A New Longitudinal Classification Method Based on Stacking Predictions for Separate Time Points

Simon Provost and Alex A. Freitas

University of Kent, Canterbury, United Kingdom
simon.gilbert.provost@gmail.com, a.a.freitas@kent.ac.uk

Abstract. Biomedical research often uses longitudinal data with repeated measurements of variables across time (e.g. cholesterol measured across time), which is challenging for standard machine learning algorithms due to intrinsic temporal dependencies. The Separate Waves (SepWav) data-transformation method trains a base classifier for each time point ("wave") and aggregates their predictions via voting. However, the simplicity of the voting mechanism may not be enough to capture complex patterns of time-dependent interactions involving the base classifiers' predictions. Hence, we propose a novel SepWav method where the simple voting mechanism is replaced by a stacking-based meta-classifier that integrates the base classifiers' wave-specific predictions into a final predicted class label, aiming at improving predictive performance. Experiments with 20 datasets of ageing-related diseases have shown that, overall, the proposed Stacking-based SepWav method achieved significantly better predictive performance than two other methods for longitudinal classification in most cases, when using class-weight adjustment as a class-balancing method.

Keywords: Longitudinal Data · Classification · Supervised Machine Learning · Age-Related Diseases · Data Transformation

1 Introduction

Longitudinal data track the same subjects over time, capturing their characteristics (features) across multiple time points, also called "waves" [3, 10, 16]. This type of data is common in biomedical research; e.g., blood pressure and cholesterol are recorded over time for various diseases and conditions [19]. Traditional supervised machine learning techniques struggle with longitudinal data due to temporal dependencies, irregular sampling, and diverse patient histories [1, 21].

There are two approaches for longitudinal data classification [16], viz. data transformation and algorithm adaptation. Data transformation converts longitudinal data into a static format, making it suitable for a wide range of standard classification algorithms. However, this approach often ignores temporal relationships. Algorithm adaptation adapts classification algorithms to account for temporal patterns in the data. Nevertheless, this approach often requires complex designs and tools, making it less accessible than data-transformation methods.

This work focuses on the data-transformation approach, specifically the Separate Waves ("SepWav") method [16]. In the context of longitudinal data, where features are recorded at various time points (waves), SepWav learns a classifier

for each wave-specific feature in longitudinal data. For example, one classifier receives first-wave blood-test data, while another receives second-wave blood-test data. The classifiers' predictions are then combined, e.g. using a majority voting mechanism, selecting the class label with the most votes from all wave-specific classifiers [6]. Although SepWav preserves wave-specific patterns compared to other transformation methods (see section 2), the simple nature of its voting mechanism may fail to leverage patterns learnt across waves.

To tackle this limitation, we propose a novel SepWav method where the simple voting mechanism is replaced by a more sophisticated stacking-based meta-classifier that integrates wave-specific class-probability predictions into a final predicted class label, aiming at improving predictive performance.

The proposed stacking-based SepWav method was evaluated in experiments with 20 datasets of ageing-related diseases from the English Longitudinal Study of Ageing (ELSA) [4]. This is an important type of biomedical dataset given global healthcare pressures from ageing populations [24].

The experiments compare our new stacking-based SepWav method against two other methods: (a) a baseline random forest without using SepWav and (b) standard SepWav with majority voting. The experiments also consider two class-balancing scenarios: (a) without using any class-balancing method, i.e., using the originally imbalanced class distributions of each dataset; and (b) using class-weight adjustment as a class-balancing method. The results show that the combination of the proposed SepWav method and class-weight adjustment produced, overall, the best predictive performance results.

The remainder of this paper is structured as follows. Section 2 discusses the background on longitudinal classification, focusing on data transformation methods. Our proposed stacking-based SepWav method is described in Section 3. In Section 4, the experimental setup is presented, followed by results and discussions in Section 5. Section 6 concludes and discusses future works.

2 Background on Longitudinal Classification

Longitudinal classification is a variant of the classification task of machine learning where some features take values at multiple time points ("waves") [3,16]. The goal is to learn a model for predicting the class label (Y) of each instance, considering the temporal evolution of feature values across the waves. More precisely, the goal is to learn a predictive model (classifier function) of the form:

$$Y \leftarrow f(X_{1,1}, X_{1,2}, \ldots, X_{1,T}, \ldots, X_{D,1}, X_{D,2}, \ldots, X_{D,T})$$

where $X_{i,j}$, $i = 1, \ldots, D$, $j = 1, \ldots, T$, is the value of the i-th feature at the j-th wave (time point) for the current instance, D is the number of features (the data's dimensionality), and T is the number of waves. Standard machine learning algorithms, which are in general designed for coping with static (no temporal dimension) data, falter when confronted with the temporal dependencies inherent in longitudinal data [1].

Note that in this work we address a type of longitudinal classification task where the features are longitudinal (measured across multiple time points), but

the class variable is static (measured at a single time point). There are other variations of longitudinal classification where the class variable is also longitudinal [9], but such classification tasks are out of the scope of this paper.

As mentioned earlier, this work focuses on the data-transformation approach, which converts longitudinal data into a data format suited for standard classification algorithms. This approach has the advantage that any standard classification algorithm can be applied to the transformed data, but the disadvantage that the data transformation method employed involves some loss of relevant temporal information, by comparison with the original longitudinal data.

One common method based on this approach is **Merge Waves Minus Time Indices** (MerWav-Time(-)) [16], called *Longitudinal Features Non Sequential* in [3]. MerWav-Time(-) "flattens" features across all waves into a single feature set, so that the value of each feature at each wave is considered a distinct feature, losing the temporal information associated with the time indices (wave ids). For example, the values $X_{i,1}$ and $X_{i,2}$ of a longitudinal feature X_i at waves 1 and 2 are treated as two independent features, ignoring the fact that they are temporal variations of the same longitudinal feature. In other words, while enabling the use of standard algorithms, this method discards temporal ordering, risking bias by treating correlated measurements as completely independent. This simplicity makes it a common baseline method in the literature, but its limitations motivate more sophisticated methods for longitudinal classification. The MerWav-Time(-) method has been used, e.g., in [8, 20, 25].

This work focuses on the **Separate Waves** (SepWav) method [16], also based on the data-transformation approach. Figure 1 illustrates the SepWav method for longitudinal classification. It trains a classifier C_j ($j = 1, \ldots, T$) using each wave j's specific features ($j = 1, \ldots, T$), where T denotes the number of waves and the number of classifiers. Each classifier is trained independently from the others using the feature set $X_{*,j}$, where the subscript $*, j$ represents all features at wave j. All classifiers predict the same class variable Y — in this work, the presence or absence of a disease in the last wave. Finally, as shown in Figure 1's box (A), these wave-specific predictions are aggregated, usually through majority voting or weighted voting. In this work, where this standard SepWav method is used as a baseline method in our experiments, majority voting is used to aggregate the binary class predictions: $P = \text{mode}(C_1(X_{*,1}), C_2(X_{*,2}), \ldots, C_N(X_{*,T}))$, where $C_j(X_{*,j})$ is the class label predicted by the classifier for wave j ($j = 1, \ldots, T$). This SepWav method has been used in [7, 22, 23].

Unlike MerWav-Time(-), SepWav maintains the integrity of wave-specific patterns by refraining from merging the features across waves, thereby reducing the bias that could arise from considering related measurements as independent.

Note that, since each base classifier in a SepWav method is trained with features from just one wave (time point), temporal patterns are considered only when integrating the predictions of the base classifiers. In a standard SepWav method this integration is performed by a simple voting mechanism, which may not capture well more complex temporal patterns involved in the aggregation of the base classifiers' predictions.

This shortcoming of the standard SepWav method motivates our proposal of a stacking-based mechanism (based on a meta-classifier) to integrate the predictions of the wave-specific base classifiers, as detailed in Section 3.

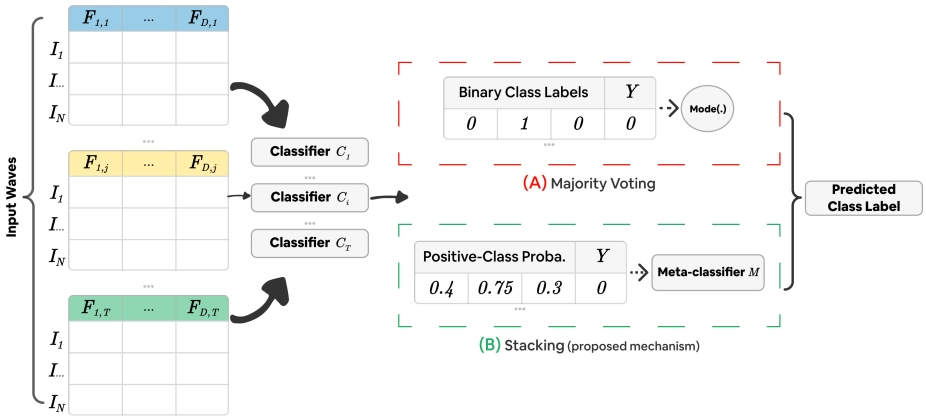

Fig. 1: Overview of the Separate Waves ("SepWav") method for longitudinal data classification. At the left end of the figure, the longitudinal dataset is decomposed into T datasets, one per wave (time point), each dataset with N instances and D features. In the middle, a separate classifier is trained for each wave. At the right end of the figure, wave-specific classifier predictions can be aggregated using one of two mechanisms: (A) Majority Voting, the usual mechanism in the literature (Section 2), or (B) Stacking, the proposed mechanism (Section 3).

3 The Proposed Stacking-Based Separate Waves Method for Longitudinal Classification

This work proposes a novel stacking-based Separate Waves (SepWav) method for longitudinal classification.

Recall that stacking is a meta-learning mechanism where first many learners are trained independently of one another, producing base predictive models (classifiers), and then a "meta-learner" is trained to produce a meta-classifier based on the outputs of the base classifiers [13], thereby improving the capacity to model intricate patterns within the data.

The basic idea of our proposed SepWav method consists of using a more sophisticated mechanism for aggregating the outputs of the base classifiers (one for each wave), based on stacking, instead of simply using majority voting for that aggregation as in the standard SepWav method. The motivation for the proposed stacking-based method is that, due to learning a meta-classifier from all the base classifiers' class-probability predictions, stacking should capture temporal dynamics and base-classifier interactions better than simpler voting methods.

Figure 1 shows the difference between a standard SepWav method (Section 2) and the proposed stacking-based SepWav method. Note that both types of SepWav methods learn a separate base classifier from each wave's feature set,

but they differ in how they aggregate those classifiers' predictions. The box (A) in the figure shows that a standard SepWav computes the majority vote among the class labels predicted by the base classifiers, whilst the box (B) shows that the proposed SepWav method uses a stacking-based mechanism, where a meta-classifier is trained to predict an instance's class based on meta-features representing the positive-class probabilities output by the base classifiers.

More precisely, the proposed SepWav method works as follows. For a dataset with T waves, we train T base classifiers (trained independently from each other), namely a classifier C_j for each wave j, $j = 1, \ldots, T$, where each C_j is trained with a subset of the training data containing all features at wave j and the class variable Y. Note that, although the classifiers are trained with different feature sets (from different waves), all classifiers are trained to predict the same class variable, which usually is a variable measured in the last (most recent) wave, as is the case in this work. Then, the values of the probability of the positive class predicted by those T classifiers, denoted V_j ($j = 1, \ldots, T$), become meta-features of the meta-training set that will be fed into the meta-classifier:

$$\mathbf{V} = [V_1, V_2, \ldots, V_T] \tag{1}$$

A meta-classifier M is then trained on $(\mathbf{V}, Y)$ to produce the final predicted class. Hence, M computes the predicted class of each new (test) instance x as:

$$P(x) = M(\mathbf{V}(x)) \tag{2}$$

This mechanism allows M weighing and integrating the base classifiers' predicted probabilities of the positive class, based on their relative relevance and their interactions, potentially detecting patterns missed by simpler voting schemes.

We use Logistic Regression [11] as the meta-classifier in our experiments, as it works well with the small number of meta-features (one per wave) for our datasets (which have 4 or 7 waves, as detailed in Section 4). Other meta-classifiers could be used in future research, particularly for datasets with much larger numbers of waves (leading to much larger numbers of meta-features).

We are aware of only one work using SepWav and stacking for longitudinal classification [22], where the class variable is longitudinal and the meta-learner is a longitudinal classification algorithm (LSTM). By contrast, in our work the class variable is static (see Section 2), and so there is no need to use a longitudinal classification algorithm as the meta-learner.

4 Experimental Setup

This section describes the datasets, the methods, and the evaluation methodology used in the experiments, and how imbalanced classes were addressed.

4.1 The Datasets Used in the Experiments

We used longitudinal datasets from [17], based on the popular English Longitudinal Study of Ageing (ELSA) database [4]. The ELSA database tracks core participants, who are 50 years of age or older and reside in the United Kingdom, through repeated interviews. In our experiments we used two types of ELSA

datasets, which we refer to as: (a) ELSA-nurse, with biomedical data collected every four years by a nurse or health professional; and (b) ELSA-core, with data from core interviews conducted every two years. These datasets derive from the work in [17], which created 20 longitudinal datasets, each combining a set of core or nurse data (features) with one of 10 age-related diseases as the class variable.

The 10 ELSA-nurse datasets contain 7,096 instances and 140 biomedical features from waves 2, 4, 6, and 8 of the ELSA study, as well as a class variable from wave 8. The ELSA-core datasets contain 8,405 instances and 171 features, nearly all from ELSA waves 1–7 (except age, from wave 8), and a class variable from wave 8. Note that, for each type of ELSA data (nurse or core), in general all 10 datasets have the same set of features (with a minor exception for diabetes-nurse data, see below), but the 10 datasets have different binary class variables, each representing the presence or absence of a different age-related disease. Missing values were imputed in a data preprocessing phase. The process of creating these datasets is described in detail in [17].

We made a simple modification in just one of those 20 datasets, the *diabetes* dataset with *Nurse* data, where we removed a single longitudinal feature: "HbA1c", since this feature is directly used to clinically decide whether a patient has diabetes (i.e., it is not fair to use this feature to predict diabetes [18]).

Table 1 shows the class distribution of each of the 20 datasets (2 datasets per row) used in the experiments. The second and third columns show the percentage of positive instances (individuals with the disease) for the Nurse and Core datasets. The final two columns show the class imbalance ratio — the ratio of the number of instances in the majority (negative) class divided by the number of instances in the minority (positive) class — for the Nurse and Core datasets.

Table 1: Class distribution in each dataset

Disease	**Posit. class % (Nurse)**	**Posit. class % (Core)**	**Class imbalance ratio (Nurse)**	**Class imbalance ratio (Core)**
Arthritis	42.57%	39.65%	1.35	1.52
Hbp	40.21%	38.72%	1.49	1.58
Cataract	32.72%	29.60%	2.06	2.38
Diabetes	13.33%	12.83%	6.50	6.80
Osteoporosis (Osteop.)	9.22%	8.45%	9.85	10.84
Stroke	5.93%	5.45%	15.86	17.35
Heart attack (Heart Att.)	5.65%	5.25%	16.70	18.06
Angina	3.64%	3.39%	26.51	28.49
Dementia	2.09%	1.92%	46.95	51.20
Parkinsons (Parkin.)	0.93%	0.89%	106.53	111.07

4.2 The Methods Evaluated in the Experiments

In our experiments, we compare three methods within the data-transformation approach for longitudinal classification. All methods use Random Forest as a base classification algorithm for a fair comparison of the different methods' results.

The baseline method, which is the most popular and simplest method to apply standard (non-longitudinal) classification algorithms to longitudinal data, is the **MerWav-Time(-) method**. As described in Section 2, this method "flattens" the sets of features from all waves into a single feature set, ignoring the time indices of the features. I.e., the values of a feature at different time points (waves) are simply treated as completely different features so that the temporal relationship between different values of the same longitudinal feature is no longer explicitly represented in the data. Then, a standard classifier (in this work, Random Forest) is trained on the flattened data. Hereafter we refer to this method as **Standard Random Forest (Std-RF)**.

The other two methods use the SepWav method, training a separate Random Forest classifier on each wave but differing in how they combine the outputs of the base classifiers. **SepWav with Majority Voting (SW-RF-MV)** aggregates the base RF classifiers' predicted class labels via majority voting.

Finally, the proposed **SepWav enhanced by Stacking with Logistic Regression (SW-RF-SLR)** uses stacking with Logistic Regression as the meta-classifier to integrate the probabilities of the positive class predicted by the base RF classifiers (see Section 3 for details).

The Random Forest algorithm, implemented in the Scikit-Learn library [14] with default hyperparameters, served as the main classifier in Std-RF and as the base classifiers in the two SepWav methods. The SepWav method is implemented in Scikit-Longitudinal, a recently designed Scikit-Learn-like library for longitudinal machine learning [15]; please see the Code Availability section.

4.3 Evaluation Methodology

The predictive accuracy of each method is assessed by two metrics: (a) the well-known Area Under the Receiver Operating Characteristic Curve (AUROC) [2]; and (b) the Geometric Mean (GMean) of sensitivity and specificity [12], where sensitivity is the proportion of positive-class instances correctly predicted as positives and specificity is the proportion of negative-class instances correctly predicted as negatives.

To estimate generalisation performance, we use a well-known 10-fold cross-validation procedure. I.e., the dataset is randomly divided into 10 equal-sized folds, and in each iteration, one fold serves as the test set, while the remaining 9 folds are merged and used as the training set. This process is repeated 10 times, with each fold used as the test set once. The reported values of the AUROC and GMean metrics are computed as their average values across the 10 test sets.

For each table of results reported in Section 5, the statistical significance of the results is analysed in two steps. First, the Friedman test [5] (which has the advantage of being non-parametric) is used to assess whether or not there is a statistically significant difference among the results of all compared methods as a

whole, based on the standard significance level $\alpha = 0.05$. Second, if the Friedman test's p-value was significant ($p < 0.05$), the Holm's post-hoc test [5] is used to compare the control method against each other method—the control method is the best method (based on the current metric). For each pairwise comparison (control vs. another method), the control method is deemed significantly better than another method if the Holm's test p-value is smaller than the corresponding adjusted alpha—see [5] for details of the calculation of adjusted alpha values, which incorporates a correction for multiple hypothesis testing.

4.4 Coping with Class Imbalance

Given the class imbalance in our datasets (see Table 1), we evaluated each method in two scenarios. In the imbalanced-classes scenario, no class-balancing method was applied, i.e., all models were trained on the original imbalanced datasets, as a baseline scenario. In the class-weight adjustment scenario, before training the classifiers, we adjusted instance weights to balance the influence of minority and majority classes using a "balanced" weighting scheme, as used in [11]. The weight w_j for each instance in class j was calculated as $w_j = \frac{n}{k \cdot n_j}$, where n is the number of training instances, k is the number of classes (2 in our datasets), and n_j is the number of instances in class j. These instance weights were applied only to the training set, leaving the test set with its original imbalanced class distribution to reflect real-world conditions.

We also performed preliminary experiments with the well-known random undersampling and SMOTE class-balancing methods, but overall their results were not very good, so they are not reported here.

5 Computational Results and Discussion

This section presents results on the predictive accuracy of the three methods mentioned in Subsection 4.2: Std-RF, SW-RF-MV, and SW-RF-SLR; using the GMean and AUROC metrics across the 20 datasets described in Section 4.1.

Subsection 5.1 compares the methods in the imbalanced-class (baseline) scenario, whilst Subsection 5.2 compares the methods in the class-weight adjustment (balanced-class) scenario. Finally, Subsection 5.3 compares all 6 combinations of 3 methods times 2 class-balancing scenarios.

5.1 Comparing Methods in the Original Imbalanced-Classes Scenario

Tables 2 and 3 present the GMean and AUROC results for each method in the scenario with imbalanced classes as in the original dataset, i.e., without applying any class-balancing technique. In these tables, the best-performing method for each dataset (in each row) is highlighted with boldface font. In addition, the row after the results for all datasets provides the average rank of each method across datasets. The methods' average ranks are calculated as follows. First, for each dataset, the best method (based on the current metric) is assigned rank 1, and the worst method is assigned rank 3. If there is a tie between two or

more methods, the ranks are divided proportionally among the tied methods. E.g., if two methods are tied as the best method, each gets rank 1.5. Then, for each method, its average rank is averaged over the 20 datasets. Finally, the last row in Tables 2 and 3 shows the p-value obtained by the Friedman test when comparing the results of all methods as a whole (see Subsection 4.3).

For the GMean metric (Table 2), Std-RF achieved the best (smallest) average rank (1.88), a little better than the average ranks for SW-RF-SLR (1.95) and SW-RF-MV (2.17). However, the Friedman test produced a p-value of 0.479, indicating that there are no statistically significant differences among the results of the 3 methods as a whole.

For the AUROC metric (Table 3), SW-RF-MV achieved the best average rank (1.65), substantially better than the average ranks for Std-RF (1.95) and SW-RF-SLR (2.40). However, the Friedman test's p-value (0.058) is slightly above the significance level threshold of 0.05, so again the differences among the results of the 3 methods do not reach statistical significance.

In summary, in the original (baseline) imbalanced-classes scenario, there were no statistically significant differences among the results of the 3 methods.

Table 2: GMean results for Std-RF, SW-RF-MV and SW-RF-SLR, without any class-balancing method

Dataset	Std-RF	SW-RF-MV	SW-RF-SLR
EC_Angina	0.0000	0.0000	0.0000
EC_Arthritis	**0.7204**	0.7002	0.7154
EC_Cataract	0.4646	**0.5369**	0.0000
EC_Dementia	0.1353	0.0000	**0.2596**
EC_Diabetes	0.1811	0.0096	**0.2736**
EC_HBP	**0.6021**	0.5744	0.5388
EC_HeartAtt.	0.0000	0.0000	0.0000
EC_Osteop.	0.0237	0.0000	**0.0475**
EC_Parkin.	0.0000	0.0000	0.0000
EC_Stroke	**0.0297**	0.0000	**0.0297**
EN_Angina	0.0000	0.0000	0.0000
EN_Arthritis	0.5903	0.7002	**0.7154**
EN_Cataract	0.4646	**0.5369**	0.3641
EN_Dementia	0.0000	0.0000	0.0000
EN_Diabetes	0.1811	**0.4640**	0.2736
EN_HBP	**0.6021**	0.5744	0.5388
EN_HeartAtt.	**0.0316**	**0.0316**	0.0000
EN_Osteop.	0.0237	0.0000	**0.0799**
EN_Parkin.	0.0000	0.0000	0.0000
EN_Stroke	**0.0297**	0.0000	**0.0297**
EC + EN Rank	**1.88**	2.17	1.95
Friedman Test	p-value = 0.479		

Table 3: AUROC results for Std-RF, SW-RF-MV and SW-RF-SLR, without any class-balancing method

Dataset	Std-RF	SW-RF-MV	SW-RF-SLR
EC_Angina	0.7542	**0.7811**	0.7504
EC_Arthritis	**0.8116**	0.8084	0.8101
EC_Cataract	0.7445	**0.7618**	0.7440
EC_Dementia	0.8544	**0.8650**	0.8402
EC_Diabetes	**0.7870**	0.7867	0.7834
EC_HBP	**0.7091**	0.7060	0.7073
EC_HeartAtt.	0.7399	**0.7659**	0.7382
EC_Osteop.	0.7507	**0.7734**	0.7554
EC_Parkin.	0.7136	**0.7598**	0.7442
EC_Stroke	0.7611	**0.7815**	0.7639
EN_Angina	**0.7542**	0.7429	0.7504
EN_Arthritis	0.6771	0.8084	**0.8101**
EN_Cataract	**0.7445**	0.7315	0.7201
EN_Dementia	0.7485	**0.7952**	0.7779
EN_Diabetes	0.7870	**0.8733**	0.7834
EN_HBP	**0.7091**	0.7060	0.7073
EN_HeartAtt.	0.7561	**0.7633**	0.7382
EN_Osteop.	0.7507	**0.7734**	0.7166
EN_Parkin.	0.5714	**0.7598**	0.3950
EN_Stroke	0.7611	0.7382	**0.7639**
EC + EN Rank	1.95	**1.65**	2.40
Friedman Test	p-value = 0.058		

5.2 Comparing Methods in the Class-Weighting Scenario

Tables 4 and 5 present the GMean and AUROC results for each method in the class-weight adjustment scenario, i.e., with class-balancing applied to the training set before running the classification algorithms (see Subsection 4.4).

For the GMean metric (Table 4), SW-RF-SLR obtained the ideal average rank of 1.00, i.e., it was the best method in all 20 datasets. The Friedman test indicates a significant difference among the methods' results (p-value < 0.001). Hence, we apply the Holm's post-hoc test, whose results are shown in the last

two rows of the table. SW-RF-SLR was significantly better than both Std-RF and SW-RF-MV, as the Holm's test produced $p < 0.001$ for each comparison, which is much less than the adjusted alpha of 0.050 and 0.025, respectively.

For the AUROC metric (Table 5), SW-RF-SLR obtained the best (lowest) average rank of 1.70, but the Friedman test's p-value of 0.074 indicates that there is no statistically significant difference among the 3 methods.

Hence, in the class-weight adjustment scenario, SW-RF-SLR was the best method in terms of both GMean and AUROC, with significant results for GMean.

Table 4: GMean results for Std-RF, SW-RF-MV and SW-RF-SLR, with class-weighting adjustment

Dataset	Std-RF	SW-RF-MV	SW-RF-SLR
EC_Angina	0.0000	0.0000	**0.6903**
EC_Arthritis	0.7191	0.6988	**0.7419**
EC_Cataract	0.4044	0.4952	**0.6696**
EC_Dementia	0.0000	0.0000	**0.7673**
EC_Diabetes	0.1636	0.0000	**0.7248**
EC_HBP	0.5840	0.5527	**0.6556**
EC_HeartAtt.	0.0000	0.0000	**0.6889**
EC_Osteop.	0.0237	0.0119	**0.6781**
EC_Parkin.	0.0000	0.0000	**0.7154**
EC_Stroke	0.0000	0.0000	**0.7017**
EN_Angina	0.0000	0.0000	**0.6565**
EN_Arthritis	0.7191	0.6988	**0.7419**
EN_Cataract	0.4937	0.4955	**0.6727**
EN_Dementia	0.0000	0.0000	**0.6300**
EN_Diabetes	0.4441	0.0000	**0.7248**
EN_HBP	0.6539	0.5527	**0.7041**
EN_HeartAtt.	0.0000	0.0000	**0.6827**
EN_Osteop.	0.0247	0.0370	**0.6485**
EN_Parkin.	0.0000	0.0000	**0.5160**
EN_Stroke	0.0000	0.0000	**0.6564**
EC + EN Rank	2.40	2.60	**1.00**
Friedman Test		p-value = <0.001	
Adjusted alpha α^*	0.050	0.025	control
p-value	**<0.001**	**<0.001**	N/A

Table 5: AUROC results for Std-RF, SW-RF-MV and SW-RF-SLR, with class-weighting adjustment

Dataset	Std-RF	SW-RF-MV	SW-RF-SLR
EC_Angina	0.7554	0.7853	**0.7897**
EC_Arthritis	0.8122	0.8096	**0.8156**
EC_Cataract	0.7454	**0.7625**	0.7421
EC_Dementia	0.8443	0.8551	**0.8781**
EC_Diabetes	0.7974	0.7898	**0.8053**
EC_HBP	0.7092	0.7079	**0.7144**
EC_HeartAtt.	0.7378	**0.7748**	0.7615
EC_Osteop.	0.7518	**0.7755**	0.7643
EC_Parkin.	0.7240	0.7712	**0.7973**
EC_Stroke	0.7642	0.7807	**0.7834**
EN_Angina	0.7554	**0.7853**	0.7443
EN_Arthritis	0.8122	0.8096	**0.8156**
EN_Cataract	0.7200	**0.7625**	0.7240
EN_Dementia	**0.8443**	0.7654	0.7525
EN_Diabetes	**0.8664**	0.7898	0.8053
EN_HBP	0.7621	0.7573	**0.7665**
EN_HeartAtt.	0.7378	**0.7623**	0.7598
EN_Osteop.	0.7169	**0.7342**	0.7219
EN_Parkin.	0.6051	**0.7712**	0.5966
EN_Stroke	0.7218	0.7807	**0.7834**
EC + EN Rank	2.40	1.90	**1.70**
Friedman Test		p-value = 0.074	

5.3 Comparing All Combinations of Classification Methods and Class-Balancing Scenarios

Subsections 5.1 and 5.2 compared classification methods in two scenarios: a class-imbalance scenario (as in the original dataset) and a scenario where a class-weight adjustment method was used to improve class balancing, respectively.

This subsection presents the results for a complementary type of analysis, from a more holistic perspective, considering all 6 combinations of 3 types of classification methods times 2 class-balancing scenarios.

Note that the motivation for this type of analysis is not to compare the proposed SW-RF-SLR method against the two other methods in a controlled scenario (which has been done in the previous two subsections), but rather to compare the different combinations of method and class-balancing scenario as a whole, considering the interaction between these two components of a classification system. The results of this type of analysis seem particularly useful for real-world applications, since in practice users should consider that interaction when deciding which classification system they would use.

For the GMean metric (Table 6), SW-RF-SLR with class-weight adjustment attained the ideal average rank of 1.00, achieving the best result in all 20 datasets. The Friedman test's p-value (< 0.001) shows that there is a significant difference among the results of all 6 combinations of methods and class-balancing scenarios as a whole. The results of Holm's post-hoc test indicate that SW-RF-SLR with class-weight adjustment significantly outperformed all other combinations of method and class-balancing scenario ($p < 0.001$ for each comparison).

Table 6: GMean results for all combinations of a classification method and a class-balancing scenario

Dataset	Imbalanced classes **Std-RF**	Class weighting **Std-RF**	Imbalanced classes **SW-RF-MV**	Class weighting **SW-RF-MV**	Imbalanced classes **SW-RF-SLR**	Class weighting **SW-RF-SLR**
EC_Angina	0.0000	0.0000	0.0000	0.0000	0.0000	**0.6903**
EC_Arthritis	0.7204	0.7191	0.7002	0.6988	0.7154	**0.7419**
EC_Cataract	0.4646	0.4044	0.5369	0.4952	0.0000	**0.6696**
EC_Dementia	0.1353	0.0000	0.0000	0.0000	0.2596	**0.7673**
EC_Diabetes	0.1811	0.1636	0.0096	0.0000	0.2736	**0.7248**
EC_HBP	0.6021	0.5840	0.5744	0.5527	0.5388	**0.6556**
EC_HeartAtt.	0.0000	0.0000	0.0000	0.0000	0.0000	**0.6889**
EC_Osteop.	0.0237	0.0237	0.0000	0.0119	0.0475	**0.6781**
EC_Parkin.	0.0000	0.0000	0.0000	0.0000	0.0000	**0.7154**
EC_Stroke	0.0297	0.0000	0.0000	0.0000	0.0297	**0.7017**
EN_Angina	0.0000	0.0000	0.0000	0.0000	0.0000	**0.6565**
EN_Arthritis	0.5903	0.7191	0.7002	0.6988	0.7154	**0.7419**
EN_Cataract	0.4646	0.4937	0.5369	0.4955	0.3641	**0.6727**
EN_Dementia	0.0000	0.0000	0.0000	0.0000	0.0000	**0.6300**
EN_Diabetes	0.1811	0.4441	0.4640	0.0000	0.2736	**0.7248**
EN_HBP	0.6021	0.6539	0.5744	0.5527	0.5388	**0.7041**
EN_HeartAtt.	0.0316	0.0000	0.0316	0.0000	0.0000	**0.6827**
EN_Osteop.	0.0237	0.0247	0.0000	0.0370	0.0799	**0.6485**
EN_Parkin.	0.0000	0.0000	0.0000	0.0000	0.0000	**0.5160**
EN_Stroke	0.0297	0.0000	0.0000	0.0000	0.0297	**0.6564**
Avg Rank	3.65	3.88	4.08	4.55	3.85	**1.00**
Friedman Test	p-value = <0.001					
Adjusted alpha α^*	0.050	0.017	0.013	0.010	0.025	control
p-value	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	N/A

For the AUROC metric (Table 7), both SW-RF-SLR with class-weight adjustment and SW-RF-MV with class-weight adjustment attained the best average rank of 2.50. The Friedman test's result, p-value < 0.001, indicates a significant difference among the results of all 6 method/class-balance combinations as a whole. Between the two joint winners, we chose SW-RF-SLR with class-weight adjustment as the control combination (to be compared against each of the other combinations), for two reasons. First, this combination has a higher number of wins than SW-RF-MV with class-weight adjustment (10 vs. 6 wins, respectively), so the former would be chosen as the best for a larger number of datasets in practice. Second, as the main contribution of this work is to introduce SW-RF-SLR, it is more important to compare a combination using this method (rather than a combination using SW-RF-MV) against each other combination.

The results of the Holm's test indicate that SW-RF-SLR with class-weight adjustment significantly outperformed Std-RF with imbalanced classes ($p = 0.002 < 0.013$) and SW-RF-SLR with imbalanced classes ($p < 0.001 < 0.010$).

In summary, the combination of SW-RF-SLR and class-weight adjustment achieved the best overall result in these experiments. More precisely, its results were significantly better than the results of all other 5 method/class-balance combinations for the GMean metric, and significantly better than the results of

two other combinations for the AUROC metric (with the caveat that for this metric the best results are shared by both the SW-RF-SLR with class-weight adjustment and the SW-RF-MV with class-weight adjustment combinations).

Table 7: AUROC results for all combinations of a classification method and a class-balancing scenario

Dataset	Imbalanced classes **Std-RF**	Class weighting **Std-RF**	Imbalanced classes **SW-RF-MV**	Class weighting **SW-RF-MV**	Imbalanced classes **SW-RF-SLR**	Class weighting **SW-RF-SLR**
EC_Angina	0.7542	0.7554	0.7811	0.7853	0.7504	**0.7897**
EC_Arthritis	0.8116	0.8122	0.8084	0.8096	0.8101	**0.8156**
EC_Cataract	0.7445	0.7454	0.7618	**0.7625**	0.7440	0.7421
EC_Dementia	0.8544	0.8443	0.8650	0.8551	0.8402	**0.8781**
EC_Diabetes	0.7870	0.7974	0.7867	0.7898	0.7834	**0.8053**
EC_HBP	0.7091	0.7092	0.7060	0.7079	0.7073	**0.7144**
EC_HeartAtt.	0.7399	0.7378	0.7659	**0.7748**	0.7382	0.7615
EC_Osteop.	0.7507	0.7518	0.7734	**0.7755**	0.7554	0.7643
EC_Parkin.	0.7136	0.7240	0.7598	0.7712	0.7442	**0.7973**
EC_Stroke	0.7611	0.7642	0.7815	0.7807	0.7639	**0.7834**
EN_Angina	0.7542	0.7554	0.7429	**0.7853**	0.7504	0.7443
EN_Arthritis	0.6771	0.8122	0.8084	0.8096	0.8101	**0.8156**
EN_Cataract	0.7445	0.7200	0.7315	**0.7625**	0.7201	0.7240
EN_Dementia	0.7485	**0.8443**	0.7952	0.7654	0.7779	0.7525
EN_Diabetes	0.7870	0.8664	**0.8733**	0.7898	0.7834	0.8053
EN_HBP	0.7091	0.7621	0.7060	0.7573	0.7073	**0.7665**
EN_HeartAtt.	0.7561	0.7378	**0.7633**	0.7623	0.7382	0.7598
EN_Osteop.	0.7507	0.7169	**0.7734**	0.7342	0.7166	0.7219
EN_Parkin.	0.5714	0.6051	0.7598	**0.7712**	0.3950	0.5966
EN_Stroke	0.7611	0.7218	0.7382	0.7807	0.7639	**0.7834**
Avg Rank	4.30	3.65	3.25	**2.50**	4.80	**2.50**
Friedman Test			p-value = <0.001			
Adjusted alpha α^*	0.013	0.017	0.025	0.050	0.010	control
p-value	**0.002**	0.052	0.205	1.000	**<0.001**	N/A

6 Conclusion

The main contribution of this work is to propose a novel stacking-based Separate Waves (SepWav) method as a type of data-transformation approach for longitudinal data classification. The proposed method uses a logistic regression algorithm as a meta-classifier in a stacking framework for combining the predictions of base classifiers (random forests) that were separately trained with data from each wave (time point), as opposed to simply combining the base classifiers' predictions across waves via voting as usual.

The proposed method was compared against two other methods: (a) a standard RF method, using the baseline data-transformation approach of ignoring the features' time indices (wave IDs), i.e., disregarding temporal dependencies in the data; and (b) a standard SepWav method where the outputs of the base classifiers are combined by majority voting as usual. Each of the 3 methods was evaluated in two class-balancing scenarios: (a) with the imbalanced class distribution in the original datasets, and (b) with class-weight adjustment.

The experiments involved 20 datasets extracted from the English Longitudinal Study of Ageing (ELSA) and two predictive accuracy metrics: the Geometric Mean of sensitivity and specificity (GMean) and AUROC.

Overall, the combination of the proposed stacking-based SepWav method using Logistic Regression as the meta-classifier (SW-RF-SLR) and class-weight ad-

justment outperformed the other 5 combinations of method and class-balancing scenario, with statistically significant results in most cases.

Note, however, that in the experiments the superiority of this method was observed only when it was used together with class-weight adjustment (as a class-balancing technique); i.e., SW-RF-SLR did not perform well in the scenario with imbalanced classes (as in the original datasets). This shows the importance of also using a class-balancing technique when classifying real-world longitudinal data with a large degree of class imbalance, as in our datasets.

These findings underscore the potential of stacking to improve longitudinal classification by effectively integrating the outputs of wave-specific base classifiers, addressing the temporal dynamics in a longitudinal dataset more robustly than the other two evaluated methods. This advancement is particularly relevant in biomedical domains, where accurate modelling of time-dependent health data can potentially enhance disease prediction and patient monitoring.

Future research could involve exploring alternative meta-classifiers or comparing the results of proposed method against the results of recurrent neural network / LSTM methods, which learn temporal dependencies in the data.

Code Availability

The SepWav methods and scripts for model evaluation are available in the Scikit-Longitudinal library [15], a Sklearn-like library for longitudinal ML.

Acknowledgments This work was conducted without external funding.

Disclosure of Interests The authors declare no competing interests.

References

1. Allam, A., Feuerriegel, S., Rebhan, M., Krauthammer, M.: Analyzing patient trajectories with artificial intelligence. J Med Internet Res **23**(12), e29812 (Dec 2021)
2. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7), 1145–1159 (1997)
3. Cascarano, A., Mur-Petit, J., Hernández-González, J., Camacho, M., de Toro Eadie, N., Gkontra, P., Chadeau-Hyam, M., Vitrià, J., Lekadir, K.: Machine and deep learning for longitudinal biomedical data: a review of methods and applications. Artificial Intelligence Review **56**(2), 1711–1771 (2023)
4. Clemens, S., Phelps, A., Oldfield, Z., Blake, M., Oskala, A., Marmot, M., Rogers, N., Banks, J., Steptoe, A., Nazroo, J.: English longitudinal study of ageing: Waves 0-8, 1998-2017 (2019)
5. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7**(1), 1–30 (2006)
6. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple Classifier Systems. pp. 1–15. Springer (2000)
7. Erdogan, B.E., Akyüz, S.Ö.: A weighted ensemble learning by svm for longitudinal data: Turkish bank bankruptcy. In: Tez, M., von Rosen, D. (eds.) Trends and Perspectives in Linear Statistical Inference. pp. 89–103. Springer (2018)
8. Finkelstein, J., Jeong, I.C.: Machine learning approaches to personalize early prediction of asthma exacerbations. Annals of the New York Academy of Sciences **1387**(1), 153–165 (2017)

9. Jie, B., Liu, M., Liu, J., Zhang, D., Shen, D.: Temporally constrained group sparse learning for longitudinal data analysis in alzheimer's disease. IEEE Transactions on Biomedical Engineering **64**(1), 238–249 (2017)
10. Kelloway, E.K., Francis, L.: Longitudinal research and data analysis. In: Research methods in occupational health psychology, pp. 374–394. Routledge (2012)
11. King, G., Zeng, L.: Logistic regression in rare events data. Political Analysis **9**(2), 137–163 (2001)
12. Lawson, J.D., Lim, Y.: The geometric mean, matrices, metrics, and more. The American Mathematical Monthly **108**(9), 797–812 (2001)
13. Pavlyshenko, B.: Using stacking approaches for machine learning models. In: 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). pp. 255–258 (2018)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
15. Provost, S., Freitas, A.A.: Scikit-longitudinal: A machine learning library for longitudinal classification in python. Journal of Open Source Software **10**(112), 8481 (2025). https://doi.org/10.21105/joss.08481
16. Ribeiro, C., Freitas, A.A.: 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
17. Ribeiro, C.E.: New Longitudinal Classification Approaches and Applications to Age-Related Disease Data. Ph.d. thesis, School of Computing (2022)
18. Sacks, D.B., Arnold, M., Bakris, G.L., Bruns, D.E., Horvath, A.R., Kirkman, M.S., Lernmark, A., Metzger, B.E., Nathan, D.M.: Guidelines and recommendations for laboratory analysis in the diagnosis and management of diabetes mellitus. Clinical Chemistry **57**(6), e1–e47 (06 2011)
19. Siegel, R.L., Giaquinto, A.N., Jemal, A.: Cancer statistics, 2024. CA: A Cancer Journal for Clinicians **74**(1), 12–49 (Jan 2024)
20. Simon, G.E., Johnson, E., Lawrence, J.M., Rossom, R.C., Ahmedani, B., Lynch, F.L., Beck, A., Waitzfelder, B., Ziebell, R., Penfold, R.B., Shortreed, S.M.: Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. American Journal of Psychiatry **175**(10), 951–960 (Oct 2018)
21. Singh, A., Nadkarni, G., Gottesman, O., Ellis, S.B., Bottinger, E.P., Guttag, J.V.: Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration. Journal of biomedical informatics **53**, 220–228 (2015)
22. Susman, A., Krishnamurthy, R., Li, Y.C., Olaimat, M., Bozdag, S., Varghese, B., Sheikh-Bahaei, N., Pandey, G.: Longitudinal ensemble integration for sequential classification with multimodal data pre-print arxiv:2411.05983 (2024)
23. Van Daalen, F., Smirnov, E., Davarzani, N., Peeters, R., Karel, J., Brunner-La Rocca, H.P.: An ensemble approach to time dependent classification. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1007–1011 (2018)
24. WHO, W.H.O.: Ageing and health (Oct 2024), https://www.who.int/news-room/fact-sheets/detail/ageing-and-health
25. Zhao, Y., Healy, B.C., Rotstein, D., Guttmann, C.R.G., Bakshi, R., Weiner, H.L., Brodley, C.E., Chitnis, T.: Exploration of machine learning techniques in predicting multiple sclerosis disease course. Plos One **12**(4), 1–13 (04 2017)