Assessment of Residual Deep Neural Networks and AdaBoost to predict adherence to digital-based active and healthy aging interventions

Sergio Pérez-Velasco[1] [0000-0002-2999-3216], Gonzalo C. Gutiérrez-Tobal[1,2] [0000-0002-1237-3424], Víctor Martínez-Cagigal[1,2] [0000-0002-3822-1787], Eduardo Santamaría-Vázquez[1,2] [0000-0002-7688-4258], and Roberto Hornero[1,2] [0000-0001-9915-2570]

[1] Biomedical Engineering Group, University of Valladolid, Valladolid, Spain
[2] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales, Nanomedicina, (CIBER-BBN), Madrid, Spain

Corresponding author: `victor.martinez@gib.tel.uva.es`

Abstract. Predicting users' adherence to digital interventions focused on active and health aging could prevent early dropouts. This is the context of the IFMBE scientific challenge held as part of the 2022 IUPESM World Congress on Medical Physics and Biomedical Engineering. The task is designed as a binary classification problem in which data from different sources of 6 consecutive weeks are used to predict high or low adherence in the next 1.5 weeks. We propose deep learning (residual deep neural networks, DNN) and adaptive boosting (AdaBoost) approaches to solve this issue. Two datasets were used in the challenge though only one was available for training the models. The geometric mean (GE) between sensitivity and specificity in the second unseen dataset was established as the reference score for the challenge. A Residual DNN model reached the highest GE (0.7500) among the ten available attempts. Our analysis suggests that the differences between the training GE and the final score may be due to overfitting caused by noise in the training dataset and to a different data distribution comparing to the unseen test dataset. Additionally, explainable artificial intelligence (XAI) techniques let us point to the days since the last app usage and the number of different days of app usage as key features to predict adherence. XAI let us also uncover the important role of hidden patterns within brain games features. All in all, our results suggest that predicting adherence using the available dataset can be accurately conducted.

Keywords: active aging, digital adherence, digitalization, residual deep learning, adaboost, explainable artificial intelligence.

1 Introduction

Sustained low natality and increased life expectancy are rising population aging in developed countries. According to World Population Prospects 2019 [1], by 2050, 1 in 6 people in the world will be over the age of 65, up from 1 in 11 in 2019, a situation that will pose great challenges in the years to come. The old-age dependency ratio (OADR) is defined as the number of old-age dependent persons (≥65 years old) per 100 persons of working age (aged 20 to 64 years). This metric approximates the implied socioeconomic impact associated with a growing proportion of the population at older ages [1]. OADR is expected to more than double in most regions of the world by 2050, peaking at countries such as Japan (expected OARD of 81 by 2050), Rep. Korea (expected OARD of 79 by 2050) or Spain (expected OARD of 78 by 2050) [1]. These data put in prospective the socioeconomic burden derived from population aging, especially considering future demographic challenges in certain regions.

In this context, developing public health strategies to improve the health state of the elderly and reduce medical expenses must be a priority. The most extended approaches to fight age-related decline are based on active and healthy aging (AHA). The goal of AHA is helping people to stay in charge of their own lives for as long as possible, as well as actively contributing to the economy and society [2]. This multidimensional concept, which is based on promoting an active lifestyle from the physical and mental perspectives, has proven to reduce and delay age-related cognitive and physical decline [2]. Unfortunately, classical AHA interventions, based on personalized therapies conducted by professionals, are unaffordable in the future demographic context.

Digitalized AHA interventions have the potential to reduce the economic burden over healthcare systems, providing a cost-effective strategy to fight against age-related decline [3]. Nevertheless, this type of technological solutions requires high rates of adherence and long-term use to be effective, but users frequently abandon the solution due to a wide range of reasons, such as accessibility barriers or high complexity. In fact, low adherence has been identified as the main limitation of digital AHA interventions to become a successful health strategy [4]. A promising method to address this problem is the identification of users at risk of lower adherence rates and usage patterns that indicate imminent dropout. Their detection can be very useful to apply tailored intervention strategies aimed at recovering from disengagement. However, given the high volume of demographic and usage data produced by each user, advanced data analysis techniques could be a must to predict dropouts reliably.

This study aimed at investigating novel methods to predict dropouts before they happen using advanced machine learning and deep learning models. Concretely, we used different flavors of AdaBoost ensemble, and a deep neural network based on residual connections. The work was part of the IFMBE Scientific Challenge 2022 held as part of the IUPESM World Congress on Medical Physics and Biomedical engineering (IUPESM WC2022). The challenge provided a dataset with demographic and usage data of a digital AHA intervention with the goal of improving both features and models to predict dropout. The challenge had 2 phases. This paper only presents the results of the second phase, where each team had 10 attempts (A1-A10) to reach the maximum accuracy in predicting user dropout.

2 Materials and Methods

2.1 Dataset

According to the Scientific Challenge IUPESM WC2022, the dataset gathers information about the activity of users in a mobile application focused on AHA [5]. The dataset comprises the app activity of more than 300 users that performed an intervention for at least 6 months in the Moving AHA (MAHA) network in Madrid, Spain. Users were asked to use the app frequently (at least twice a week) according to their needs.

Data includes questionnaires and app usage [5]. The former is composed by users answers to different surveys regarding quality of life and acceptance of the MAHA app, both in the baseline T1 (i.e., before using the app) and final evaluation T2 (i.e., after 6 months of usage). As questionnaires, the following 5 datasets were included: (i) sociodemo, sociodemographic characteristics of participants and dates of entering and termination; (ii) self-perception questionnaire (SPQ), overall perception regarding quality of life, physical activity, social life and the impact of the MAHA app in their lives; (iii) unified theory of acceptance and use of technology (UT-AUT), motivation regarding app usage intention and behavior; (iv) EQ-5D-3L, users' health state; and (v) UCLA, evaluation of users' loneliness. On the other hand, data regarding app usage includes logs of the different applications included in the MAHA app during the intervention: (i) brain games, (ii) physical activity exercises, (iii) finger tapping (coordination exercises), (iv) mindfulness app, and (v) digital phenotyping (timestamps about user's usage patterns while navigating).

2.2 Feature extraction

The goal of the Scientific Challenge IUPESM WC2022 was to predict the adherence of the user during the forthcoming 3 scheduled data acquisitions (i.e., 1.5 weeks) given a window of 12 scheduled data acquisitions (i.e., 6 weeks of app usage).

Concerning the feature extraction process, we considered two tricky aspects of the database that are worth to mention. First, although users were asked to use the MAHA app twice a week, many of them did not follow that instruction. Data is highly unstructured, resulting in large temporal windows without any data for most of the applications. This phenomenon encouraged us to develop features to consider the effective time users employed using the app inside the 6 weeks window. Second, a lot of the information provided in the dataset belongs to T2, i.e., post-evaluation features. We think that using these T2 data to train a model to predict the forthcoming adherence would not be a realistic approach as they include information that would not be available in a practical implementation in the training data. Moreover, sociodemo includes a "status" variable that indeed indicates if the user abandoned the app before the experimental period (i.e., dropout), if the experimental period was finished, or even if user is still using the technology.

Next, we summarize the features extracted from questionnaires, apps, and customized metrics, also enumerated in Table 1 with self-explanatory labels. Features that belong to T2 are indicated in italic:

1. **Sociodemo:** gender, age, educational level, technology level, living environment (rural/urban), living status (alone/accompanied), level of physical/cognitive decline ("ucs"), type of device, number of days in program and "*status*".
2. **SPQ:** perception regarding quality of life (Q1), physical activity (Q3), social life (Q5); as well as how the MAHA app modified the previous items at T2 (*Q2, Q4, Q6*).
3. **UT-AUT:** users' motivation at T2 in terms of *effort, performance, attitude, social influence, facilitating conditions, self-efficacy, anxiety,* and *behavioral intention* regarding the MAHA app usage.
4. **EQ-5D-3L:** users' health state in terms of mobility, self-care, usual activities, pain/discomfort, and anxiety/depression for T1 and *T2*.
5. **UCLA:** mean of the UCLA values that measure loneliness both in T1 ("ucla") and T2 ("*ucla_post*").
6. **Brain games:** number of tries for each difficulty (easy, normal, medium, hard), number of successfully solved tasks for each difficulty, average difficulty, and number of different days in which users used this app.
7. **Physical activity:** number of tries and solved tasks involving upper/lower extremities or gait, and number of different days in which users used the app.
8. **Finger tapping:** number of tries, number of solved tasks, number of bilateral (both hands involved) tries; mean and standard deviation reaction time, and mean accuracy across tries; and number of different days in which users used this app.
9. **Mindfulness:** number of tries, number of solved tasks, average duration of tries, and number of different days in which users used this app.
10. **Digital phenotyping:** number of timestamps, number of different days that produced timestamps and number of days since last timestamp.
11. **Additional features:** number of days since the last interaction with the app ("days_since_last"), number of different days in which the user tried any of the applications ("all_independent_days") and number of days that passed since the start of the program ("days_since_start_program").

To extract the training observations, all possible time windows of 7.5 weeks long (6 weeks for training, 1.5 weeks to obtain the adherence label) between each user date of entering and finalization were computed. Features were extracted for each window, and lastly duplicated observations were removed. As indicated in the rules of the Scientific Challenge WC2022, adherence was marked as low (0) if the number of effective acquisitions was less than 2 during the 1.5-week period, and high (1) otherwise. A planned acquisition was effectively implemented by a participant if at least one of the variables scheduled for measurement was received.

The number of features and observations varied between our different attempts. All of them used T1 features, while attempts A5-A7 added T2 features as well. Attempts A7-A10 incorporated features extracted from "digital phenotyping" and "additional features" sources. Details regarding the number of observations and most useful features for prediction will be discussed in the next sections.

Table 1. Summary of extracted features.

Source	Extracted features
Sociodemo	gender, age, educational_level, technology_level, living_environment, living_conditions, living_status, ucs, device, days_in_program, *status*
SPQ	spq_q1, spq_q3, spq_q5, *avg_spq_q2, avg_spq_q4, avg_spq_q6*
UT-AUT	*effort, performance, attitude, social_influence, facilitating_cond, self_efficacy, anxiety, behavioral_intention*
EQ-5D-3L	eq_mobility, eq_selfcare, eq_usualactivities, eq_pain, eq_anxiety, *eq_mobility_post, eq_selfcare_post, eq_usualactivities_post, eq_pain_post, eq_anxiety_post*
UCLA	avg_ucla, *avg_ucla_post*
Brain games	bg_ntries, bg_nsolved, bg_ntries_easy, bg_nsolved_easy, bg_ntries_normal, bg_nsolved_normal, bg_ntries_medium, bg_nsolved_medium, bg_ntries_hard, bg_nsolved_hard, bg_avg_difficulty, bg_days
Physical activity	p_ntries, p_nsolved, p_nupper, p_nupper_solved, p_nlower, p_nlower_solved, p_ngait, p_ngait_solved, p_days
Finger tapping	f_ntries, f_nsolved, f_nbilateral, f_avg_meanrt, f_avg_stdrt, f_avg_acc, f_days
Mindfulness	mind_ntries, mind_nsolved, mind_avgduration, mind_days
Digital phenotyping	accesses_pheno, days_phenotyping, days_since_last_pheno
Additional features	all_independent_days, days_since_last, days_since_start_program

* T2 features are displayed in italic.

2.3 Models: Residual Deep Learning and AdaBoost

Model 1. Residual Deep learning. Deep learning has been developed in the last years as the state-of-the-art technique for classification tasks in a wide variety of fields, like computer vision or natural language processing [6]. The deep neural network (DNN) that we implemented includes the mechanism of residual connections [7] that allows the information for traveling through the layers of the DNN. It also makes use of bottleneck features as presented in the ResNeXt architecture [8] as a regularization mechanism in the residual connections to prevent overfitting. Other regularization elements of the Residual DNN are the use of dropout, global average pooling before the final projection head used for classification, and early stopping of the training when validation loss does not improve after several epochs. As shown in Fig. 1, the DNN is structured as follows:

- **Stem block.** First layer of the network with a convolution with filter size of 3 and stride of 2 that reduces the dimensionality of the input features by the stride value and creates 50 filters with those condensed features. The convolution is followed by batch normalization, and *elu* activation function to introduce non-linearity and spatial dropout for regularization.

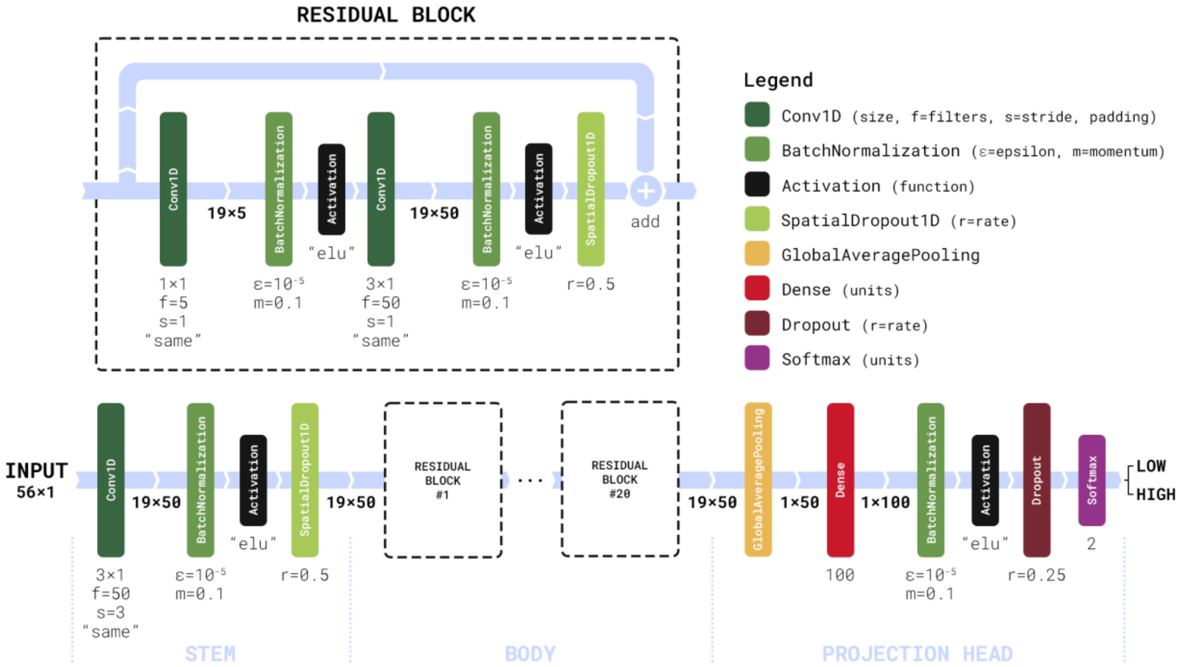

Fig. 1. Structure of the Residual DNN approach (attempt A10), composed by three main parts: (1) stem, which takes the data from 6 weeks, reduces the dimensionality and increments the number of channels; (2) body, which applies 20 residual blocks and ends keeping one value for each filter; and (3) projection head, which projects the features in dense layers to return the prediction for the next 1.5 weeks as low (0) or high (1) adherence.

- **Body.** It is composed of connections of residual blocks that includes 2 consecutive convolution operations each. The first one is the bottleneck and has filter size of 1 with only 5 filters, reducing by 10 the number of filters of the previous layers. It is followed by batch normalization and *elu* activation. Then, these bottleneck features are expanded with a convolution of filter size of 3 and 50 filters. This convolution is followed by batch normalization and *elu* activation, but also by spatial dropout which makes inactive some filters in training. The values of these 50 filters are added to the values of the 50 filters before the bottleneck operation, completing the residual operation. These residual blocks are repeated 20 times. At the end of the body, a global average pooling is performed to keep one value from each filter.
- **Projection head.** All the 50 features extracted in the body of the DNN are then connected to a dense layer with double the features. The new dense layer is followed by batch normalization, *elu* activation and dropout to finally be classified by a softmax layer as ‘HIGH’ (1) or ‘LOW’ (0) adherence.

In total, four models were trained with this method: one with 41 features (attempt A2), one with 73 features (A7), and two with 56 features (A8 and A10).

Model 2. AdaBoost. Adaptive boosting (AdaBoost) is a common yet successful choice when addressing classification problems that was originally developed by Freund and Schapire [9]. It is an ensemble-learning boosting method in which multiple base (or 'weak') classifiers of the same type are combined so that each new learner complements the predictions conducted by the learners from previous iterations [10]. This is achieved by re-weighting the observations miss-classified in past runs, thus providing them with higher chances to be rightly classified in the current iteration. The final classification task is conducted based on the vote of all classifiers, these weighted votes contributing more as the errors of the corresponding classifiers are lower [10]. In this study, we used decision stumps as 'weak' classifiers as recommended to minimize overfitting [11]. It also provides us with the ability to conduct an automatic de facto feature selection as only one feature is used at each iteration, while easing the measurement of the contribution of each feature to the final classification. One model was trained using this method and 13 features (attempt A9).

Model 3. Gentle AdaBoost. Gentle AdaBoost is an AdaBoost improvement developed by Friedman et al. [12] to reduce the generalization error by introducing Newton stepping [11], [12]. Five models were trained using this method: one with 35 features (attempt A1), two with 41 features (A3 and A4), and 2 with 61 features (A5 and A6).

2.4 Statistical analysis, validation strategy, and explainable artificial intelligence

In accordance with the rules of the challenge, the geometric mean (GE) of the sensitivity (Se) and specificity (Sp) was used as the reference to measure the performance of our models:

$$GE = \sqrt{Se \cdot Sp}, \tag{1}$$

where Se accounts for the percentage of rightly classified high adherent observations and Sp accounts for the percentage of rightly classified low adherent observations. In order to estimate the possible GE to be reached in the unseen dataset, the available data was used along with different validation strategies. A bootstrap 0.632+ procedure was used to estimate the GE in the AdaBoost and Gentle AdaBoost models while varying the number of stumps in the ensemble [13]. As this method is computationally exhaustive for the Residual DNN algorithm, a 5-fold cross-validation methodology was implemented instead. Additionally, two explainable artificial intelligence (XAI) techniques were used. Relative importance ($\hat{I}^2$) was used to measure the contribution of each feature to the decisions of the AdaBoost and Gentle Adaboost models [14], as it is an easy choice when analyzing decision trees like stumps. Similarly, SHAP values (for SHapley Additive exPlanations) were used to mimic this analysis in the Residual DNN models [15].

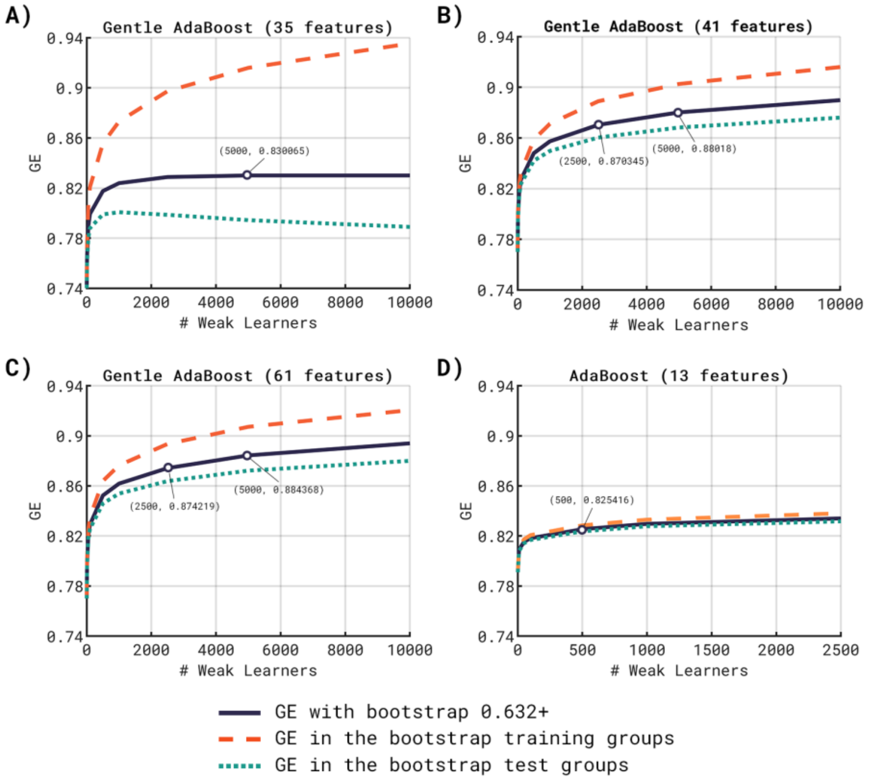

Fig. 2. GE estimation and hyperparamenter optimization for all the ensemble-learning methods used in the study. A) Gentle AdaBoost with 35 features (attempt A1). B) Gentle AdaBoost with 41 features (A3 and A4). C) Gentle AdaBoost with 61 features (A5 and A6). D) AdaBoost with 13 features (A9).

3 Results

3.1 Ensemble learning hyperparameter optimization

An optimization process was conducted along with the GE estimation in order to select the number of weak learners for the ensemble-learning methods. As mentioned above, a bootstrap 0.632+ method was used for this purpose. Accordingly, 100 bootstrap replicates of the original database were formed by resampling with replacement, thus deriving in 100 new bootstrap-based training sets and their corresponding test sets. Fig. 2 displays the averaged GE for these training and test groups for increased numbers of learners in each ensemble-learning method used. The GE composite estimation is also shown, which is weighted according to the original 0.632+ method by Efron and Tibshirani [13]. Comparing these curves, higher estimated GEs are obtained with Gentle

Table 2. Test results for all phase II attempts.

Attempt	Model	Features	# obs.	# learners/ residual blocks*	Test score (GE)
A1	Gentle AdaBoost	35	4339	5000	0.5798
A2	Residual DNN	41	11404	10	0.7142
A3	Gentle AdaBoost	41	11404	2500	0.6910
A4	Gentle AdaBoost	41	11404	5000	0.6910
A5	Gentle AdaBoost	61	11404	2500	0.5805
A6	Gentle AdaBoost	61	11404	5000	0.5344
A7	Residual DNN	73	12611	30	0.6750
A8	Residual DNN	56	12611	20	0.7493
A9	AdaBoost	13^{+}	12611	500	0.6847
A10	**Residual DNN**	**56**	**12611**	**20**	**0.7500**

* Num. learners (for AdaBoost attempts) or residual blocks (for Residual DNN attempts).

$^{+}$ The 13 features were selected from the 73 features set after observing the relative importance of each of them in a tentative AdaBoost model trained with 1000 learners. In order to try to reduce overfitting, only the 13 features that gathered the 95% of the relative importance were included in the model.

AdaBoost for 41 and 61 features. However, AdaBoost (13 features) shows significantly less difference between the averaged training and test GEs, maybe being an early indicator of less overfitting.

3.2 Classification performance

Table 2 displays the results achieved for each attempt of phase II in terms of GE. The machine-learning method, the number of features used for its training, the number of training observations (in accordance with these features), and the number of weak learners (if ensemble method) or residual blocks (if DNN) are also shown. As observed, the Residual DNN model from attempt A10 reached the highest performance. In general, the Residual DNN method outperformed the ensemble-learning approach either Gentle or plain AdaBoost.

One of the main differences between attempts resulted from the changes in the feature extraction process. The increase in performance from A1 to A4 was achieved after adding the features that include the number of independent days in which the applications were used, as well as the number of days since the last use of any application (e.g., “bg_days”, “p_days”, “f_days”, “mind_days”, “all_independent_days”, and “days_since_last”). Then, the drop in performance in A5, A6 and A7 was observed after including the information collected at the end of the program (T2 features). This information was not used at the beginning of the phase II as it would not be useful to detect high or low adherence in a real scenario. The real purpose of the classification task, as defined in the challenge rules, is to predict the adherence of users when the experiment is ongoing, and this T2 information would be only available at the end of the process. Accordingly, in the last attempts we started from the 41 features present in attempts

A2-A4 but separating the information of the different games in difficulties or types, aggregating the number of days that the user has been on the program, and adding the information from phenotyping. The minor difference between A8 and A10 is only due to a change in the stride of the stem convolution of the Residual DNN, which is 3 in A10 and 2 in A8.

Regarding the ensemble-learning methods, the comparison between the results in Table 2 and the hyperparameter optimization process shown in Fig. 2, confirms that these models experiment a high degree of overfitting. Interestingly, AdaBoost with only 13 features and 500 learners reached almost the same performance that Gentle AdaBoost with 41 features and 2500 to 5000.

3.3 Explaining the machine learning models

The most efficient ensemble-learning model in terms of number of features used, number of learners, and GE scored in the test set was AdaBoost (A9). Accordingly, we used it to estimate the importance of the features to predict user's adherence by computing $\hat{I}^2$ normalized to 100 [14]. In order to get a similar measure of the importance in the predictions of the features from the best Residual DNN model (A10), we used the SHAP approach, which makes use of the shapley values [15]. In this case, a kernel explainer method was used to obtain the order of features according to its contribution to the decision on users' adherence [15]. Fig. 3 displays the relative importance of the features in the AdaBoost model and the SHAP value of the features of the Residual DNN. In the latter case, only the 15 features (out of 56) with the highest SHAP value are shown for simplicity. Although the initial feature set for the two models is not the same, up to 5 features are in both importance rankings. Moreover, the two models coincide in giving the maximum importance to the number of days since the last interaction of the user with the app. In addition, the number of different days in which the app has been used (within the 6 considered weeks for each observation) ranked high in both models (third in AdaBoost and tenth in Residual DNN). This finding agrees with the increase in performance shown in models from attempts A2-A4, where these features were included, compared with A1. An interesting difference between the models is the importance given by Residual DNN to the information from brain games (8 features), which is not equal in AdaBoost (2 features).

4 Discussion

In this study, we have developed up to 10 machine-learning models from three different approaches (Residual DNN, AdaBoost, and Gentle AdaBoost) to classify the MAHA dataset observations into high or low adherence. High maximum GE was reached, thus showing not only high performance in the classification task but also a compromise between sensibility and specificity. The greatest increase in performance among the attempts was achieved after careful feature engineering from the original data, as well as using deep learning instead of the more traditional ensemble learning methods.

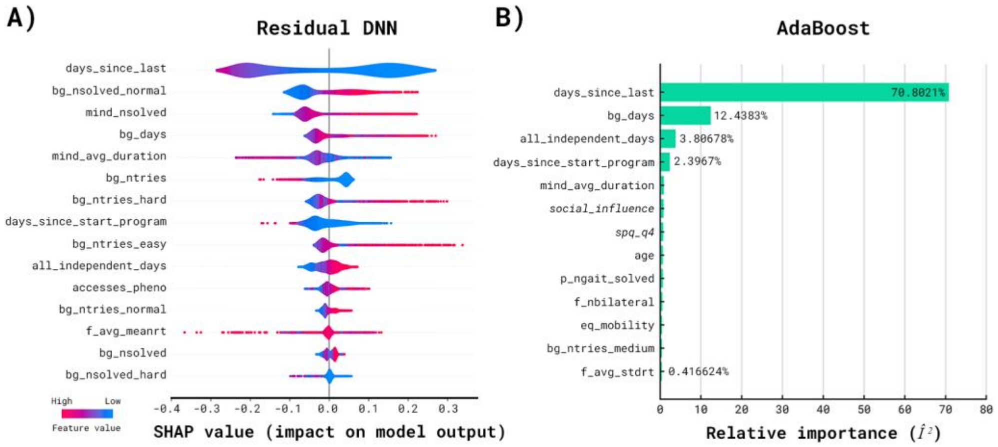

Fig. 3. A) SHAP values of the 15 most relevant features for the DNN (attempt A10), sorted in descending order of importance. Blue and red colors indicate low and high feature value, respectively. SHAP value (in X axis) indicates whether the feature with low/high value is related to HIGH adherence if positive (i.e. $x > 0$), and low adherence if negative (i.e. $x < 0$). B) Normalized relative importance of the 13 most relevant features for the AdaBoost (attempt A9) in descending order.

When developing the models, a first important general issue was the observed difference between the estimated GE in the available dataset and the score achieved in the unseen test set, the latter being remarkably lower. One possible explanation for this performance decrease is overfitting. Despite using bootstrap 0.632+ to estimate GE, both ensemble-learning approaches suffered from this problem to some extent. Gentle AdaBoost was originally developed as an improvement of plain AdaBoost with the significant drawback of being particularly sensible to overfitting in the presence of noisy datasets [11]. Accordingly, one would expect the training dataset to be affected by a substantial amount of noise. This idea would be also supported by the need of several regularization methods when developing the Residual DNN. We had to use the bottleneck convolutions in the residual operations [8], a stride of the same size as the kernel in the stem layer, a spatial dropout, a global average pooling operation, and the use of the early stopping technique. Another possible reason for the lower score achieved in the test dataset would be some differences in its characteristics (e.g., classes distribution) comparing to the training dataset. As plain AdaBoost is well-known for its generalization ability [10], the remarkable lower score in the test set could be not only due to overfitting but also to possible differences in the characteristics of both datasets. However, future analysis would be needed to assess both these differences and the presence of noise in the training data. Finally, we also had to cope with class imbalance in the dataset derived from extracting observations from each independent day. Accordingly, we first discarded any observation where all features (apart from the days in the program) were repeated. This greatly reduced the size of the dataset, but we ended up with data of greater quality, as identified in our exploratory experiments. Moreover, when training the Residual DNN and AdaBoost models, we further balanced the classes weights in the loss function.

The use of the XAI methods (SHAP and relative importance) let us infer interesting additional information regarding the features used and the way in which the models conduct their decisions. Accordingly, the initial intuition that a higher number of days since the last use of any app ("days_since_last") is useful to predict low adherence is corroborated by the negative SHAP value and the corresponding red color in the Residual DNN, as well as the high relative importance reached in AdaBoost. Similarly, showing a higher number of days where the user opens the application ("all_independent_days") implies that the models will tend to predict higher adherence. Additionally, less evident patterns in the features extracted can be also suggested. For example, a high number of brain games solved in hard and normal difficulties will make the model predict respectively low and high adherence, suggesting that having games that are not challenging enough for participants will result in lower adherence to the program. Interestingly, the longer the mindfulness app duration, the lower predicted adherence. This may suggest that a balance needs to be maintained between difficulty and user engagement. Of note, the fact that the Residual DNN model responsible for the best attempt includes up to 8 features from brain games highly ranked, and AdaBoost only 2, could be suggesting the importance of this kind of information. It could be also indicating that the convolutional layers of the DNN are obtaining hidden adherence-related patterns from these features. In contrast, AdaBoost would not be able to take advantage of this hidden information because of its use of features in a more individual way. This behavior of the Residual DNN would be highlighting one of the main advantages of deep-learning techniques, that is, obtaining useful information beyond the assumptions of human beings on data. As observed, the use of XAI methods increased the usefulness of the models by providing insights about the classification task beyond predicting the users' adherence. In this case, by uncovering not evident yet useful information about the role of brain games.

A final remark can be done regarding the decrease in performance observed when including the T2 features (attempts A5, A6, A7, and A9). This could be suggesting that this information is not important or even is missing in the test set. The idea would be also supported by the relative importance shown in Fig 3.B, where 2 of these features are ranked 5th and 6th on the training dataset. The unimportant nature of these features for predicting adherence in the test set would support our assumption that they should not be used, while highlighting the usefulness of our proposal in a realistic context.

5 Conclusions

Residual DNN outperformed Gentle and plain AdaBoost when predicting adherence to a digital intervention focused on active and health aging. High values in the geometric mean of sensitivity and specificity were reached, thus highlighting the concurrent usefulness of the data and the model. The differences in the scores of the available and unseen datasets may indicate high noise in the first one and a different data distribution in the second one. Moreover, T2 features showed no usefulness to generalize the models in the unseen dataset. In contrast, XAI techniques showed that the number of days since the last use of the app and the number of different days in which the users open the app play key roles in the decisions taken by the models. Similarly, XAI techniques uncovered the importance of the features obtained from the brain games. All in all, our results suggest that predicting adherence using the MAHA dataset can be accurately conducted.

Acknowledgments

This research has been developed under the grants PID2020-115468RB-I00, RTC2019-007350-1 and PDC2021-120775-I00 funded by 'Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación/10.13039/501100011033/' and European Regional Development Fund (ERDF) A way of making Europe; under the R+D+i project 'Análisis y correlación entre la epigenética y la actividad cerebral para evaluar el riesgo de migraña crónica y episódica en mujeres' ('Cooperation Programme Interreg V-A Spain-Portugal POCTEP 2014–2020') funded by 'European Commission' and ERDF; and by 'Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN)' through 'Instituto de Salud Carlos III' co-funded with ERDF funds. E. S.-V. and S.P-V. were in receipt of a PIF grant by the 'Consejería de Educación de la Junta de Castilla y León'.

References

[1] United Nations, "World Population Ageing 2019: highlights," 2019.

[2] Deary, I. J., et al. (2009). Age-associated cognitive decline. British medical bulletin, 92(1), 135-152.

[3] Parra, C., et al. (2014). Information technology for active ageing: A review of theory and practice.

[4] Law, C. K., et al. (2020). Physical exercise attenuates cognitive decline and reduces behavioural problems in people with mild cognitive impairment and dementia: a systematic review. *Journal of physiotherapy*, 66(1), 9-18.

[5] Fico, et al. (2022). The MAHA dataset: understanding and improving adherence to digital interventions for Active and Healthy Ageing. *Presented at IUPESM World Congress on Medical Physics and Biomedical Engineering*, Singapore.

[6] Vaswani, A., et al., (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[7] He, K., et al., (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

[8] Xie, S., et al., (2017). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492-1500.

[9] Freund, Y., and Schapire, R. E., (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.

[10] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann.

[11] Wu, S., and Nagahashi, H., (2014). A new method for solving overfitting problem of gentle AdaBoost. *Fifth International Conference on Graphic and Image Processing (ICGIP 2013)*. 9069, SPIE.

[12] Friedman, J., Hastie, T., and Tibshirani, R., (2000). Additive logistic regression: a statistical view of boosting. *The annals of statistics* 28(2), 337-407.

[13] Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438), 548-560.

[14] Friedman, J., and Meulman, J., (2003). Multiple additive regression trees with application in epidemiology. *Statistics in medicine*, 22(9), 1365-1381.

[15] Lundberg, S. M., and Lee, S. I., (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.