BIT.UA at iDPP: Predictive Analytics on ALS Disease Progression Using Sensor Data with Machine Learning

Jorge Miguel Silva[1], José Luis Oliveira[1]

[1]IEETA/DETI, LASI, University of Aveiro, Portugal

Abstract

This technical report details our participation in the iDPP CLEF 2024 challenge, focusing on predictive modeling of ALS progression using sensor data from smartwatches. We competed in Tasks 1 and 2, aiming to predict clinical and self-assessed ALS Functional Rating Scale-Revised (ALSFRS-R) scores. Our methodology centered on machine learning techniques, primarily employing an ensemble of Random Forest classifiers. Our best model, which utilized temporal analysis, achieved a Mean Absolute Error (MAE) of 0.25 in Task 1 and 0.326 in Task 2, with corresponding Root Mean Square Errors (RMSE) of 0.544 and 0.608, respectively. This demonstrates the model's effective leverage of time-series data across different assessment settings While the temporal model excelled in capturing the nuances of ALS progression through sensor data, variations in performance between Tasks 1 and 2 highlighted the challenges posed by the subjective nature of self-assessments. Our results contribute to understanding the potential and limitations of current predictive models, emphasizing the importance of sophisticated time-series analysis in improving prognostic tools for ALS.

Keywords

ALS, Disease Progression, Sensor Data, Machine Learning, Predictive Modeling

1. Introduction

Amyotrophic Lateral Sclerosis (ALS) is a progressive neurodegenerative disorder that primarily affects motor neurons in the brain and spinal cord, leading to severe physical disability and ultimately, respiratory failure [1]. Epidemiological studies indicate that ALS affects approximately 2 per 100,000 individuals annually, with most patients experiencing onset between the ages of 40 and 70 [2]. Despite the advancement in understanding ALS pathophysiology, its etiology remains largely idiopathic, posing significant challenges in its diagnosis and management [3]. Consequentially there is a need for reliable diagnostic tools and methodologies to track disease progression and evaluate therapeutic outcomes effectively.

The ALS Functional Rating Scale-Revised (ALSFRS-R) is a validated instrument used to determine the degree of functional impairment in ALS patients. It encompasses twelve domains of daily activities, each rated on a scale from 0 to 4, where higher scores denote greater functional independence. The ALSFRS-R provides a quantifiable measure of disease progression, correlates significantly with quality of life measures and is extensively used in clinical trials and practice [4]. However, the inherent subjectivity of self-reported measures and the episodic nature of clinical evaluations necessitate the exploration of objective, continuous monitoring methods to supplement these assessments.

In this context, wearable technology offers unprecedented opportunities for real-time, objective monitoring of physiological parameters. Sensor data collected via smartwatches and specialized mobile applications can provide detailed insights into the physical activity patterns and autonomic functions of patients, potentially enabling early detection of disease progression markers [5]. Moreover, these data could facilitate the development of predictive models that anticipate changes in ALSFRS-R scores, thereby offering a proactive approach to disease management.

The 2024 Intelligent Disease Progression Prediction (iDPP) challenge at the Conference and Labs of the Evaluation Forum (CLEF) [6, 7] encompasses a series of tasks aimed at leveraging sensor and

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

jorge.miguel.ferreira.silva@ua.pt (J. M. Silva); jlo@ua.pt (J. L. Oliveira)
http://jorgemfs.com (J. M. Silva)
0000-0002-6331-6091 (J. M. Silva); 0000-0002-6672-6176 (J. L. Oliveira)

environmental data to predict disease progression in ALS and Multiple Sclerosis (MS). Specifically, this research focuses on two tasks within the ALS domain: predicting ALSFRS-R scores from longitudinal sensor data and evaluating patient self-assessment scores derived through mobile health applications. These tasks are designed to address the pressing need for tools that can predict disease trajectory and patient-reported outcomes from objective data sources [8].

This work describes the participation of our team in iDPP at CLEF 2024, specifically in Tasks 1 and 2.

2. Related Work

Advancements in predictive modeling for ALS have utilized machine learning and deep learning to forecast disease understanding and progression effectively.

Seminal work by Hothorn *et al.* [9] used the PRO-ACT database to develop a Random Forest algorithm as part of the DREAM Phil Bowen ALS Prediction Prize4Life Challenge, highlighting the importance of past disease progression. Subsequently, Tang *et al.* [10] expanded this approach by using both model-based and model-free methods to predict changes in ALSFRS-R scores, demonstrating the capability of machine learning in capturing complex interactions within ALS data. Further refining model capabilities, Faghri *et al.* [11] applied machine learning techniques to identify distinct clinical subgroups within the ALS population, enhancing the understanding of ALS's heterogeneity and supporting the development of targeted interventions.

On the other hand, recent studies like those by Johnson *et al.*[12] and Vieira *et al.*[13] have demonstrated the potential of wearable technology and sensor data in monitoring ALS. These technologies enable both active and passive data collection, providing continuous monitoring of ALS progression, which is highly correlated with traditional ALSFRS-R scores.

Deep learning has notably advanced ALS research through diverse applications, for example, Van der Burgh *et al.* [14] combined MRI data with deep learning to predict patient survival times with high accuracy and Sengur *et al.* [15] and Yin *et al.* [16] applied convolutional neural networks to classify EMG signals and genetic data enhancing the understanding of complex genetic interactions in ALS. More recent work by Müller *et al.* [17] and Pancotti *et al.* [18] utilized advanced neural networks to model disease progression and predict future ALSFRS-R scores.

The BRAINTEASER project [19] exemplifies the successful integration of clinical assessments with real-time sensor data, creating comprehensive models that detect nuanced changes in ALS progression, thereby facilitating more responsive patient care. The iDPP@CLEF challenges in 2023 [20] and the current challenge [8] demonstrate the difficulties in integrating environmental data into predictive models, highlighting the complex nature of environmental influences on ALS progression that remains to be fully understood.

3. Methodology and Experimental Setup

3.1. Data Collection

This study uses data from smartwatches and mobile health applications, focusing on physiological metrics such as heart rate, steps, and activity levels, alongside patient-reported outcomes. The data captures continuous physiological data, giving detailed information on many features related to patient activity and health status over time [6, 7]. The datasets employed are crucial for developing predictive models for ALS progression. The primary datasets included:

- **train-static.csv**: Contains demographic and static information about the patients, including variables such as age, sex, and baseline health metrics.
- **train-sensor.csv**: Comprises time-series data from various sensors, capturing daily physiological and activity metrics.
- **train-alsfrs.csv**: Provides ALS Functional Rating Scale-Revised (ALSFRS-R) scores, which serve as target variables for the models.

- Corresponding test datasets are used for evaluating model performance.

3.2. Data Preprocessing

To ensure the quality and consistency of the input data, a rigorous preprocessing pipeline was implemented. The preprocessing steps included handling missing values, feature engineering, and data cleaning.

3.2.1. Handling Missing Values

Given the longitudinal nature of the data, missing values occurred because some patients did not follow the guidelines every day. To address this issue, we used imputation techniques.

Specifically, Iterative and Simple Imputation were employed to ascribe missing values in both static and sensor datasets [21]. IterativeImputer performs multivariate imputation by modeling each feature with missing values as a function of other features in a round-robin fashion, making it particularly effective for datasets with complex interactions between variables. Additionally, in some cases, mean and median imputation were experimented with for numeric columns, depending on their distribution and specific characteristics. This approach ensures that the imputed values are reasonable and do not introduce significant biases. Finally, rows with more than half of their values missing were excluded to maintain data integrity. This threshold was chosen to balance retaining as much data as possible while ensuring the remaining data was of high quality.

3.2.2. Feature Engineering

Several approaches were used for feature engineering to enhance the predictive power of the models by creating new features from the raw data to capture underlying patterns and trends more effectively. Particularly, we aggregated sensor data using statistical measures such as mean, median, standard deviation, minimum, maximum, and range. These aggregations were performed over specified time windows to capture temporal dynamics. On the other hand, we used the tsfresh library [22] for time-series feature extraction. Tsfresh provided a wide range of feature extraction methods tailored for time-series data, which helped capture complex temporal patterns of data. Finally, historical data was incorporated by including previous ALSFRS-R scores as features to provide historical context and improve model accuracy. This approach leverages the temporal aspect of the data, acknowledging that past scores are indicative of future outcomes.

3.3. Feature Selection

The initial feature extraction process generated a vast number of features. To enhance model performance and reduce computational complexity, feature selection was performed based on relevance and significance.

Correlation analysis was conducted to analyze the correlation of features with target variables. Highly correlated features were prioritized as they are more likely to have a significant impact on model performance.

Preliminary models were trained to evaluate the importance of each feature. This approach helped identify which features contributed most to the predictive accuracy of the models. Features with low importance scores were discarded to streamline the model and improve efficiency.

3.4. Model Development

3.4.1. Algorithm Selection

The choice of a machine learning algorithm was crucial for the success of our predictive models. We ultimately selected the Random Forest Classifier (RFC) for its robustness and superior performance in handling multi-output regression tasks, very important for predicting multiple ALSFRS-R scores simultaneously. Initially, we considered several algorithms, including linear regression, support vector machines (SVM), and gradient boosting machines. The RFC's ability to handle missing labels in classification further justified our choice. As an ensemble learning method, the RFC constructs multiple decision trees during training and averages their predictions. This approach is particularly advantageous for managing large datasets with high dimensionality and is resistant to overfitting.

To manage the prediction of multiple ALSFRS-R scores, a Multi-output Classifier was used. This meta-estimator fitted a separate Random Forest Classifier for each target variable, allowing for simultaneous prediction of all 12 ALSFRS-R score components. By doing so, we ensured that each target is treated independently, which is beneficial given the varying degrees of correlation between different ALSFRS-R components.

3.4.2. Pipeline Construction

A comprehensive data processing and modeling pipeline was constructed to streamline the workflow and ensure reproducibility. This pipeline was designed to handle all necessary preprocessing steps before feeding the data into the classifier. The key components of the pipeline included:

Preprocessing steps involved normalization and encoding of features to prepare the data for modeling. For numerical features, the StandardScaler was used to scale them to have a mean of zero and a standard deviation of one. Scaling is crucial for algorithms like Random Forest that are sensitive to the scales of input features. For categorical features such as sex, the OneHotEncoder was applied, transforming them into a format suitable for machine learning algorithms.

The last step in the pipeline was the application of the Multi-Output Classifier with the RFC as the underlying estimator. This step involved fitting the model to the training data and using it to make predictions.

3.5. Hyperparameter Tuning

For this study, hyperparameter tuning was performed using Randomized Search, a method that allows for more efficient exploration of the hyperparameter space compared to an exhaustive grid search. Randomized Search CV samples a fixed number of parameter settings from the specified distributions and evaluates them using cross-validation.

The hyperparameters tuned included the number of estimators, which refers to the number of trees in the forest. This parameter influences the overall performance and stability of the predictions. Maximum depth, the maximum depth of the trees, controls the complexity of the model and helps prevent overfitting. Minimum samples split, the minimum number of samples required to split an internal node, impacts the model's ability to generalize to new data. Minimum samples leaf, the minimum number of samples required to be at a leaf node, similarly affects the model's complexity and generalization. Maximum features, the number of features to consider when looking for the best split, can significantly affect the model's performance by reducing overfitting and improving generalization. Lastly, bootstrap, which determines whether bootstrap samples are used when building trees, affects the diversity of the trees in the forest.

3.6. Validation Strategy

To ensure the robustness of the model, first opted for a multi-label stratified shuffle split was used for cross-validation. This method maintains the balance of multiple target variables across the training and validation folds, ensuring that each fold is representative of the overall data distribution. However, due to label misrepresentation, a simple multi-label shuffle split was used in most cases.

3.7. Approaches for ALSFRS-R Score Prediction

The following tables summarize the distinct methodological (Table 1) and ensemble approaches (Table 2) utilized for predicting ALSFRS-R scores in Tasks 1 and 2 of the iDPP CLEF 2024 challenge.

Method	Description
Mean	Employs the average of observed values from the sensor data to establish a baseline prediction, simplifying the model by reducing noise and variance in the data.
Median	Uses the median of the sensor data to provide a robust prediction that is less sensitive to outliers, aiming to capture the central tendency of the ALSFRS-R scores.
More Metrics	Integrates a wider array of statistical measures (mean, median, standard deviation, min, max) from the sensor data to create a comprehensive profile for prediction.
Temporal	Leverages advanced time-series analysis techniques to capture and utilize the temporal patterns in the sensor data, emphasizing changes over time in the ALSFRS-R scores.

Table 1
Table describing methodologies implemented for Task 1 and 2 of iDPP.

Ensemble Method	Description
Minimum	Uses the minimum value among all model predictions for each score, aiming to provide a conservative estimate of the ALSFRS-R scores.
Average	Calculates the average of predictions from all models, balancing the influence of each model to mitigate extreme predictions and smooth outliers.
Maximum	Selects the maximum value from the predictions, reflecting an optimistic estimation which may capture potential over-performances in patient assessment scores.

Table 2
Table describing ensembles implemented for Task 1 and 2 of iDPP.

Each method and ensemble strategy is designed to address specific characteristics of ALS progression data. These approaches range from basic statistical summaries to complex time-series models and integrated ensemble methods, providing a solution set for diverse data analysis solutions.

3.8. Dataset Splits

The data was split into training, validation, and test sets, ensuring temporal consistency and patient-specific segregation to prevent data leakage and ensure reliable evaluation. We first separated the train datasets into an 80-20 split, where 80% of the data was used for training and the remainder for validation. Upon receiving the test data, training was done using all data.

3.9. Hardware and Environment

Experiments were conducted on a virtual machine with a Ubunto OS equipped with 32 GB of memory configurations to handle the computational demands of feature extraction and model training.

4. Results

In this section, we present the results for Task 1 and Task 2 of the iDPP CLEF 2024 challenge, covering both the validation and test phases. For both tasks, the same methodological approaches—Mean, Median, More Metrics, and Temporal Analysis—were consistently applied. Additionally, for the test predictions, we employed an ensemble approach that combined the results of the Mean, Median, More Metrics, and Temporal Analysis methods. This ensemble approach utilized the minimum, average, and maximum values of each predicted score to enhance the robustness and accuracy of our final model outputs. This approach allowed us to directly compare the effectiveness of each methodology across different stages of the evaluation process and assess the robustness of our models under varying conditions inherent in clinically-assessed and self-assessed ALS Functional Rating Scale-Revised (ALSFRS-R) scores.

4.1. Validation Results

The validation results for Task 1 and Task 2 of the iDPP CLEF 2024 challenge, presented in Tables 3 and 4 respectively, illustrate the performance of four methodologies—Mean, Median, Combination of several metrics (More Metrics), and Temporal—across twelve parameters of the ALS Functional Rating Scale-Revised (ALSFRS-R). These methodologies are quantified through Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), providing insights into model accuracy and consistency across two distinct tasks: clinically-assessed and self-assessed ALSFRS-R scores.

Method	**Mean**		**Median**		**More Metrics**		**Temporal**	
Metric	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Q1	0.2963	0.6086	0.2593	0.5774	0.2963	0.6086	0.1852	0.4303
Q2	0.1111	0.3333	0.1481	0.3849	0.1481	0.3849	0.0370	0.1925
Q3	0.1481	0.3849	0.1481	0.3849	0.1481	0.3849	0.1111	0.3333
Q4	0.4074	0.6383	0.4074	0.6383	0.3333	0.5774	0.3333	0.5774
Q5	0.3333	0.6939	0.4815	0.7935	0.4815	0.7935	0.4444	0.7698
Q6	0.5185	0.8165	0.4815	0.7935	0.4815	0.7935	0.5926	0.8607
Q7	0.3704	0.6086	0.3333	0.5774	0.3333	0.5774	0.3704	0.7698
Q8	0.2222	0.5443	0.2222	0.5443	0.2222	0.5443	0.2963	0.6667
Q9	0.3333	0.7454	0.3333	0.7454	0.2593	0.6939	0.6667	1.1863
Q10	0.1852	0.4303	0.1852	0.4303	0.1852	0.4303	0.1111	0.3333
Q11	0.2222	0.6086	0.2593	0.6383	0.2222	0.6086	0.1111	0.3333
Q12	0.3704	0.9027	0.3333	0.8819	0.3333	0.8819	0.1852	0.6939
Average	0.2932	0.6096	0.2994	0.6158	0.2870	0.6066	0.2870	0.5956

Table 3
Task 1 validation results obtained from bitua team using different methodologies.

In Task 1, Temporal Analysis demonstrates the most consistent superior performance, notably achieving the lowest MAE and RMSE in several quarters, exemplified by its notable performance in Q2 with an MAE of 0.0370 and RMSE of 0.1925. This method effectively leverages time-series data to capture the dynamic progression of ALS, indicating its robust capability to handle the complexities inherent in clinical data. Conversely, the combined metrics approach, which integrates a broader array of statistical features, shows comparable MAE but slightly less RMSE in some parameters.

The simpler statistical approaches of the Mean and Median methodologies exhibit higher error rates, especially in parameters associated with complex symptomatology like respiratory functions in later quarters. These methods, although computationally simpler, struggle with modeling nuanced changes over time, leading to reduced predictive reliability.

For Task 2, the challenges shift towards handling the greater subjectivity of self-assessment data, which introduces variability in reporting and perception by patients themselves. Here, the Temporal method again shows its strength in specific quarters but with increased variability across all parameters, reflecting the challenges of capturing subjective self-assessments over time. The combination of metrics (More Metrics) method provides slightly more stable results compared to the Temporal approach, though none of the methodologies achieve the lower error metrics seen in Task 1.

Both tasks reveal that performance disparities across different ALSFRS-R parameters highlight the complexities of predictive modeling in ALS. Lower MAEs and RMSEs in the initial parameters suggest clearer data patterns early in the disease progression, whereas increased errors in later quarters reflect the inherent challenges with more complex or sporadically reported symptoms.

Method	**Mean**		**Median**		**More Metrics**		**Temporal**	
Metric	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Q1	0.1600	0.4472	0.1600	0.4472	0.1600	0.4472	0.1429	0.4286
Q2	0.2400	0.5292	0.2000	0.4899	0.1800	0.4690	0.3061	0.7693
Q3	0.0400	0.2000	0.0200	0.1414	0.0400	0.2000	0.0816	0.2857
Q4	0.1200	0.4000	0.1600	0.4899	0.1600	0.4899	0.2449	0.5345
Q5	0.5000	0.8124	0.4800	0.7746	0.4600	0.7616	0.4082	0.6999
Q6	0.3200	0.6000	0.2600	0.5477	0.3400	0.6481	0.3061	0.5890
Q7	0.2400	0.5292	0.2200	0.4690	0.2600	0.5099	0.3673	0.6389
Q8	0.1600	0.4472	0.1800	0.4690	0.1400	0.4243	0.1429	0.4286
Q9	0.3000	0.5831	0.2600	0.5477	0.2400	0.5292	0.2041	0.4518
Q10	0.2000	0.5657	0.2200	0.6164	0.1800	0.5477	0.2653	0.6227
Q11	0.2600	0.7874	0.2800	0.8000	0.2600	0.7874	0.4898	1.2617
Q12	0.0600	0.3162	0.0800	0.3464	0.0800	0.3464	0.0816	0.3499
Average	0.2167	0.5181	0.2100	0.5116	0.2083	0.5134	0.2534	0.5884

Table 4
Task 2 validation results obtained from bitua team using different methodologies.

The comparative analysis of the two tasks underscores the crucial role of sophisticated modeling techniques, particularly those capable of handling temporal dynamics, to enhance the accuracy of predictive models in ALS monitoring.

4.2. Global Results

The global results of the participation in the challenge iDPP of CLEF 2024 are shown in Table 5 and Table 6. In both cases, the results of our team in the test dataset mirror the validation results.

#	**Team**	**Method**	**MAE**	**RSME**
1	fcool	fcool_T1_locf	0.202380952	0.491212504
1	idppexplorers	idppexplorers_T1_naiveSubmission	0.202380952	0.491212504
1	mandatory	mandatory_T1_d1	0.202380952	0.491212504
1	unipd	UNIPD_t1_hold	0.202380952	0.491212504
2	idppexplorers	idppexplorers_T1_EN	0.222222222	0.50478526
3	compbiomedunito	RandomForest_MonoWindow	0.234126984	0.518945141
4	bitua	bitua_T1_ensemble_max	0.25	0.544324192
6	bitua	bitua_T1_temporalAnalysis	0.333333333	0.606889042
15	bitua	bitua_T1_moremetrics	0.388888889	0.683397063
19	bitua	bitua_T1_median	0.400793651	0.700515112
20	bitua	bitua_T1_mean	0.400793651	0.706631136
23	bitua	bitua_T1_ensemble_avg	0.412698413	0.709317604
39	bitua	bitua_T1_ensemble_min	0.48015873	0.795525562
Median	-	-	**0.403489069**	**0.707744917**
Mean	-	-	**0.412698413**	**0.709317604**

Table 5
Task 1 results obtained from bitua team, compared to top results and mean and average results.

For Task 1, our most successful method, the $bitua\_T1\_ensemble_max$, employed an ensemble model that selected the maximum score from a set of predictions, achieving a Mean Absolute Error (MAE) of 0.25 and a Root Mean Square Error (RMSE) of 0.544324192. This method ranked fourth overall, suggesting that an optimistic perspective on ALS progression, where higher ALSFRS-R scores indicate better functional ability, may align more closely with certain patient trajectories.

Following this, the $bitua\_T1\_temporalAnalysis$ method, which incorporated advanced time-series analysis tools like tsfresh for feature extraction, demonstrated the ability to capture the dynamic aspects of ALS progression with an MAE of 0.333333333 and an RMSE of 0.606889042. This method which used the temporal patterns in the sensor data demonstrates the importance of observing these results over time in the classification of ALSFRS-R scores.

Conversely, our simpler approaches, specifically $bitua\_T1\_median$ and $bitua\_T1\_mean$, which condensed predictions into median or mean values, resulted in less precise outcomes with MAEs and RMSEs around 0.40 and 0.70, respectively. These methods, while straightforward, struggled to model the complexities of ALS progression adequately, reflecting the challenges in using reductionist approaches for such multifaceted medical data.

On the opposite end of our ensemble spectrum, the $bitua\_T1\_ensemble_min$, which utilized the minimum scores from ensemble predictions, produced the least effective outcomes among our entries, with an MAE of 0.48015873 and an RMSE of 0.795525562. This approach is likely overly pessimistic and thus struggled to achieve good results.

Comparatively, the performance of $bitua\_T1\_ensemble\_max$ closely approached that of the competition leaders, who all achieved an MAE of 0.202380952 and an RMSE of 0.491212504. Although slightly behind, the marginal differences of 0.0476 in MAE and 0.0531 in RMSE from the top results demonstrate competitive capability and underscore the potential of our ensemble and temporal analysis strategies.

Moreover, these methods significantly outperformed the median competition results (MAE of 0.403489069 and RMSE of 0.707744917), marking a clear advantage over the average performance. The distinctions in both MAE and RMSE underline the effectiveness of our approach and highlight our method's robustness compared to the broader field.

#	Team	Method	MAE	RSME
1	fcool	fcool_T2_locf	0.287878788	0.577431447
1	unipd	UNIPD_t2_hold	0.287878788	0.577431447
2	compbiomedunito	RandomForest_MonoWindow	0.310606061	0.601358746
3	bitua	bitua_T2_ensemble_max	0.325757576	0.608658838
4	bitua	bitua_T2_moremetrics	0.371212121	0.654232661
5	bitua	bitua_T2_mean	0.393939394	0.686437506
6	bitua	bitua_T2_median	0.401515152	0.708103457
8	bitua	bitua_T2_ensemble_avg	0.424242424	0.714777806
9	bitua	bitua_T1_temporalAnalysis	0.431818182	0.720780408
13	bitua	bitua_T2_ensemble_min	0.5	0.818368472
Median	-	-	**0.585532747**	**0.896730391**
Mean	-	-	**0.515151515**	**0.837021651**

Table 6
Task 2 results obtained from bitua team, compared to top results and mean and average results.

The best-performing method in Task 2 was also the $bitua\_T2\_ensemble\_max$, achieving a MAE of 0.325757576 and a RMSE of 0.608658838. Although less accurate compared to its performance in Task 1, this method still demonstrated robustness. It ranked third among all entries, suggesting that maximizing predictions can effectively capture optimistic patient-reported outcomes, likely reflecting better self-perceived functional abilities.

Following closely, the $bitua\_T2\_moremetrics$ method, which incorporated a broader array of metrics (mean, median, standard deviation, minimum and maximum values), posted an MAE of 0.371212121 and an RMSE of 0.654232661.

The variation in model performance between Tasks 1 and 2 can largely be attributed to the nature of the target variables. Task 1 focused on medically assigned ALSFRS-R scores, which are likely more consistent and standardized due to their clinical origin. In contrast, Task 2 dealt with self-assessment scores that can be more subjective and influenced by the patient's perception and mood at the time of evaluation. This inherent subjectivity and variability in self-assessment could make modeling more challenging, potentially explaining why the same methods yielded different levels of accuracy across the tasks.

Moreover, patient self-assessment might not always align closely with clinical evaluations, leading to discrepancies that can affect the performance of models trained primarily with clinical score patterns in mind. This misalignment could particularly impact methods like the $bitua\_T1\_temporalAnalysis$, which performed significantly worse in Task 2.

Despite these challenges, the $bitua\_T2\_ensemble\_max$ method's performance was quite competitive when compared to the leading entries in Task 2, fcool and unipd, both of which achieved an MAE of 0.287878788 and an RMSE of 0.577431447. The differences in MAE and RMSE were 0.038 and 0.031, respectively, indicating that even with the subjective nature of self-assessments, our method remained robust.

Furthermore, this method significantly outperformed the median and mean results across all teams, where the median MAE was 0.585532747 and the median RMSE was 0.896730391. This performance indicates not only the effectiveness of the ensemble max strategy in capturing the higher ends of patient self-reported outcomes but also its overall reliability in a more variably reported dataset.

5. Conclusion

This study has demonstrated the efficacy of using sensor data from smartwatches and mobile health applications to predict ALS disease progression, substantiated through our participation in the iDPP CLEF 2024's Tasks 1 and 2. These tasks emphasize the potential of machine learning models, particularly the Random Forest Classifier within an ensemble framework, to forecast the ALS Functional Rating Scale-Revised (ALSFRS-R) scores both from a clinical and a patient-centered perspective.

Our findings reveal that ensemble methods, particularly the maximization strategy, were the most effective across both tasks, though with varying degrees of success. Task 1, focusing on clinically assigned ALSFRS-R scores, benefited from the objective consistency these scores typically maintain, thereby allowing our models to achieve a higher predictive accuracy. Conversely, Task 2 involved predicting patient self-assessment scores that inherently presented more variability and subjectivity, posing greater challenges for our predictive models. The performance dip in Task 2 illustrates the complexities and the nuanced differences between patient-perceived symptoms and clinically evaluated symptoms.

These outcomes highlight the critical role of precise data handling and feature engineering in enhancing model performance. The integration of advanced time-series analysis tools and the strategic handling of missing data were crucial in developing models that could effectively interpret the complex nature of ALS progression. However, the discrepancy in performance between the two tasks also points to the need for models that can better accommodate the subjective variabilities inherent in self-assessments.

Future research should focus on refining the algorithms used in predictive modeling of ALS progression. This involves enhancing data preprocessing techniques and incorporating advanced machine learning frameworks that can better handle the variability and complexities of ALS data. Improving feature engineering is essential, particularly by integrating sophisticated temporal features such as rolling means and time-windowed aggregates, which can provide a deeper understanding of disease dynamics. Additionally, applying deep learning techniques, especially those tailored for time-series analysis like Long Short-Term Memory (LSTM) networks, could significantly enhance model performance. Incorporating anomaly detection for outlier identification and signal processing techniques will further refine the models' predictive capabilities. Exploring hybrid models that combine machine learning flexibility with rule-based logic could also offer robust solutions by merging data-driven insights with clinical expertise, leading to more accurate and responsive predictive models for ALS progression.

6. Funding

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. It also has received support from the EC under grant agreement 101081813, Genomic Data Infrastructure.

References

[1] S. Boillée, C. V. Velde, D. W. Cleveland, Als: a disease of motor neurons and their nonneuronal neighbors, Neuron 52 (2006) 39–59.

[2] L. C. Wijesekera, P. Nigel Leigh, Amyotrophic lateral sclerosis, Orphanet journal of rare diseases 4 (2009) 1–22.

[3] B. Oskarsson, T. F. Gendron, N. P. Staff, Amyotrophic lateral sclerosis: an update for 2018, in: Mayo clinic proceedings, volume 93, Elsevier, 2018, pp. 1617–1628.

[4] J. M. Cedarbaum, N. Stambler, E. Malta, C. Fuller, D. Hilt, B. Thurmond, A. Nakanishi, B. A. S. Group, A. complete listing of the BDNF Study Group, et al., The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function, Journal of the neurological sciences 169 (1999) 13–21.

[5] J. W. van Unnik, M. Meyjes, M. R. J. van Mantgem, L. H. van den Berg, R. P. van Eijk, Remote monitoring of amyotrophic lateral sclerosis using wearable sensors detects differences in disease progression and survival: a prospective cohort study, Ebiomedicine 103 (2024).

[6] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. Di Nunzio, P. Fariselli, J. García Dominguez, A. G. Marta Gromicho, E. Longato, S. Madeira, U. Manera, S. Marchesin, L. Menotti, G. Silvello, E. Tavazzi, E. Tavazzi, I. Trescato, M. Vettoretti, B. D. Camillo, N. Ferro, Overview of iDPP@CLEF 2024: The Intelligent Disease Progression Prediction Challenge, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, September 9th to 12th, 2024, 2024.

[7] G. Birolo, P. Bosoni, G. Faggioli, H. Aidos, R. Bergamaschi, P. Cavalla, A. Chiò, A. Dagliati, M. de Carvalho, G. Di Nunzio, P. Fariselli, J. García Dominguez, A. G. Marta Gromicho, E. Longato, S. Madeira, U. Manera, S. Marchesin, L. Menotti, G. Silvello, E. Tavazzi, E. Tavazzi, I. Trescato, M. Vettoretti, B. D. Camillo, N. Ferro, Intelligent Disease Progression Prediction: Overview of iDPP@CLEF 2024, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, 2024.

[8] Brainteaser, Intelligent disease progression prediction (idpp) challenge, https://brainteaser.dei.unipd.it/challenges/idpp2024/, 2024. Accessed: 2024-05-24.

[9] T. Hothorn, H. H. Jung, Randomforest4life: a random forest for predicting als disease progression, Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration 15 (2014) 444–452.

[10] M. Tang, C. Gao, S. A. Goutman, A. Kalinin, B. Mukherjee, Y. Guan, I. D. Dinov, Model-based and model-free techniques for amyotrophic lateral sclerosis diagnostic prediction and patient clustering, Neuroinformatics 17 (2019) 407–421.

[11] F. Faghri, F. Brunn, A. Dadu, A. Chiò, A. Calvo, C. Moglia, A. Canosa, U. Manera, R. Vasta, F. Palumbo, et al., Identifying and predicting amyotrophic lateral sclerosis clinical subgroups: a population-based machine-learning study, The Lancet Digital Health 4 (2022) e359–e369.

[12] S. A. Johnson, M. Karas, K. M. Burke, M. Straczkiewicz, Z. A. Scheier, A. P. Clark, S. Iwasaki, A. Lahav, A. S. Iyer, J.-P. Onnela, et al., Wearable device and smartphone data quantify als progression and may provide novel outcome measures, NPJ Digital Medicine 6 (2023) 34.

[13] F. G. Vieira, S. Venugopalan, A. S. Premasiri, M. McNally, A. Jansen, K. McCloskey, M. P. Brenner, S. Perrin, A machine-learning based objective measure for als disease severity, NPJ digital medicine 5 (2022) 45.

[14] H. K. van der Burgh, R. Schmidt, H.-J. Westeneng, M. A. de Reus, L. H. van den Berg, M. P. van den

Heuvel, Deep learning predictions of survival based on mri in amyotrophic lateral sclerosis, NeuroImage: Clinical 13 (2017) 361–369.

[15] A. Sengur, Y. Akbulut, Y. Guo, V. Bajaj, Classification of amyotrophic lateral sclerosis disease based on convolutional neural network and reinforcement sample learning algorithm, Health information science and systems 5 (2017) 1–7.

[16] B. Yin, M. Balvert, R. A. van der Spek, B. E. Dutilh, S. Bohte, J. Veldink, A. Schönhuth, Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype, Bioinformatics 35 (2019) i538–i547.

[17] M. Müller, M. Gromicho, M. de Carvalho, S. C. Madeira, Explainable models of disease progression in als: Learning from longitudinal clinical data with recurrent neural networks and deep model explanation, Computer Methods and Programs in Biomedicine Update 1 (2021) 100018.

[18] C. Pancotti, G. Birolo, C. Rollo, T. Sanavia, B. Di Camillo, U. Manera, A. Chiò, P. Fariselli, Deep learning methods to predict amyotrophic lateral sclerosis disease progression, Scientific reports 12 (2022) 13738.

[19] Brainteaser, Scientific publications, https://brainteaser.health/resources/scientific-publications/, 2024. Accessed: 2024-05-23.

[20] G. Faggioli, A. Guazzo, S. Marchesin, L. Menotti, I. Trescato, A. Helena, R. Bergamaschi, G. Birolo, P. Cavalla, A. Chiò, et al., Overview of idpp@ clef 2023: the intelligent disease progression prediction challenge, in: CEUR WORKSHOP PROCEEDINGS, volume 3497, CEUR-WS, 2023, pp. 1123–1164.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[22] M. Christ, N. Braun, J. Neuffer, A. W. Kempa-Liehr, Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package), Neurocomputing 307 (2018) 72–77.