temporal patterns in the data itself has the advantage that those patterns are generic enough to be used with existing powerful classification algorithms, without requiring the design of new and more complex algorithms to exploit them. In this article, we propose 6 different types of constructed temporal features (3 of them being novel contributions), calculated from the values of the different feature measurements taken over time, and investigate whether adding those constructed temporal features to the original longitudinal dataset improves the classification model's predictive accuracy. Our experiments involved 20 real-world longitudinal datasets created from a human-ageing study, and showed that the proposed approach of adding the constructed temporal features to the original feature set produced better classifiers overall.**

I. Introduction

Longitudinal datasets contain multiple measures of a set of features taken over different time points, following the same group of individuals (instances). Because standard classification 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 explicitly add a representation of this information as additional features in the dataset.

Longitudinal studies of human ageing commonly produce large longitudinal datasets, with thousands of features and instances, measured over several time points [1]. In these studies, a cohort of participants has their data collected over fixed (typically years long) time intervals, about many aspects of their lives, to further our understanding of the impact of ageing on individuals and the society.

In this article, we propose Constructed Temporal Features (CTFs) for the supervised machine learning problem of longitudinal classification. We evaluate the performance of the proposed CTFs on real-world datasets created from the English Longitudinal Study of Ageing (ELSA) [2]. For our experiments, we trained classifiers for 20 datasets created to predict the diagnosis of 10 age-related diseases at the most recent wave (time point) of the ELSA study, using data collected over a period of up to 12 years.

As related work, Niemann et al. [3] generated evolution features by comparing instance clustering results at different time points in a longitudinal dataset. In addition, Buizza et al. [4] created longitudinal pattern features by comparing distances and means related to two subsequent images (PET/CT scans). Both works are quite different from our study since they focus on specific types of CTFs involving clustering results and images, which are out of the scope of this work.

In a more similar context to ours, Pomsuwan and Freitas [5] have also used CTFs for longitudinal data, using datasets created from the ELSA database. Two features used in their study, Monotonicity and Diff, are also used in our study, and are described in detail in Section II. However, their study focused on proposing a new longitudinal feature selection algorithm, rather than on the CTFs; and they did not present any discussion regarding the impact of the CTFs on the predictive accuracy of their classifiers.

Our work differs from these related studies as it is the first that focuses specifically on the creation of CTFs for longitudinal data (including the proposal of new types of CTFs) and the evaluation of their impact on the predictive accuracy of a classification algorithm (Random Forests).

Constructing temporal features in a preprocessing step is only one of the possible approaches for considering temporal patterns in classification problems. Other strategies include Structural Pattern Detection [6], Recurrent Neural Networks (often Long-Short Term Memory) [7], and Deep Learning [8]. Although we are using only standard classifiers in our experiments, our approach may be combined with more sophisticated algorithms tailored to longitudinal analysis.

By itself, CTF creation has the advantages of being simple to implement and adapt, and generating interpretable features that clearly represent temporal patterns. This is in contrast e.g. to deep neural networks, where the constructed features are not directly interpretable by users.

We performed experiments adding 6 types of CTF to 20 real-world longitudinal datasets, and compare them using three performance metrics. Our experiments compared three dataset compositions: a baseline using only the original features, a feature set composed only of CTFs, and our proposed approach of combining the original features and CTFs in a single dataset. Our results showed that the Random Forest classifiers tended to perform better with our proposed approach.

II. The Proposed Constructed Temporal Features

We define six types of Constructed Temporal Features (CTFs), which can be added to longitudinal datasets to increase predictive accuracy. We describe each CTF's calculation for the numeric and (if applicable) ordered nominal features in our datasets.

We use the term "conceptual feature" to refer to the abstract definition of a feature, without specifying the time point (wave) where the feature was measured. For instance, *cholesterol* is a conceptual feature; whilst, in concrete terms, the dataset will contain different *cholesterol* features which are distinguished by the time points (waves) where they were measured. The proposed CTFs are calculated using values of different measurements of the same conceptual feature.

Regarding the novelty of the proposed CTFs, Monotonicity and Diff have been applied to similar longitudinal datasets in [5], as mentioned earlier. We could not find an example of the Ratio feature being applied to longitudinal datasets, but it is a small variation from the Diff feature. The Percentile and the two CTFs based on the age mean/mode, DiffAgeMean and AvgDiffAgeMean, are novel contributions.

A. Monotonicity

A monotonic increase or decrease of a feature's values over its consecutive measurements in a longitudinal dataset may be a temporal pattern useful for predicting the value of the class variable. To represent these patterns, we created a Monotonicity CTF with three possible values [5]: $+1$ indicating a monotonic increase in the feature's values across all its measurements, -1 indicating a monotonic decrease and 0 indicating no monotonicity pattern.

The calculation of a Monotonicity feature value is shown in Equation 1, where $F_{i,t}$ denotes the value of the i-th feature at time point (wave) t. Note that it first checks if all feature values are equal; if so, it assigns a 0 value. Hence, the rules for the -1 and 1 values apply only if there has been at least one change to the feature's values over time.

$$\begin{aligned} Monotonicity(F_i) = 0 : F_{i,0} = F_{i,1} = \cdots = F_{i,T} \\ OR \begin{cases} -1 : F_{i,0} \leq F_{i,1} \leq \cdots \leq F_{i,T} \\ 1 : F_{i,0} \geq F_{i,1} \geq \cdots \geq F_{i,T} \\ 0 : \text{ otherwise} \end{cases} \end{aligned} \tag{1}$$

Note that we do not use a strict monotonicity definition with $<$ and $>$ operators. Rather, we use a more flexible monotonicity definition with the $\leq$ and $\geq$ operators. Hence, if the feature's value increased or decreased at least once, as long as the feature's values do not change in the opposite direction in other waves, we consider this a monotonic change. The motivation for this more flexible definition is that it can be applied to both numeric and ordered nominal features. Our datasets contain several ordered nominal features taking between 2 and 8 possible values, and the above strict definition of monotonicity would not be flexible enough to cope with such ordered nominal features. For example, if a dataset has 4 waves but the feature can take only two ordered values, say ''low" and ''high", it is impossible to detect a monotonic change according to the strict definition, but a sequence of feature values such as ''low", ''low", ''high", ''high" would be recognised as a monotonic increase, a potentially useful pattern for classification.

B. Difference Between Last Two Measurements

Feature measurements taken closer in time to the class wave (the last wave) arguably have more impact on the model's output, as they are likely more closely related to the class variable than measurements of the same feature taken further in the past (earlier waves). In this context, we consider that the most recent changes to a feature's value may represent an important temporal trend. Thus, we created the Diff CTF to measure the numerical difference between the conceptual feature's last and second to last measurements.

The calculation of the Diff CTF is shown in Equation 2, where T is the index of the class variable's wave (the last wave). For ordered nominal features, Diff represents the degree of difference between the nominal values. This is only possible because all nominal features in our datasets are ordered, so we can assign numerical values to them and calculate the difference between these values as a degree of difference. Note that this degree of difference measurement is precise only for cases where the response options in the nominal features are equidistant, and we do not make that assumption, as we did not design the data. However, after inspecting all nominal features in our datasets, we decided that their values can be considered similarly distant enough that the Diff calculation would be acceptable. The same decision was made for Percentile, DiffAgeMean and AvgDiffAgeMean, where we also calculate degrees of difference for nominal features.

$$Diff(F_i) = F_{i,T} - F_{i,T-1} \tag{2}$$

C. Ratio Between Last Two Measurements

The Ratio CTF functions similarly to the Diff CTF. However, instead of the difference, it calculates the result of dividing the value of the conceptual feature's last measurement by its second to last measurement. This CTF cannot be calculated for nominal features, as that would require an assumption of equidistance between the possible values, which is not guaranteed in our datasets. Hence, in this work the Ratio CTF is used only for numeric features.

Note that the Ratio CTF can capture patterns quite different from patterns captured by the Diff CTF. For example, Diff has the same value (0.2) for the feature value pairs (0.2, 0.4) and (0.6, 0.8), whilst Ratio has value 2 for the former pair and 1.33 for the latter.

The calculation of the Ratio CTF is shown in Equation 3. To avoid a division by zero error, before performing the division we add 1 to both feature values.

$$Ratio(F_i) = \frac{F_{i,T} + 1}{F_{i,T-1} + 1} \tag{3}$$

D. Last Measurement's Difference from Age-based Mean/Mode

The subject's age is the most relevant feature in our datasets. In preliminary experiments with missing value replacement in these datasets, we have found that the mean (or mode) value of subjects of the same age as the current instance is a good estimation for an expected value of a feature [9]. Therefore, we propose a CTF to calculate the difference between a feature's value in the last wave and its ''expected" value, which is an age-based mean/mode.

The calculation of the DiffAgeMean CTF is shown in Equation 4. To calculate the expected value for a feature F_i's last measurement ($F_{i,T}$) for each subject, we get the value of that subject's age at wave T ($Age_{i,T}$) and calculate the mean over all measurements of F_i, over all waves and subjects, where a subject's age equals $Age_{i,T}$. For nominal features, the expected value is the mode among individuals of the same age, and we measure the degree of difference from that mode.

$$DiffAgeMean(F_i) = F_{i,T} - Exp(F_i, Age_{i,T}) \tag{4}$$

E. Average Difference from Age-based Mean/Mode

As an expansion of the DiffAgeMean feature, we calculate the DiffAgeMean for all different measurements of a feature, then average these results (dividing the sum of DiffAgeMean's by the number of measurements), to get an average difference from the expected values. Note that at each wave of the study, the subject's age changes, so we need to recalculate the expected value for each measurement of the current feature. The AvgDiffAgeMean CTF is calculated as shown in Equation 5. Again, for nominal features, we use the mode as the expected values, instead of the mean.

$$\text{AvgDiffAgeMean}(\text{F}_i) = \frac{\sum_{k=1}^{T} F_{i,k} - Exp(F_{i,k}, Age_{i,k})}{T} \tag{5}$$

F. Age-based Percentile

This proposed CTF is also based on the measurement taken from subjects with the same age as the current subject ($Age_{i,T}$). However, instead of choosing one expected value, we rank all values of the current conceptual feature from all subjects with $age = Age_{i,T}$, and compute in what Percentile the current subject's last measurement for the conceptual feature is. We consider the last measurement of the feature, as it is the most relevant.

This CTF was inspired by the percentile feature used in [10], but that work did not use any other variable to compute percentiles and did not use longitudinal data. By contrast, we compute age-based percentiles and adapt DiffAgeMean's calculation to cope with a feature's multiple measurements across time points in longitudinal datasets. Thus, the Age-based Percentile CTF indicates what percentage of the other subjects with the same age as the current subject had measurements with lower values than the current subject's measurement.

For example, the Percentile value of 30% means that only about 30% of the subjects of the same age as the current subject have feature values lower than the current subject's value. The temporal aspect of the Percentile CTF is the calculation of the Ranks, which happens over all different measurements of the current feature.

The calculation of the Percentile CTF is shown in Equation 6. Note that the term $Age_{i,T}$ in this equation is indexed by T because we compute the rank of a subject's feature value at wave T. However, when computing the rank, we consider any measurement from subjects of that age, regardless of the wave. In Equation 6, $NValues(F_{i,T}, Age_{i,T})$ is the number of values used to compute the ranking. This CTF is calculated for numeric and ordered nominal features in the same way.

$$Percentile(F_i) = \frac{Rank(F_{i,T}, Age_{i,T})}{NValues(F_{i,T}, Age_{i,T})} \tag{6}$$

III. Methodology

A. The Elsa-nurse and Elsa-core Datasets

The English Longitudinal Study of Ageing (ELSA) is one of the most prominent populational studies of ageing [2], [11]. The ELSA has thousands of respondents from inhabitants of United Kingdom households, which take part in a core interview every two years, answering questions about various aspects of their lives. In addition, 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.

For our study, we used the datasets from the ELSA-core questionnaires for waves 1-8 (2002-2016) and ELSA-nurse questionnaires for waves 2, 4, 6 and 8 (2004-2016). We created one ELSA-core and one ELSA-nurse dataset for each of the 10 age-related diseases we are interested in predicting, totalling 20 longitudinal datasets.

The class variable in each dataset refers to the presence (negative class) or absence (positive class) of a reported diagnose for one of the 10 age-related diseases, for each instance (ELSA participant), in ELSA's 8th (most recently published) wave. For all 10 diseases, the positive class is the majority, with an increased class imbalance for rarer diseases, such as Dementia and Parkinson's Disease. The class labels were created from specific questions in the ELSA-core questionnaire about the diagnosis of each target age-related disease. For more information on the creation of the class variables, please see [5]. It is important to highlight that the ELSA and TILDA participants themselves are reporting the diagnosis of the target diseases in the 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 (false negatives), and that some patients wrongly reported their positive diagnosis (false positives).

The proposed Constructed Temporal Features (CTFs) represent temporal information that is generated by measuring the same conceptual feature over several consecutive time points. Naturally, only features that were measured more than once over the course of the study are valid candidates for CTF creation, which excludes demographics such as gender.

For the ELSA-core datasets, we used predictive features from waves 1-7, excluding wave 8 so that the class variable would be in the future with respect to the features. For the ELSA-nurse datasets, as we have only 4 waves, we do include wave 8's features although the classes are from that same wave. The 10 datasets have different class variables (representing different age-related diseases), but all 10 ELSA-nurse datasets have the same set of features, as do all 10 ELSA-core datasets. Details of the composition of these datasets are shown in Table I.

TABLE I
THE ELSA-NURSE AND ELSA-CORE DATASETS.

Data Source	Classes (diseases)	Instances	Feature Waves	Predictive Features	Numeric CTFs	Nominal CTFs
ELSA-nurse	10	7097	4	140	28	13
ELSA-core	10	8405	7	171	5	22

B. Preprocessing and experimental setup

For our experiments with the proposed CTFs, we created classification models using the Random Forest (RF) algorithm [12]. This work is the first to test CTFs for longitudinal data using the RF algorithm, which is among the state-of-the-art classification algorithms [13]. RFs handle well datasets with a high ratio of features to instances, which are prone to overfitting [14]. This is desirable as our proposal can add up to 6 CTFs for each conceptual feature in the dataset.

Because of the class imbalance problem mentioned earlier, training sets were balanced using the Balanced Random Forest (BRF) method [15]. 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 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 [16], [17], including a study that used datasets similar to the Elsa-nurse datasets used in our experiments [5].

The RFs were trained and tested using the Weka toolkit[1], with the default parameters $ntrees = 100$ (number of trees) and $mtry = \lfloor log_2(d) \rfloor + 1 = 8$ (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.

The RF classifiers were evaluated using three metrics: Sensitivity (True Positive Rate), Specificity (True Negative Rate) and GMean (geometric mean between Sensitivity and Specificity). These metrics were chosen partially based on [18, Chapter 4], who claim that for imbalanced biomedical data, models should be evaluated using metrics that consider their ability to predict each class separately (i.e., Sensitivity and Specificity) and at least one "global" metric of performance over both classes. We chose the GMean as a global measure, because it assigns equal importance to the prediction of both minority-class and majority-class instances, unlike more common global performance measures such as Accuracy.

All datasets had their missing values replaced in a data preprocessing step. The experiments used the well-known 10-fold cross-validation procedure, and we compared the results of three feature sets for each metric using two statistical significance tests, as follows.

First, we applied the Friedman's test, a rank-based non-parametric version of ANOVA with repeated measures [19]. If this test indicated the results are significantly different, we then applied the Nemenyi post-hoc test, a pairwise non-parametric test to determine whether or not different pairs of models have equivalent performance. Both tests were applied with the usual significance level $\alpha = 0.05$.

IV. Results and Discussion

A. Controlled Experiments with Baseline Datasets

For these experiments, our objective was to evaluate the potential increase in predictive accuracy for RF classifiers learned from baseline datasets containing the proposed CTFs as added features. First, we identify all conceptual features that can be used in the creation of the CTFs, called the set of ‘‘eligible” features. This set consists of all conceptual features that had at least two measurements (across waves), as we only calculated the CTFs for those. To have a fair evaluation of the proposed CTFs in a controlled experiment, we created datasets from the Elsa-nurse and Elsa-core databases with the following feature sets:

- **Baseline**: All measurements of the eligible conceptual features in the original dataset; no CTF.
- **CTFs-only**: The six proposed CTFs (AvgDiffAgeMean, DiffAgeMean, Diff, Monotonicity, Age-based Percentile, Ratio), created for each eligible conceptual feature; no original feature.
- **Baseline+CTFs**: Both the above feature sets combined, i.e. both original and CTF features.

All result Tables shown in this section and in section IV-B have the same structure: each column corresponds to the feature set that composes the dataset (Baseline, CTFs only, and both the Baseline and the CTFs combined), and each row shows the results for one class and data source (EN for Elsa-nurse datasets and EC for Elsa-core datasets), with the best result in boldface. In the last three rows, we show the number of wins for each feature set (number of datasets where the feature set got the best results) over the 10 Elsa-nurse datasets, over the 10 Elsa-core datasets, and over all 20 datasets, respectively. When ties happened, one 'win' was divided among the tied feature sets.

Tables II and III show the Sensitivity and Specificity values for all datasets. For both measures, the best results were clearly obtained by the Baseline+CTFs feature set, which obtained in total 11.33/20 wins for Sensitivity and 10/20 wins for Specificity. The CTFs-only feature set was overall the second best method, and it performed particularly well for Specificity, where it obtained in total 7/20 wins, against only 3/20 wins of the baseline feature set.

[1] Available at: https://www.cs.waikato.ac.nz/ml/weka/

TABLE II

Sensitivity results. Baseline datasets are comprised only of the features used to create the CTFs.

	Baseline	CTFs-only	BL+CTFs
EN_Arthritis	0.669	0.655	**0.670**
EN_HBP	**0.653**	0.647	0.650
EN_Cataract	**0.615**	0.576	0.601
EN_Diabetes	0.843	0.841	**0.846**
EN_Osteoporosis	**0.655**	0.629	0.643
EN_Stroke	0.667	0.679	**0.678**
EN_Heart Attack	**0.698**	**0.698**	**0.698**
EN_Angina	0.680	0.672	**0.683**
EN_Dementia	0.737	0.729	**0.752**
EN_Parkinson's	0.604	0.630	**0.634**
EC_Arthritis	0.741	**0.752**	0.750
EC_HBP	0.625	**0.640**	0.634
EC_Cataract	0.601	**0.626**	0.617
EC_Diabetes	0.671	**0.691**	0.674
EC_Osteoporosis	0.690	0.680	**0.691**
EC_Stroke	0.689	0.685	**0.692**
EC_Heart Attack	0.673	0.654	**0.669**
EC_Angina	**0.710**	0.691	0.706
EC_Dementia	0.757	0.771	**0.773**
EC_Parkinson's	0.685	0.714	**0.715**
Nwins EN datasets	3.33	0.33	**6.33**
Nwins EC datasets	1	4	**5**
Total Nwins	4.33	4.33	**11.33**

TABLE III

Specificity results. Baseline datasets are comprised only of the features used to create the CTFs.

	Baseline	CTFs-only	BL+CTFs
EN_Arthritis	0.594	**0.604**	0.593
EN_HBP	0.747	0.730	**0.751**
EN_Cataract	0.723	**0.754**	0.729
EN_Diabetes	0.865	**0.870**	0.863
EN_Osteoporosis	0.699	0.716	**0.717**
EN_Stroke	**0.710**	0.693	0.701
EN_Heart Attack	**0.731**	0.696	0.718
EN_Angina	0.678	0.709	**0.713**
EN_Dementia	**0.716**	0.709	0.703
EN_Parkinson's	0.636	0.591	**0.652**
EC_Arthritis	0.717	0.721	**0.726**
EC_HBP	0.662	**0.671**	0.669
EC_Cataract	0.675	0.750	**0.751**
EC_Diabetes	0.750	0.748	**0.759**
EC_Osteoporosis	0.635	0.661	**0.638**
EC_Stroke	0.694	**0.758**	0.747
EC_Heart Attack	0.689	**0.721**	0.705
EC_Angina	0.723	0.765	**0.772**
EC_Dementia	0.727	**0.832**	0.776
EC_Parkinson's	0.720	0.720	**0.747**
Nwins EN datasets	3	3	**4**
Nwins EC datasets	0	4	**6**
Total Nwins	3	7	**10**

TABLE IV

GMean results. Baseline datasets are comprised only of the features used to create the CTFs.

	Baseline	CTFs-only	BL+CTFs
EN_Arthritis	**0.630**	0.629	**0.630**
EN_HBP	**0.699**	0.688	**0.699**
EN_Cataract	**0.667**	0.659	0.662
EN_Diabetes	0.854	**0.855**	0.854
EN_Osteoporosis	0.676	0.671	**0.679**
EN_Stroke	0.688	0.686	**0.689**
EN_Heart Attack	**0.714**	0.697	0.708
EN_Angina	0.679	0.690	**0.698**
EN_Dementia	0.726	0.719	**0.727**
EN_Parkinson's	0.620	0.610	**0.643**
EC_Arthritis	0.729	0.737	**0.738**
EC_HBP	0.643	**0.656**	0.651
EC_Cataract	0.637	**0.685**	0.681
EC_Diabetes	0.709	**0.719**	0.715
EC_Osteoporosis	0.662	**0.670**	0.664
EC_Stroke	0.692	**0.721**	0.719
EC_Heart Attack	0.681	**0.687**	**0.687**
EC_Angina	0.716	0.727	**0.738**
EC_Dementia	0.742	**0.801**	0.774
EC_Parkinson's	0.702	0.717	**0.731**
Nwins EN datasets	3	1	**6**
Nwins EC datasets	0	**6.5**	3.5
Total Nwins	3	7.5	**9.5**

Table IV shows the GMean results. Again, for both these measures, the best overall result was obtained by Baseline+CTFs, which obtained in total 9.5/20 wins for GMean. The second best results were obtained by CTFs-only, with 7.5/20 wins. In general, the CTFs-only feature set performs better on the Elsa-core datasets than on the Elsa-nurse datasets, as shown in the second and third to last rows in the Table. In summary, the best result was obtained by the feature set containing both eligible original features and the proposed CTFs, for all three metrics.

Regarding the statistical tests, the Friedman test showed a significant difference between the results in two cases, so we performed the Nemenyi post-hoc test for those. For the Sensitivity metric (Friedman *p-value* $= 0.0363$), the Nemenyi test detected a significant difference between the BL+CTFs and the CTFs-only feature sets (Nemenyi *p-value* $= 0.0467$). For the GMean metric (Friedman *p-value* $= 0.0137$), there was a significant difference between the BL+CTFs and BL feature sets (Nemenyi *p-value* $= 0.0157$).

B. Full Dataset Results

In this second set of experiments, the new experiments in this current section compare the following three feature sets:

- **Baseline:** all original features with all available measurements (even features measured just once, which are not eligible for generating CTFs), no CTF;
- **CTFs+inel:** the six proposed types of CTF, created for each eligible conceptual feature, and the original features ineligible for generating CTFs;
- **Baseline+CTFs:** Both the above feature sets combined, i.e. all original and CTF features.

Tables V, VI and VII show the results for Sensitivity, Specificity and GMean for all datasets.

The best overall results were again obtained by Baseline+CTFs, although now its superiority is not so clear as in the previous Section. The main reason why the winner Baseline+CTFs has less impressive results in these new experiments seems to be because now all feature sets (including the Baseline) include features which are not used to construct CTFs but have good predictive power – e.g., the age and gender features are among the top-ranked features in the random forest models' feature importance measurements. Hence, the addition of the CTFs had a smaller impact on the resulting models for the BL+CTFs and CTFs+inel feature sets – by comparison with the experiments in the previous Section.

In this sections' experiments, the Friedman's test did not reject the null hypothesis for any of the metrics.

TABLE V
SENSITIVITY RESULTS. FULL ELSA-NURSE (EN) AND ELSA-CORE (EC) DATASETS AS THE BASELINE.

	Baseline	CTFs+inel	BL+CTFs
EN_Arthritis	**0.671**	0.658	**0.671**
EN_HBP	**0.651**	0.644	0.650
EN_Cataract	**0.620**	0.593	0.605
EN_Diabetes	0.841	0.836	**0.845**
EN_Osteoporosis	**0.649**	0.633	0.643
EN_Stroke	0.670	0.681	**0.674**
EN_Heart Attack	**0.700**	0.694	**0.700**
EN_Angina	**0.684**	0.673	0.681
EN_Dementia	0.729	**0.748**	0.743
EN_Parkinson's	0.628	**0.650**	0.627
EC_Arthritis	0.743	0.747	**0.756**
EC_HBP	0.621	**0.642**	0.637
EC_Cataract	0.612	**0.623**	0.613
EC_Diabetes	0.674	**0.686**	0.682
EC_Osteoporosis	0.692	0.687	**0.702**
EC_Stroke	0.676	0.683	**0.696**
EC_Heart Attack	0.674	0.669	**0.680**
EC_Angina	**0.710**	0.699	0.706
EC_Dementia	0.756	0.765	**0.766**
EC_Parkinson's	0.677	**0.714**	0.694
Nwins EN datasets	**5**	2	3
Nwins EC datasets	1	4	**5**
Total Nwins	6	6	**8**

TABLE VI
SPECIFICITY RESULTS. FULL ELSA-NURSE (EN) AND ELSA-CORE (EC) DATASETS AS THE BASELINE.

	Baseline	CTFs+inel	BL+CTFs
EN_Arthritis	0.586	**0.609**	0.594
EN_HBP	**0.749**	0.724	0.745
EN_Cataract	0.723	**0.751**	0.736
EN_Diabetes	0.866	0.863	**0.868**
EN_Osteoporosis	0.696	**0.723**	0.716
EN_Stroke	**0.724**	0.698	0.720
EN_Heart Attack	**0.738**	0.711	0.713
EN_Angina	**0.702**	**0.702**	0.686
EN_Dementia	0.709	0.696	**0.736**
EN_Parkinson's	0.712	0.652	**0.727**
EC_Arthritis	0.721	**0.722**	0.720
EC_HBP	**0.678**	0.677	0.673
EC_Cataract	**0.777**	0.760	0.764
EC_Diabetes	0.751	0.743	**0.758**
EC_Osteoporosis	0.670	**0.696**	0.666
EC_Stroke	**0.751**	**0.751**	0.747
EC_Heart Attack	**0.717**	0.714	**0.717**
EC_Angina	**0.765**	**0.765**	0.761
EC_Dementia	**0.845**	**0.845**	0.807
EC_Parkinson's	0.707	0.693	**0.733**
Nwins EN datasets	**3.5**	**3.5**	3
Nwins EC datasets	**4**	3.5	2.5
Total Nwins	**7.5**	7.0	5.5

TABLE VII
GMEAN RESULTS. FULL ELSA-NURSE (EN) AND ELSA-CORE (EC) DATASETS AS THE BASELINE.

	Baseline	CTFs+inel	BL+CTFs
EN_Arthritis	0.627	**0.633**	0.632
EN_HBP	**0.698**	0.683	0.696
EN_Cataract	**0.670**	0.667	0.667
EN_Diabetes	0.854	0.849	**0.856**
EN_Osteoporosis	0.672	0.677	**0.679**
EN_Stroke	**0.697**	0.690	**0.697**
EN_Heart Attack	**0.719**	0.702	0.707
EN_Angina	**0.693**	0.687	0.684
EN_Dementia	0.719	0.722	**0.740**
EN_Parkinson's	0.669	0.651	**0.675**
EC_Arthritis	0.732	0.734	**0.738**
EC_HBP	0.649	**0.659**	0.655
EC_Cataract	**0.690**	0.688	0.685
EC_Diabetes	0.712	0.714	**0.719**
EC_Osteoporosis	0.681	**0.692**	0.684
EC_Stroke	0.712	0.716	**0.721**
EC_Heart Attack	0.695	0.691	**0.698**
EC_Angina	**0.737**	0.731	0.733
EC_Dementia	0.799	**0.804**	0.787
EC_Parkinson's	0.691	0.703	**0.713**
Nwins EN datasets	**4.5**	1	**4.5**
Nwins EC datasets	2	3	**5**
Total Nwins	6.5	4	**9.5**

C. *Feature Importance Analysis*

To further evaluate the impact of adding CTFs to the baseline dataset, we used a feature importance metric to analyse how often the proposed CTFs were selected as the best features in the RF models. We used the feature importance metric implemented in the Weka data mining tool, which is based on the average class-impurity decrease over all nodes where the feature was selected.

We selected the top 10 features with highest average impurity decrease for the RF produced in each fold in the cross-validation process, totalling 100 top-ranking features for each dataset. The selection was done over our RF runs using the BL+CTFs feature set defined in Section IV-B. Table VIII shows how many times original (baseline) and constructed features were selected for each dataset. For each dataset source (EN for ELSA-nurse and EC for ELSA-core) the datasets are shown in increasing order of class imbalance.

TABLE VIII
FEATURE IMPORTANCE: NUMBER OF FEATURES OF EACH CATEGORY IN THE TOP 100 FEATURES (CONSIDERING THE 10 TOP FEATURES IN EACH OF THE 10 CROSS-VALIDATION FOLDS), FOR EACH DATASET.

Dataset	Original Features Selected	Constructed Features Selected
EN_Arthritis	96	4
EN_HBP	98	2
EN_Cataract	93	7
EN_Diabetes	82	18
EN_Osteoporosis	90	10
EN_Stroke	88	12
EN_HeartAttack	72	28
EN_Angina	74	26
EN_Dementia	64	36
EN_Parkinson's	55	45
EC_Arthritis	2	98
EC_HBP	2	98
EC_Cataract	5	95
EC_Diabetes	16	84
EC_Osteoporosis	14	86
EC_Stroke	25	75
EC_HeartAttack	22	78
EC_Angina	29	71
EC_Dementia	33	67
EC_Parkinson's	36	64
Total	**996**	**1004**

The Table shows an interesting trend of fewer CTFs being selected for the ELSA-nurse datasets, with more CTFs being selected as the class imbalance ratio increases, and the opposite happens for the ELSA-core datasets. Overall, 18.8% and

and 81.6% of the best ranked features were CTFs, for the ELSA-nurse and ELSA-core datasets, respectively.

We did not show the specific numbers for each CTF due to space constraints, however it is important to note that the type of CTF selected most often was Percentile (41.2%), followed by Monotonicity (22.3%) and DiffAgeMean (16.2%). Both Percentile and DiffAgeMean focus on the most recent measurement of a feature, and compare it to the measurements of other individuals of the same age of the respondent. Monotonicity aims to identify upwards or downwards trends in the values of a feature over all measurements. Note that these temporal trends would be ignored by the classification algorithm applied to the original dataset, so adding the proposed CTFs to the dataset in a preprocessing phase is an effective and computationally non-expensive approach.

V. Conclusion

We have proposed several new types of Constructed Temporal Features (CTFs) and investigated whether adding CTFs to longitudinal datasets increases predictive accuracy. In our experiments, we used 20 real-world datasets created from the ELSA. To assess the effect of adding the proposed CTFs to longitudinal datasets, we ran two sets of experiments.

First, we ran a controlled experiment to measure the impact of the CTFs in predictive accuracy. These experiments compared three different feature sets: (a) a baseline set with only the original features used for constructing CTFs, (b) the proposed CTFs only (no original features), and (c) an extended feature set with both feature sets (a) and (b). The results were a very clear increase in the Sensitivity, Specificity, and GMean metrics for the third approach.

In the second set of experiments, we included in all three feature sets the original ELSA features that were ineligible for CTF creation. These include highly predictive features such as age and gender, which improved the learned RF models. In these experiments, the trend towards better predictive accuracy in the BL+CTFs feature set persisted, although not as strong.

The Percentile, Monotonicity and DiffAgeMean CTFs were the most commonly selected types of CTF, totalling about 40% of the best-ranked features overall. Percentile and DiffAgeMean are new contributions of this work, whilst Monotonicity was proposed in [5] for numerical features only, whilst in this work they were also extended to ordered nominal features.

Future work could involve adding the proposed CTFs to different longitudinal datasets, as well as proposing other types of CTFs and variations of existing CTFs. In addition, we would like to ask healthcare professionals to analyse our proposed CTFs and give feedback about their clinical validity and interest, possibly proposing new features based on what type of temporal pattern would be considered relevant for clinical or healthcare research purposes.

Acknowledgment

The English Longitudinal Study of Ageing was developed by researchers based at the University College London, NatCen Social Research, and the Institute for Fiscal Studies. The data are linked to the UK Data Archive and freely available at https://discover.ukdataservice.ac.uk.

References

[1] C. Ribeiro and A. A. Freitas, "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*, 2019, 5 pages.

[2] J. Banks, G. Batty, K. Coughlin, and et al., "English longitudinal study of ageing: Waves 0–8, 1998–2017.[data collection]," 2019.

[3] U. Niemann, T. Hielscher, M. Spiliopoulou, H. Völzke, and J.-P. Kühn, "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*. IEEE, 2015, pp. 121–126.

[4] G. Buizza, I. Toma-Dasu, M. Lazzeroni, C. Paganelli, M. Riboldi, Y. Chang, Ö. Smedby, and C. Wang, "Early tumor response prediction for lung cancer patients using novel longitudinal pattern features from sequential pet/ct image scans," *Physica Medica*, vol. 54, pp. 21–29, 2018.

[5] T. Pomsuwan and A. A. Freitas, "Feature selection for the classification of longitudinal human ageing data," in *Proc. IEEE International Conference on Data Mining Workshops (1st Workshop on Data Mining for Aging, Rehabilitation and Independent Assisted Living (ARIAL))*. IEEE, 2017, pp. 739–746.

[6] M. A. Morid, O. R. L. Sheng, G. Del Fiol, J. C. Facelli, B. E. Bray, and S. Abdelrahman, "Temporal pattern detection to predict adverse events in critical care: Case study with acute kidney injury," *JMIR medical informatics*, vol. 8, no. 3, e14272, 2020.

[7] M. Aghili, S. Tabarestani, M. Adjouadi, and E. Adeli, "Predictive modeling of longitudinal data for alzheimer's disease diagnosis using rnns," in *International Workshop on PRedictive Intelligence In MEdicine*. Springer, 2018, pp. 112–119.

[8] J. Luo, M. Ye, C. Xiao, and F. Ma, "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*, 2020, pp. 647–656.

[9] C. E. Ribeiro and A. A. Freitas, "A data-driven missing value imputation approach for longitudinal datasets," *Artificial Intelligence Review*, vol. 1, no. 31, 2021.

[10] R. Al-Otaibi, R. B. Prudêncio, M. Kull, and P. Flach, "Versatile decision trees for learning over multiple contexts," in *Proc. European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2015)*. Springer, 2015, pp. 184–199.

[11] J. Abell, N. Amin-Smith, J. Banks, and et al., *The Dynamics of ageing: Evidence from the English Longitudinal Study of Ageing 2002-2016 (Wave 8)*. London: Institute for Fiscal Studies, 2018. [Online]. Available: https://www.ifs.org.uk/publications/13510

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[14] E. Scornet, G. Biau, J.-P. Vert *et al.*, "Consistency of random forests," *The Annals of Statistics*, vol. 43, no. 4, pp. 1716–1741, 2015.

[15] C. Chen, A. Liaw, L. Breiman *et al.*, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, no. 1-12, 24, 2004.

[16] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Information Sciences*, vol. 250, pp. 113–141, 2013.

[17] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *Journal of Artificial Intelligence Research*, vol. 19, pp. 315–354, 2003.

[18] J. D. Malley, K. G. Malley, and S. Pajevic, *Statistical learning for biomedical data*. Cambridge University Press, 2011.

[19] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.