Computer Methods and Programs in Biomedicine 269 (2025) 108899

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: https://www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine

Stabilizing machine learning for reproducible and explainable results: A novel validation approach to subject-specific insights

Gideon Vos [a], Liza van Eijk [b], Zoltan Sarnyai [c], Mostafa Rahimi Azghadi [a,*]

[a] College of Science and Engineering, James Cook University, James Cook Dr, Townsville, 4811, QLD, Australia
[b] College of Health Care Sciences, James Cook University, James Cook Dr, Townsville, 4811, QLD, Australia
[c] College of Public Health, Medical, and Vet Sciences, James Cook University, James Cook Dr, Townsville, 4811, QLD, Australia

ARTICLE INFO

Dataset link: https://github.com/xalentis/Reproducibility

MSC:
68T01
92-08

Keywords:
Machine learning
Reproducibility
Explainable A.I.
Precision medicine

ABSTRACT

Introduction: Machine Learning (ML) is transforming medical research by enhancing diagnostic accuracy, predicting disease progression, and personalizing treatments. While general models trained on large datasets identify broad patterns across populations, the diversity of human biology, shaped by genetics, environment, and lifestyle, often limits their effectiveness. This has driven a shift towards subject-specific models that incorporate individual biological and clinical data for more precise predictions and personalized care. However, developing these models presents significant practical and financial challenges. Additionally, ML models initialized through stochastic processes with random seeds can suffer from reproducibility issues when those seeds are changed, leading to variations in predictive performance and feature importance. To address this, this study introduces a novel validation approach to enhance model interpretability, stabilizing predictive performance and feature importance at both the group and subject-specific levels.

Methods: We conducted initial experiments using a single Random Forest (RF) model initialized with a random seed for key stochastic processes, on nine datasets that varied in domain problems, sample size, and demographics. Different validation techniques were applied to assess model accuracy and reproducibility while evaluating feature importance consistency. Next, the experiment was repeated for each dataset for up to 400 trials per subject, randomly seeding the machine learning algorithm between each trial. This introduced variability in the initialization of model parameters, thus providing a more comprehensive evaluation of the machine learning model's features and performance consistency. The repeated trials generated up to 400 feature sets per subject. By aggregating feature importance rankings across trials, our method identified the most consistently important features, reducing the impact of noise and random variation in feature selection. The top subject-specific feature importance set across all trials was then identified. Finally, using all subject-specific feature sets, the top group-specific feature importance set was also created. This process resulted in stable, reproducible feature rankings, enhancing both subject-level and group-level model explainability.

Results: We found that machine learning models with stochastic initialization were particularly susceptible to variations in reproducibility, predictive accuracy, and feature importance due to random seed selection and validation techniques during training. Changes in random seeds altered weight initialization, optimization paths, and feature rankings, leading to fluctuations in test accuracy and interpretability. These findings align with prior research on the sensitivity of stochastic models to initialization randomness. This study builds on that understanding by introducing a novel repeated trials validation approach with random seed variation, significantly reducing variability in feature rankings and improving the consistency of model performance metrics. The method enabled robust identification of key features for each subject using a single, generic machine learning model, making predictions more interpretable and stable across experiments.

Conclusion: Subject-specific models improve generalization by addressing variability in human biology but are often costly and impractical for clinical trials. In this study, we introduce a novel validation technique for determining both group- and subject-specific feature importance within a general machine learning model, achieving greater stability in feature selection, higher predictive accuracy, and improved model interpretability. Our proposed approach ensures reproducible accuracy metrics and reliable feature rankings when using models incorporating stochastic processes, making machine learning models more robust and clinically applicable.

* Corresponding author.
E-mail address: mostafa.rahimiazghadi@jcu.edu.au (M. Rahimi Azghadi).

https://doi.org/10.1016/j.cmpb.2025.108899
Received 16 December 2024; Received in revised form 9 April 2025; Accepted 5 June 2025
Available online 21 June 2025

1. Introduction and related work

Machine learning (ML) and other forms of Artificial Intelligence (AI) have emerged as transformative statistical tools that are revolutionizing various fields of science, including medical research. These technologies excel at identifying complex patterns in vast datasets, often revealing insights that remain elusive to human researchers. In the realm of medicine, ML is increasingly utilized for tasks such as drug discovery [1–3], patient diagnostics [4–6], and the analysis of genomic data [7–9].

Despite the growing adoption of machine learning (ML) in medical research, its real-world effectiveness remains limited by challenges in model generalizability and reproducibility [10,11]. While ML models trained on large datasets can identify broad population-level patterns, they often struggle to adapt to individual patient differences driven by genetics, environment, and lifestyle [10,12,13]. This lack of adaptability leads to inconsistent clinical predictions, reducing the reliability of ML-based tools in personalized medicine [14–16].

Reproducibility remains a major issue, as even minor modifications to model parameters, dataset partitions, or random seed initialization can result in drastically different outcomes [17]. Without rigorous validation techniques, ML models risk producing misleading results that fail to hold up under independent verification. A study by Kapoor et al. [15] found that among 294 medical ML studies, many suffered from data leakage, lack of independent validation, and overoptimistic performance reporting, further exacerbating the reproducibility crisis.

Additional challenges during model building include class imbalance, outlier management and bias mitigation. As a result, there is an urgent need for validation frameworks that enhance the stability and interpretability of ML models, ensuring that predictions remain reliable across diverse clinical settings. Such inconsistencies raise concerns about the reliability of ML-based clinical decision support systems, particularly when feature importance rankings vary significantly between experiments.

As researchers increasingly rely on ML to guide critical decisions in healthcare, the need for robust validation, model interpretability, and the ethical application of these technologies becomes paramount to ensure that the advancements they offer are both meaningful and beneficial. A 2023 survey by Nature [17] found that while the use of ML is becoming increasingly common in science, 58% of the 1600 respondents raised concerns that ML techniques can introduce bias or discrimination in data, while 53% noted that ill-considered use can lead to non-reproducible research.

Explainable AI (XAI) holds significant promise in making machine learning systems more transparent, trustworthy, and actionable [18,19], potentially mitigating some of these reproducibility issues by clarifying model decision-making processes. Yet, the influence of ML recommendations on physician behavior remains poorly characterized. Nagendran et al. [20] investigated how clinicians' decisions may be influenced by additional information provided by XAI techniques and found that ML-generated explanations had a strong influence on medical prescriptions. However, clinicians, researchers, and regulators alike demand that XAI not only provide recommendations but also justify its algorithmic reasoning to ensure clinical confidence in its outputs [20].

As machine learning models become more complex, the need for interpretability becomes even more critical in fields such as healthcare, where decisions based on these models can directly influence patient outcomes. Many traditional validation techniques, such as cross-validation and train–test splits, fail to provide sufficient stability in feature importance rankings or performance metrics across trials [21]. To address this challenge, XAI supports two approaches to ensure explainability [21]: (i) ante-hoc explainability, which involves constructing models that are inherently transparent, and (ii) post-hoc explainability, which seeks to provide insights into complex, "black-box" models after they have been trained. Post-hoc techniques, such as feature importance analysis or visualization tools, aim to shed light on how these models arrived at their predictions, making them more interpretable without sacrificing performance. However, the effectiveness of these methods is often undermined by instability in feature selection due to random variations in model initialization and dataset partitioning [22].

A prerequisite of using XAI effectively for explainability is robust model generalization. Generalization refers to a model's ability to perform well on new, unseen data, ensuring that it captures the core patterns of the problem domain rather than overfitting to the specifics of the training set. Without proper generalization, interpretability loses value and offers little meaningful insight. However, the notion of generalization itself may vary by context and demographics. A system that achieves the highest possible level of generalization is ideal, but an emphasis on overly broad generalization in healthcare applications may overlook crucial patient-specific variations, reducing its clinical utility [22].

Several factors can impact a model's ability to generalize effectively across different contexts, collectively posing significant barriers to reproducibility. These include changes in clinical practice over time, patient demographic variation, and differences in hardware and software used both for data collection and model training [22]. Additional challenges during model building include class imbalance, outlier management, bias mitigation, and potential training data leakage, where information from the test or validation dataset is inadvertently included in the training dataset, leading to artificially inflated performance. Kapoor et al. [15] found that data leakage alone was responsible for overestimating ML model performance in numerous medical studies. Such inconsistencies raise concerns about the reliability of ML-based clinical decision support systems, particularly when feature importance rankings vary significantly between experiments.

The primary objective of this study is to systematically address these issues by developing a validation method that enhances both reproducibility and model interpretability. Specifically, we aimed to reproduce the findings of a previously published study [23], which provided source code, hardware and software specifications, and data via the Yale Open Data Access (YODA) Project. By leveraging multiple random trials with varying seed values, we evaluated how random initialization influences model performance and feature importance rankings.

Prior studies [24,25] have demonstrated that changing the random seed can lead to significant differences in model outputs, sometimes inflating performance estimates by up to twofold [24]. Our analysis confirmed these findings, specifically when using machine learning models that incorporate stochastic processes during initialization, resulting in both performance and feature importance sensitivity and inconsistency. This variability poses a major challenge in clinical settings, where consistent and reliable predictions are critical for decision-making. Inconsistent feature importance rankings may undermine clinicians' confidence in AI-driven insights, limiting adoption in real-world healthcare applications.

To address this, we introduce a novel random trial validation method that systematically varies random seeds across multiple iterations, aggregates feature rankings and stabilizes both model accuracy and feature selection. By stabilizing feature importance rankings across multiple randomized trials, the method enables clinicians to identify the most relevant biomarkers or predictors for disease classification, diagnosis, and treatment planning. This improved interpretability can help bridge the gap between ML models and clinical expertise, allowing practitioners to trust AI recommendations rather than viewing them as "black-box" outputs. Furthermore, trustworthy ML models have the potential to support personalized medicine, where treatment decisions are

Table 1
Datasets utilized in this study.

Dataset	Sample size	Features	Ordinals	Cardinality	Domain
1. YODA RCT [23]	1513	–	–	–	Medical
2. Breast Cancer [28]	683	10	10	91	Medical
3. Diabetes [29]	351	35	3	8150	Medical
4. College [30]	777	18	1	6249	Non-medical
5. Cars [31]	32	11	5	171	Non-medical
6. Glaucoma [32]	196	63	1	8960	Medical
7. Glass [33]	214	10	1	945	Non-medical
8. Diamonds [34]	250	10	3	544	Non-medical
8. Diamonds [34]	500	10	3	737	Non-medical
8. Diamonds [34]	2000	10	3	1273	Non-medical
8. Diamonds [34]	5000	10	3	2077	Non-medical
9. Alzheimer's Disease [35]	48	23	–	448	Medical

tailored to individual patients based on reproducible and interpretable AI-generated insights. By improving model transparency, our approach can facilitate regulatory approval of AI-based decision support systems and accelerate their integration into clinical workflows.

2. Methods

2.1. Reproducibility

The datasets used in this study are listed in Table 1. For the first experiment, five international randomized controlled trial (RCT) datasets for evaluating the comparative efficacy of anti-psychotic medications for treating schizophrenia were utilized, as per the original study [23] published in the journal *Science*. These datasets are available from the YODA project as accession numbers NCT00518323, NCT00334126, NCT00085748, NCT00078039, and NCT00083668.

The aim of the first experiment was to reproduce the findings from the original study by re-running the pre-processing, analysis and visualization routines provided using the R source code made publicly available [26]. Due to the resource intensive requirements of the original pre-processing and model building routines, only the Random Forest (RF) [27] models were selected and run for a single set of outcome criteria, the Remission in Schizophrenia Working Group (RSWG).

Once the initial analysis was completed, all random seeds in the supplied source code were changed to a single number (42), and the process was re-run to test for results stability. We observed that changes in the random seed affected both feature importance and performance. To further evaluate our findings, we extended experimentation to additional well-studied and diverse public datasets (Table 1, 2–8).

2.2. Effect of random seed and validation techniques on performance and feature importance

The next set of experiments utilized the RF algorithm in order to build models for predicting the relevant binary outcome labeled within each dataset. These experiments were designed to evaluate how altering random seeds during the initialization of the RF algorithm influences accuracy metrics and the variance in feature importance reported by the model. Furthermore, the effect of different validation techniques on model performance and feature importance was examined.

The datasets shown in Table 1 were selected due to the varying sample sizes, feature attributes, and accessibility as open datasets. Table 1 further details the number of features, ordinal variables (categorical variables with a meaningful order), and cardinality (the number of unique values a categorical variable can take) for each of the nine datasets. These attributes were included in the table to highlight that their counts had no observable correlation with the stability of feature importance during subsequent experiments and provided sufficient variation for results comparison. All experimentation was done using R [36] version 4.4.1.

For each experiment and related dataset, a number of validation techniques were applied including an 80%/20% train and test split, leave-one-subject-out (LOSO) validation, 10-fold cross validation, and leave-one-out cross-validation (LOOCV), with each training and validation round repeated using two different random seeds to initialize the RF algorithm (42 and 43). The selection of validation techniques was applied to verify whether any specific technique provides more stable predictive accuracy and feature importance ranking and are considered standard techniques typically applied when training and validating machine learning model performance. For experimentation we selected two random seeds (42 and 43), 42 being a composite and even, and 43 being a prime and odd.

2.3. Proposed randomized trials validation approach

The RF algorithm uses a user-specified random seed during bootstrap sampling and random feature selection to create multiple subsets of the training data [37] and ensure reproducibility between model training sessions [37,38]. However, a study by Henderson et al. [39] found that altering the random seed could inflate the estimated model performance by as much as 2-fold, relative to what a different set of random seeds would yield. Additionally, Peng et al. [40] noted that system-specific factors including software library versions and hardware specifications can influence the consistency of results when machine learning models are re-run, by potentially impacting the underlying random number generator. Initial testing using both SHAP [41] and LIME [42] as potential feature importance calculators showed both methods are sensitive to varying random seeds. Given this and the longer processing time required for calculations, we reverted to using built-in feature importance methods provided by standard RF algorithm.

In order to compensate for these noted limitations related to random seed selection and its potential impact on reproducibility, we addressed the problem through a repeated randomized validation method as detailed in Fig. 1.

The proposed method first splits the training dataset by subject (see the first column in Fig. 1). For each of the 400 trials per subject (Fig. 2), the system random number generator which is used by the model for key stochastic processes is initialized using the sum of the subject index and trial number. A Random Forest model is then trained using stratification on data from all other subjects and tested on the current subject (LOSO validation), and repeated for up to 400 trials. Note that experimentation using trial counts ranging from 50 to 1000 showed an optimal maximum of 400, at which point feature importance stabilized and no longer showed variance. If the model correctly predicts the outcome in each trial, the most important features are recorded. After completing all 400 trials for a subject (column 2 in Fig. 1), the recorded feature importance sets across the 400 trials (column 3 in Fig. 1) are grouped and ranked using aggregation (column 4 in Fig. 1) to identify the top sets that contributed most to achieving 100% accuracy, per

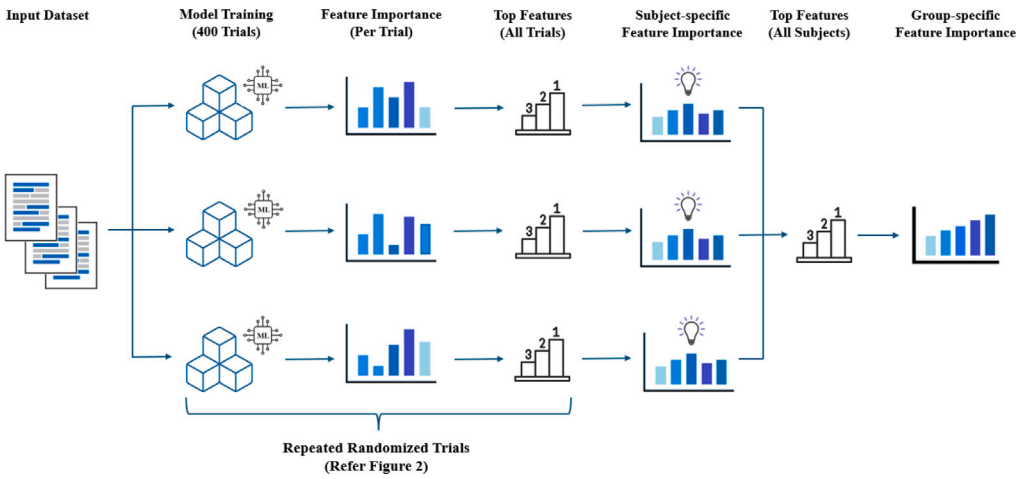

Fig. 1. Proposed randomized trial validation approach for subject- and group-specific feature importance and model performance stabilization.

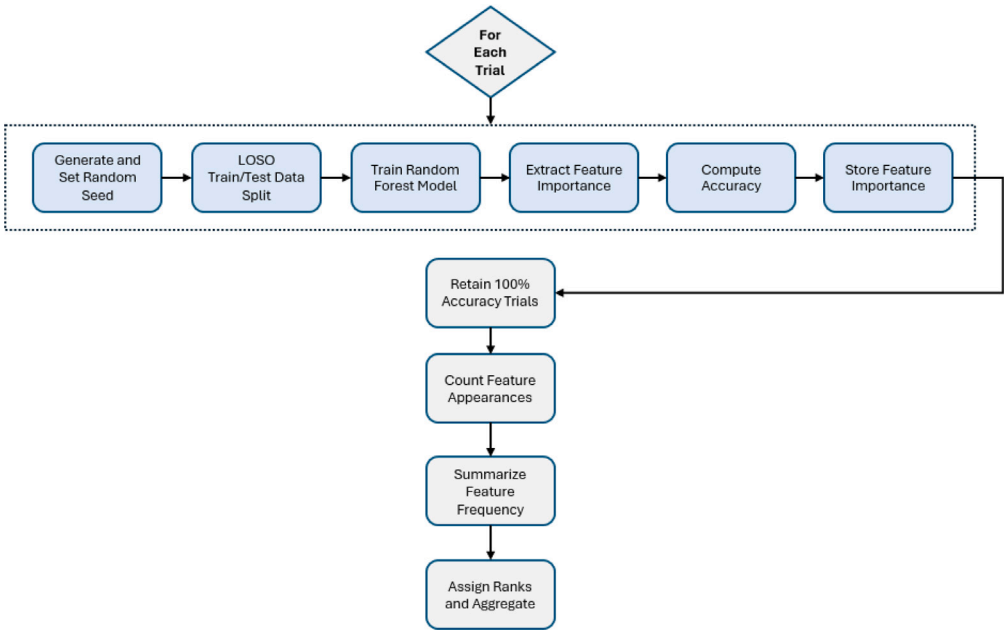

Fig. 2. Proposed randomized trial process for feature importance ranking.

subject (column 5 in Fig. 1). Upon completion of all subjects across all trials, the same ranking method (column 6 in Fig. 1) is applied to find the feature sets that occur most often across the dataset. This specifies the overall feature importance for the group (column 7 in Fig. 1).

Although the random seed in each trial is initialized using the sum of the subject index and trial number, potential concerns about systematic bias across subjects due to seed proximity are naturally mitigated by utilizing LOSO validation. Specifically, each subject's model is trained only on data from other subjects, and the random seed influences only the internal stochastic processes of the model. As the training data for each subject is entirely disconnected from their own data, the seed's effect is isolated within models that do not see the test subject during training, preventing cross-subject contamination. The use of 400 independent trials per subject further introduces a wide range of random initialization states, ensuring variability across experiments. To enhance bias-mitigation, a hashing function can further be utilized on the combination of subject index and trial number.

The proposed method was tested with datasets 1 to 8 (Table 1) to confirm its effectiveness across diverse domains, beyond just medical datasets. Additionally, for dataset 8, multiple sample sizes were selected and run to compare results on a single dataset of varying sizes.

Finally, to validate the use of the proposed method to achieve stable performance metrics and feature importance (per subject and per group) in a medical dataset, the dataset from [35] was utilized. This dataset 9 consists of 34 healthy controls and 14 subjects identified as having Alzheimer's disease, and was previously analyzed [43] to assess the relative significance of clinical observations, neuro-psychological tests, and specific blood plasma biomarkers (inflammatory and neurotrophic) [35].

3. Results

3.1. Challenges in reproducibility

Fig. 3 shows a comparison between the results of two scenarios (Within-trial no validation and Leave-one-trial-out) reported in the original study by Chekroud et al. [23], and those found in our reproduction (as explained in Section 2.1) after adjusting the original random seed numbers in the published source code [26]. For the within-trial no validation scenario, we note a substantial difference in both Chronic #2 and Older Adult subsets, with less substantial differences in the Leave-one-trial-out scenario. However, balanced accuracy reported across

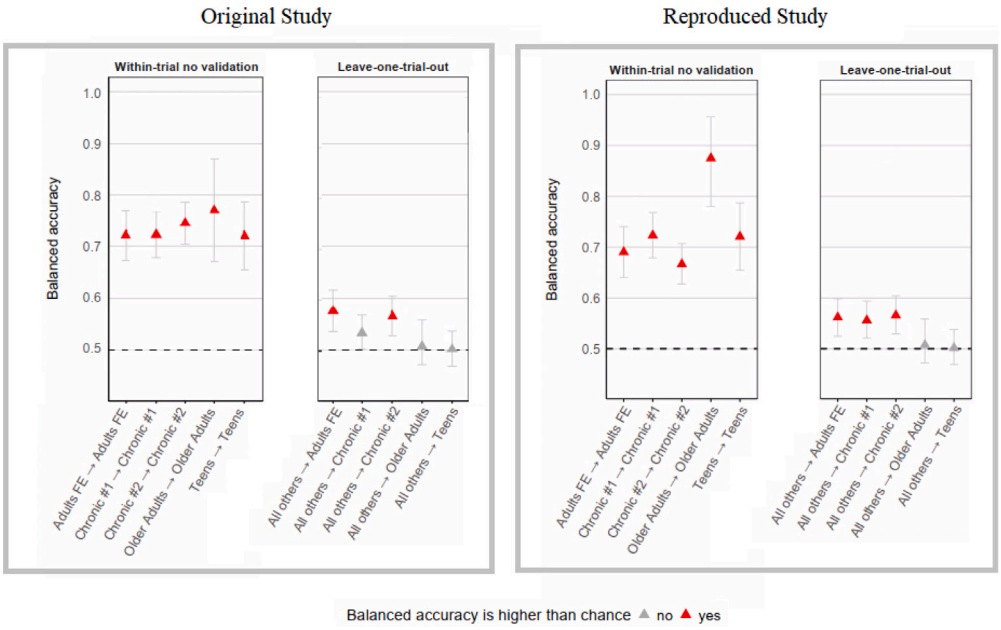

Fig. 3. Original published results from [23] vs. reproduced results with a different random seed using the published source code [26].

both scenarios showed virtually no difference between the original study (0.737, 0.537) and our reproduced study (0.735, 0.539).

Further adjusting the random seed again showed differences within the individual scenarios. These findings are consistent with those of the original study, and could potentially result in challenges in reproducing study results, even under ideal conditions where the source code, data, and hardware platforms have been duplicated, as was the case in this particular study [23]. Based on the results of the original study [23], the authors suggested that machine learning models predicting treatment outcomes (in schizophrenia) are highly context-dependent and may have limited generalizability. A potential approach to address variations in experimental context and patient demographics is the development of patient-specific models [44–48]. However, this solution introduces greater complexity, logistical challenges, and increased costs related to real-world implementation and scaling. Therefore, a new general method is needed to improve reproducibility and generalizability across diverse contexts, and this is the focus of this work.

Bouthillier et al. [49] identified three types of study reproducibility:

- Methods Reproducibility: A method is reproducible if reusing the original code leads to the same results.
- Results Reproducibility: A result is reproducible if a re-implementation of the method generates statistically similar values.
- Inferential Reproducibility: A finding or a conclusion is reproducible if one can draw it from a different experimental setup.

3.2. Random trial validation

Based on the above criteria, our initial experiment using the code and data [26] from [23] fell short in all three aspects when considering not only overall balanced accuracy and classification quality metrics, but also specific scenarios within the study (Fig. 3). To further highlight the impact of random seed choice for model initialization, we performed four experiments on a single dataset for diabetes classification using different random seeds and validation methods (Table 2). While scores attained for the first random seed remained consistent, a different random seed produced significantly different results irrespective of validation method. These inconsistencies become particularly relevant for feature importance ranking, alongside interpretability, beyond mere predictive performance [50]. Cohen's d statistic for both methods shows large effect sizes (two standard deviations), indicating that the choice of random seed has a strong impact on model performance in both validation methods.

3.3. Comparison with other methods

Breiman [37] initially introduced an ad hoc, computationally efficient feature importance calculation method for the RF algorithm known as "out-of-bag" variable importance (OOB VIMP) which remains the default in most implementations and is widely used in the research community, despite significant limitations. Wallace et al. [51] evaluated OOB VIMP's and proposed "knockoff VIMPs", an improved method which facilitates a direct and interpretable estimate of the value of a feature within a model, while however, still resulting in individual variables that can be challenging to interpret.

Fig. 4 provides a comparison of feature importance scores attained using industry standard methods including SHAP, LIME and the built-in method provided by the RF algorithm (OOB VIMP). A single model was trained using RF with a consistent random seed applied for initialization. The scores per feature produced by each individual method differ widely, providing no consistency and likely leading to significant challenges in interpretability and trustworthiness of the model's predictions.

Henderson et al. [39] in their study on RF feature importance metrics in medicine, suggested the use of proper significance testing and multiple trials with varying random seeds when comparing predictive performance, with random seed selection explicitly part of the algorithm. This averaging of multiple runs over different random seeds can give insight into the population distribution of the algorithm performance in an environment [39]. Building upon this and the aforementioned works, we developed a new validation approach aimed at stabilizing reproducibility, enhancing generalization, and improving the explainability of machine learning models.

3.4. Evaluating random trial validation

To thoroughly assess the effectiveness of our proposed randomized validation approach, we conducted experiments using a selection of well-researched open datasets (Table 1, datasets 2–8) as detailed in

Table 2
Accuracy scores based on varying random seeds and validation methods.

Validation method	Random seed	Accuracy	Precision	Recall	F1-score	Cohen's d
80/20	42	91.42%	88.88%	88.88%	88.88%	1.99
80/20	43	97.14%	96.29%	96.29%	96.29%	1.99
10-Fold CV	42	91.42%	88.88%	88.88%	88.88%	2.00
10-Fold CV	43	92.85%	89.28%	92.59%	90.90%	2.00

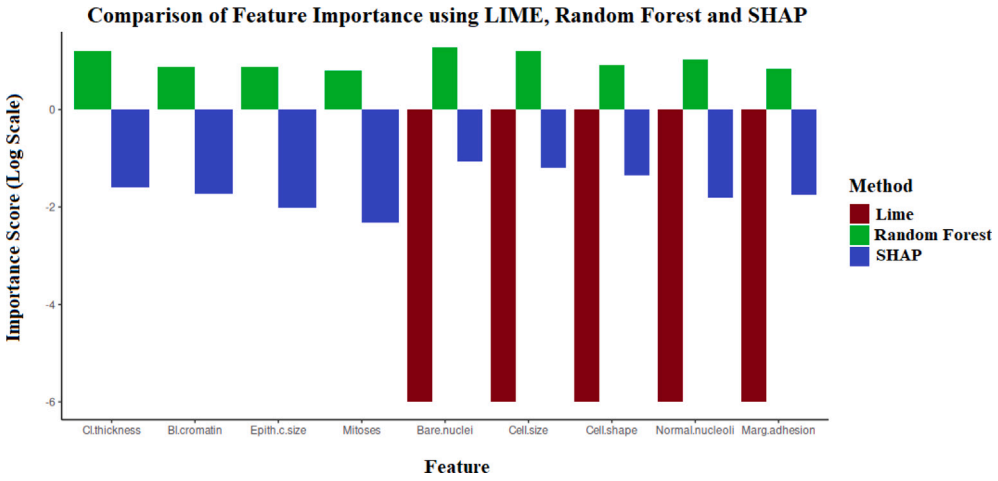

Fig. 4. Feature importance results on the Breast Cancer dataset [28] using common methods including SHAP, LIME and Random Forest built-in algorithm.

Sections 2.2 and 2.3. Among these seven datasets, three are related to health care, whereas four non-health care datasets were chosen to eliminate any potential concerns related to population diversity. Below, we discuss the results from all health-related datasets.

Fig. 5 details the outcome of experimentation using RF on the breast cancer dataset (Table 1, #2) [28]. Where a 80%/20% train/test split method was used for validation, the cell size feature consistently ranked as the most important, irrespective of random seed choice. However, feature ranking for the next four most important features differed significantly (see the first row of Fig. 5). The same effect can be observed when other validation methods were utilized (10-Fold CV, LOSO) with varying random seeds, where bare nuclei ranked as the top most important feature. Using the proposed random trial validation method described in Section 2.3, we reach feature importance stability within 256 trial iterations across all subjects as a group. Additionally, we reach a stable feature importance set per subject (not shown) using our proposed technique shown in Fig. 1.

Experimentation on the diabetes dataset [29] (Fig. 6) yielded matching feature importance ranks irrespective of random seed choice when using an 80%/20% train/test split for validation (see the first row of Fig. 6). However, the results are markedly different compared to those of the 10-Fold CV method (second row of Fig. 6). Additionally, there is significant variation in feature importance rank using two different seed values, further demonstrating the instability in machine learning on this dataset. By applying the proposed random trial validation method, we reached feature importance stability within 400 trials. Importantly, subject-level feature importance (for Subject 1 as a sample) differs from overall group feature importance.

Similar observations were noted for experiments performed on datasets 4 to 8 (Table 1), irrespective of sample size, which was tested by using dataset 8 (diamond classification). Because of space constraints, these findings are not included in the main text but can be reviewed and replicated through our open-source code and data available at https://github.com/xalentis/Reproducibility.

Dataset 9 (Alzheimer's disease) [35] was used to validate the proposed random trial approach in a practical, real-world scenario. Fig. 7 shows the feature importance plots generated using the default OOB VIMP approach implemented in the RF algorithm. When comparing the outcomes from an 80%/20% train/test split to those from 10-Fold CV, we observe variations in feature importance rankings, including four completely distinct features among the top five, aside from FAST, which is ranked as the most critical in both cases. This highlights how easily a model's explainability and stability can be influenced by merely choosing a different validation method.

3.5. Comparative analysis and validation insights into feature importance

In a prior study by Besga et al. [43] utilizing Support Vector Machines (SVM), CART Decision Trees, and RF to classify healthy controls from subjects diagnosed with Alzheimer's disease using dataset 9, a Welch's t-test p-value for each behavioral, biological biomarker, and the aggregate neuro-psychological feature was calculated. Results from this study indicated FAST, apathy (TA3), executive functions (EF), attention (A) and memory (M) as the most statistically significant features ($p < 0.001$) followed by total sleep (TS), total anxiety (TA2) and total dysphoria/depression (TDD).

Based on the results from [43], Fig. 7 prompts an inquiry into which of the two validation strategies better highlights the most relevant top-tier features. The 80%/20% validation scheme ranks the FAST feature among its top five features with $p < 0.001$; however, it also includes TA1 and TD1, which are absent from the top eight statistically significant features identified in [43]. In contrast, the 10-fold CV approach includes three features in its top five that are statistically significant at $p < 0.001$, while the other two features (TC and TD2) do not appear in the top eight significant features according to the same study.

To assess the significance of features in the Alzheimer's dataset [35], Spearman correlations were computed between the 23 features and the two classes (Normal and Alzheimer's), as depicted in Fig. 8. This figure is consistent with the results of [43] and offers an additional approach to corroborate our proposed randomized trial method.

Fig. 9 shows the feature importance obtained using the proposed random trial method for the group (left) and for an individual subject (right). The feature rankings demonstrate a strong correlation with previous research [43] and the findings shown in Fig. 8, particularly for FAST, M, TA3 and A. In contrast to the outcomes shown in Fig. 7, our proposed method successfully identifies four statistically significant

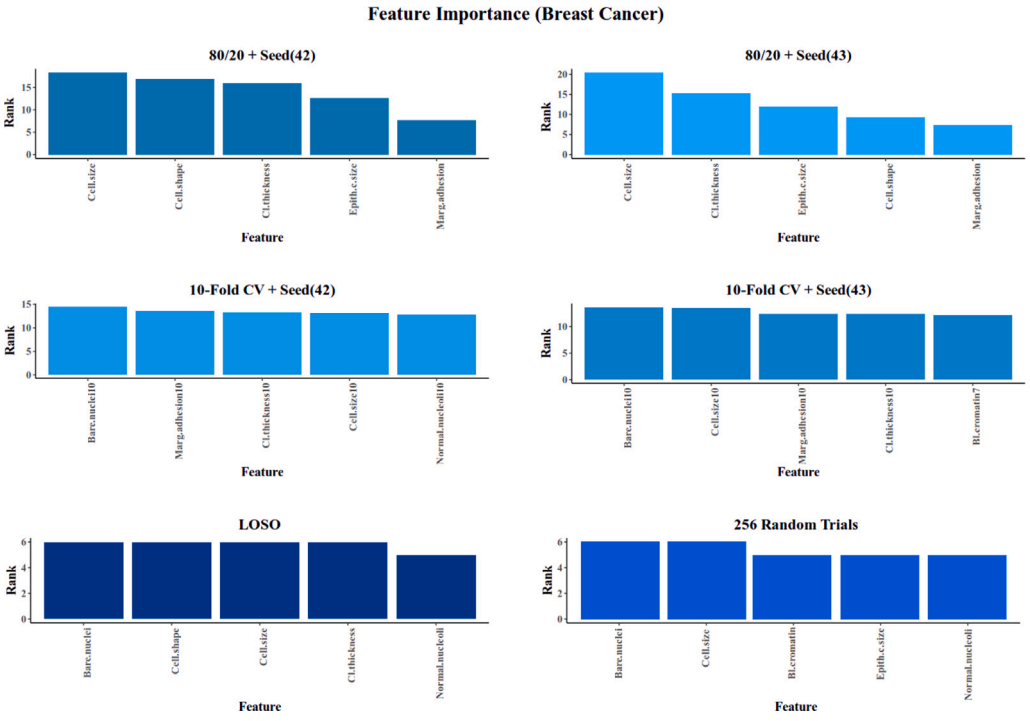

Fig. 5. Experimental results on the Breast Cancer dataset [28]. The figure shows how modifying the cross-validation technique and/or random seed can result in different feature importance sets, undermining model generalization, stability, and explainability. The figure also shows a stabilized feature importance set, using our proposed random trial validation technique.

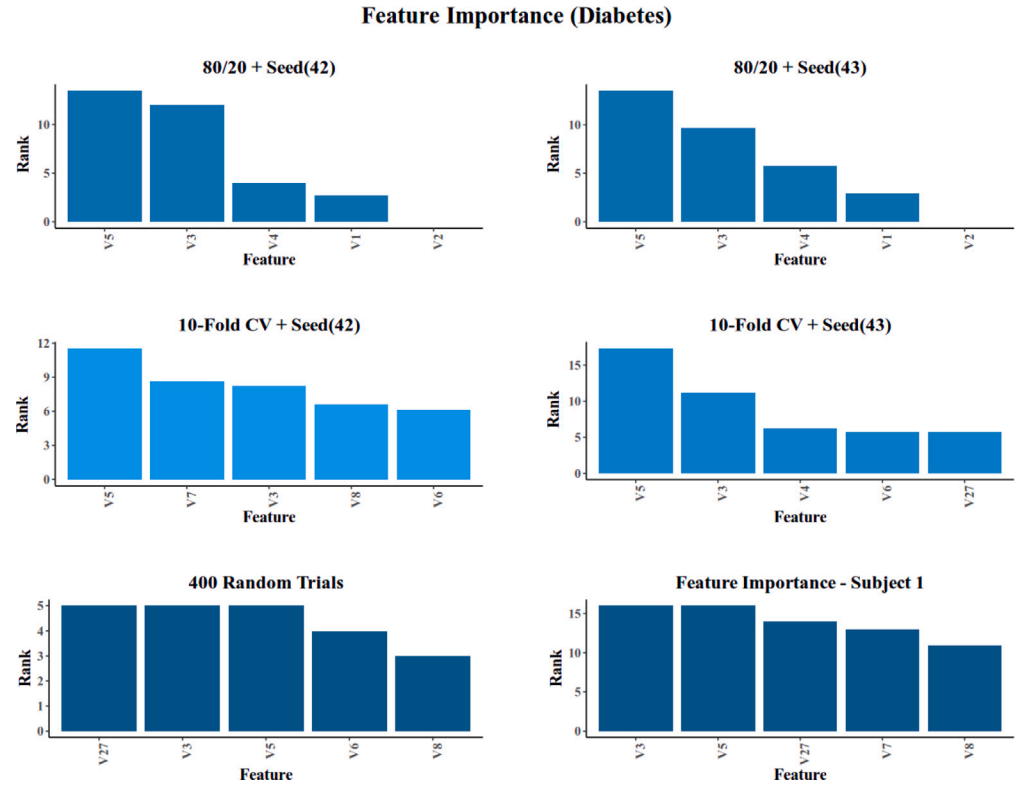

Fig. 6. Experimental results on the Diabetes dataset [29]. The figure shows how modifying the cross-validation technique and/or random seed can result in different feature importance sets, undermining model generalization, stability, and explainability. The figure additionally presents a stabilized feature importance set for the entire dataset subjects (third row, left column) and for a sample individual subject 1 (right column), employing our proposed validation method of random trials.

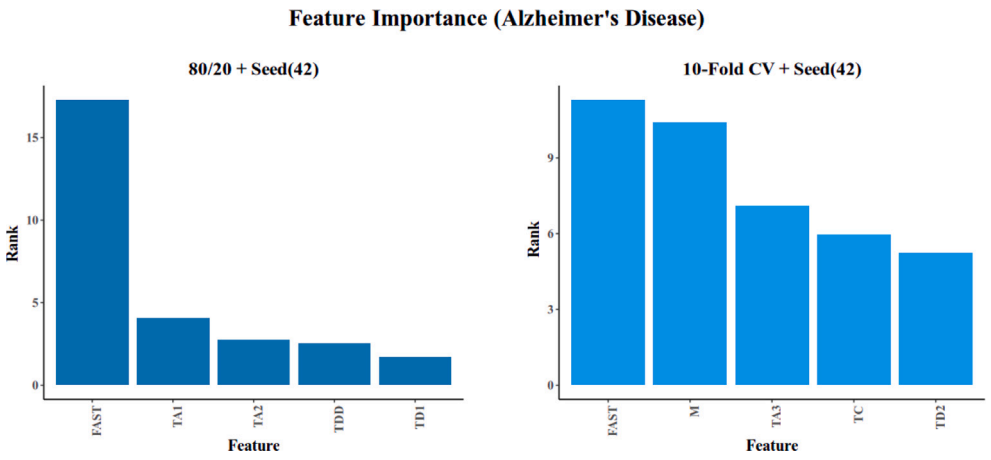

Fig. 7. Experimental results on the Alzheimer's disease dataset [35]. The figure shows how modifying the cross-validation technique even when the random seed is kept the same results in different feature importance sets, undermining model generalization, stability, and explainability.

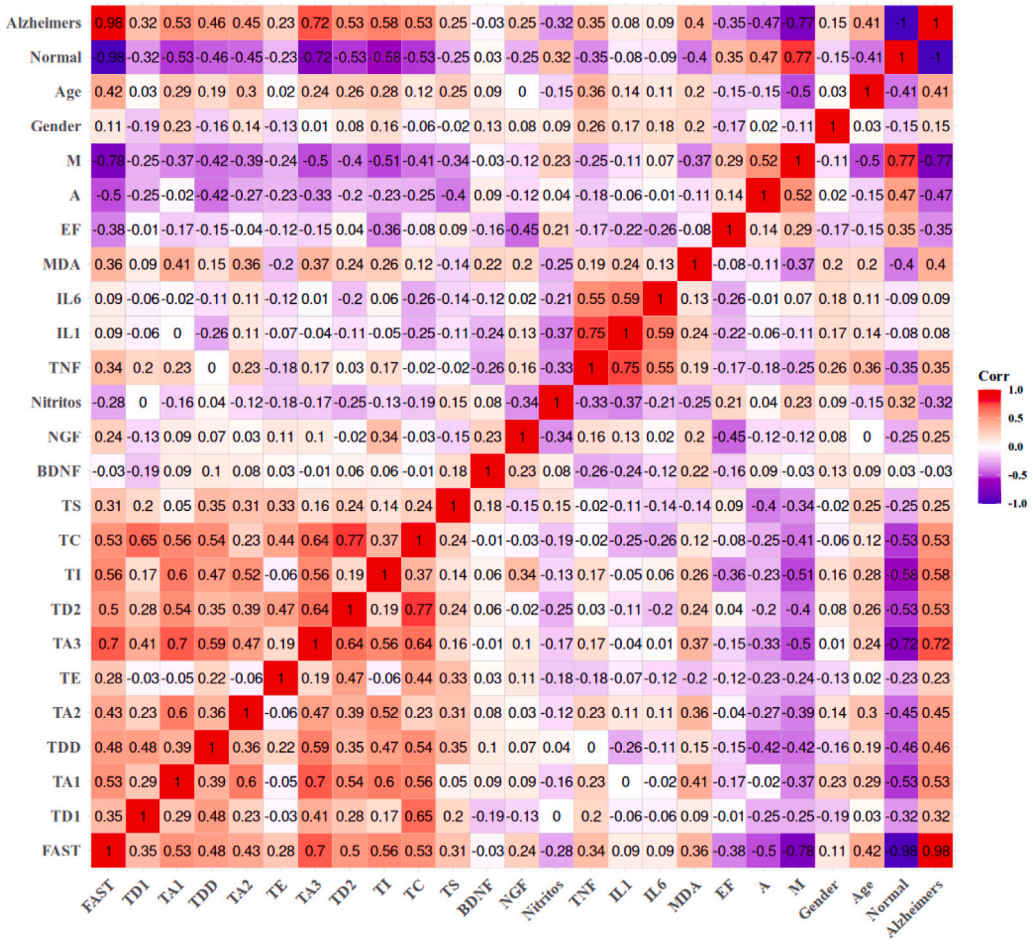

Fig. 8. Spearman correlations between 23 features and two classes (i.e. Normal and Alzheimer's) within dataset [35].

features with $p < 0.001$ among its top-5 rankings, assigning them all high ranks. This result is not achieved in any of the experiments depicted in Fig. 7. Additionally, our proposed method yields stable group and subject-level feature importance that correlates well with prior clinical findings on biomarkers significant in Alzheimer's disease [43], irrespective of random seed choice for algorithm initialization. For this particular dataset, all individual subject-level feature importance ranks corresponded with those at the group level.

3.6. Comparative evaluation of validation techniques: accuracy and computational efficiency

Table 3 provides a comparison of accuracy scores achieved for two sample datasets included in this study (the Breast Cancer [28] and the Alzheimer's Disease [35]), using a variety of common validation methods, along with our proposed validation method using random trials. Our proposed approach attained similar accuracy levels on both datasets when matched against the three traditional validation

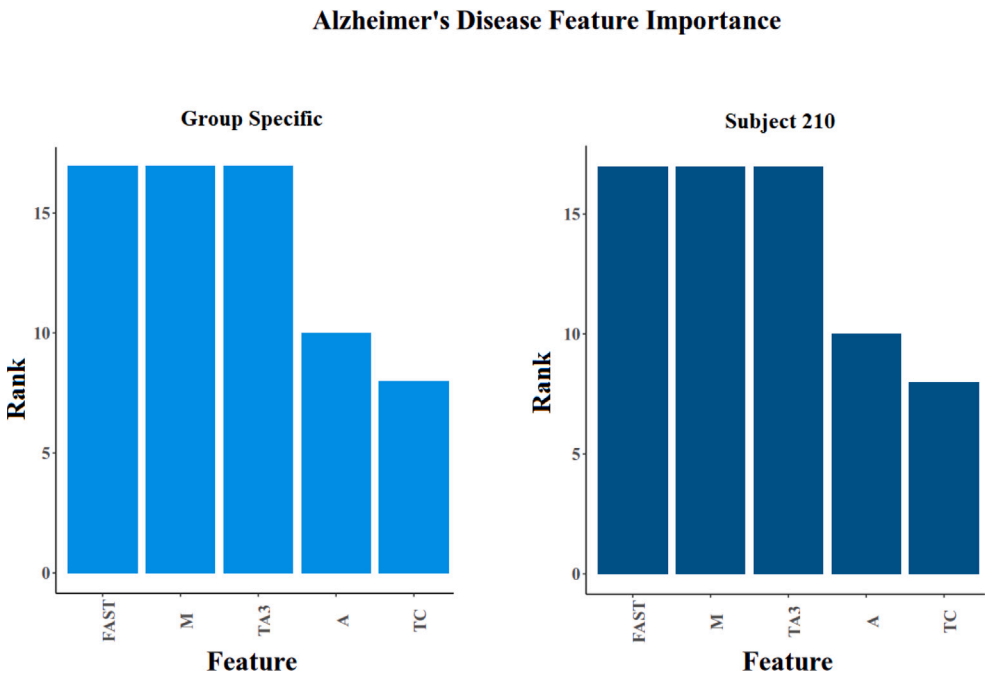

Fig. 9. Experimental results on the Alzheimer's disease dataset [35] using the proposed randomized trial method.

Table 3
Validation method accuracy comparison.

Dataset	Sample size	Validation	Accuracy
2. Breast Cancer [28]	630	80/20	100.00%
2. Breast Cancer [28]	630	10-Fold CV	97.00%
2. Breast Cancer [28]	630	LOOCV	97.00%
2. Breast Cancer [28]	630	Random Trials	**99.50%**
9. Alzheimer's Disease [35]	48	LOSO	100.00%
9. Alzheimer's Disease [35]	48	Random Trials	**100.00%**

techniques. Nevertheless, Table 3 illustrates the impact of the chosen validation technique on model performance. It is important to highlight that, given the limited sample size of 48 in the Alzheimer's Disease dataset [35], only LOSO and our proposed validation methods were employed.

Table 4 provides a summary of the execution times (in minutes) for each validation experiment conducted on datasets 2–8 in this study. Over 400 randomized trials, our proposed method significantly reduced the computational time required compared to the standard LOSO approach, while taking longer time than the 10-fold cross-validation (CV). The 80%/20% method resulted in much shorter experimentation times. Although our approach has a longer execution time compared to 10-fold CV and 80%/20% methods, it offers the advantages of yielding stable, reproducible accuracy scores and reliable feature importance assessments on both a group and subject level.

Table 5 provides a comparison of traditional validation methods with the proposed method. The commonly used 80/20 split and cross-validation shows high susceptibility to random seed variation, leading to significant fluctuations in feature importance rankings and undermining model explainability. These inconsistencies can erode trust in ML outputs, particularly in clinical contexts where stable identification of critical features is essential for decision-making and regulatory acceptance.

4. Conclusion and discussion

4.1. Key findings

In this study, we introduced a novel validation technique for determining and stabilizing both group-level and subject-specific feature importance within a single, generalized machine learning framework. This approach addresses the inherent variability in human biology, a key factor that often complicates the reproducibility and interpretability of machine learning results, even when the same hardware and software settings are applied.

Existing research often focuses on building highly accurate group-level, or general machine learning models. While these models frequently achieve and report high accuracy scores, a significant limitation arises when their performance is tested on new, unseen data [52]. In many cases, these models fail to generalize effectively due to inherent subject-specific differences in the training data, such as variations in physiological responses, individual baseline characteristics, and sensor placement inconsistencies.

4.2. Clinical implications

Both group-level and subject-specific models play crucial roles in clinical decision-making. While general models provide broad insights applicable to large populations, subject-specific models enable tailored treatment and monitoring, enhancing patient outcomes. The integration of both approaches, leveraging general models for population-based risk assessments and subject-specific models for personalized care, can significantly enhance AI-driven medicine.

While subject-specific models offer greater personalization, they require more data for each individual, which may not always be available. Additionally, these models need continuous updating and validation to ensure they remain accurate as a patient's health evolves. In contrast, group-level models benefit from larger datasets but may not be as precise for individual subjects. In clinical practice, ensuring interoperability between these models, validating them across different patient groups, and addressing ethical concerns related to personalization such as bias are key challenges that must be carefully managed.

4.3. Methodological implications

We performed an array of experiments on several open datasets to evaluate the performance of our approach. Using a sample dataset, we

Table 4
Validation method execution time comparison.

Dataset	Sample size	400 Random Trials (mins.)	LOSO (mins.)	10-Fold CV (mins.)	80/20 (s)
2. Breast Cancer [28]	683	7.2	15	4	0.1
3. Diabetes [29]	351	1.6	2	1	0.2
4. College [30]	777	2.4	7	2	0.6
5. Cars [31]	32	0.07	0.38	0.05	0.01
6. Glaucoma [32]	196	1.2	1	1	0.2
7. Glass [33]	214	0.6	0.42	0.4	0.1
8. Diamonds [34]	250	2	2	1	0.3
8. Diamonds [34]	500	4	7	2	0.6
8. Diamonds [34]	2000	16	99	11	3
8. Diamonds [34]	5000	38.4	600	29	7
9. Alzheimer's Disease [35]	48	6.1	0.14	0.064	0.02

Table 5
Comparison of different validation techniques and their impact on feature importance stability.

Validation technique	Feature stability	Computational cost
80/20 Train-Test Split	Unstable	Low (0.1 s–0.6 s)
10-Fold CV	Moderate	Moderate (0.4 min–11 min)
LOOCV	Moderate	High (0.4 min–29 min)
LOSO	Unstable	Very High (0.14 min–600 min)
Proposed Method	High	Moderate to High (1.6 min–38.4 min)

demonstrated that the proposed method achieves high Spearman correlation levels between expected and predicted feature importance, aligning with established biomarkers for Alzheimer's disease, thereby underscoring its clinical relevance. Moreover, we showed that the method delivers predictive accuracy and runtime performance comparable to and often superior to widely used validation techniques, offering a stable and interpretable alternative for biomedical applications.

The proposed method inherently accounts for time-varying characteristics of feature importance through its design of repeated trials and aggregation over multiple instances. By conducting multiple independent trials per subject and recording feature importance at each trial, temporal fluctuations and variability in feature relevance are captured across the evaluation process. The subsequent aggregation step does not merely select features based on a single snapshot but rather ranks features based on their consistent contribution across time, smoothing out transient variations. This longitudinal consideration ensures that the final ranked feature sets reflect stable and robust patterns, rather than being influenced by momentary shifts in data characteristics.

4.4. Limitations and future work

Importantly, feature importance does not necessarily imply a direct cause of the outcome, and our proposed method does not resolve problems such as bias, confounding variables or data distribution disparity. A further limitation is its higher computational demand compared to widely used techniques such as 10-fold cross-validation and the 80/20 validation split. However, our method demonstrated significantly improved computational efficiency compared to the commonly used LOSO technique on both high-powered server systems (10-core, 128 GB RAM) as well as desktop and laptop type systems (4-core, 8 GB RAM), which is prevalent in medical machine learning research. Despite its increased computational cost, the enhanced stability in reproducibility and explainability offered by our approach provides a valuable trade-off, making it a worthwhile option in medical AI, where these factors are critical.

While innovations in medical machine learning, like the proposed approach, have the potential to make a significant impact, their effectiveness will be limited if the study findings are not reproducible and accessible for further research. Open access to data and code is vital for advancing scientific research, particularly in health care, where reproducibility and transparency are essential for fostering trust in AI-driven solutions. By supporting the replication of results and encouraging further exploration, open access facilitates the development of robust, generalizable, and clinically impactful AI models, which is the primary aim of our study. To support further research in this area, the full R source code used in this study is available on GitHub at https://github.com/xalentis/Reproducibility.

CRediT authorship contribution statement

Gideon Vos: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Conceptualization. **Liza van Eijk:** Writing – review & editing, Supervision. **Zoltan Sarnyai:** Writing – review & editing, Supervision. **Mostafa Rahimi Azghadi:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

Data availability

This study, carried out under YODA Project #2024-0052, used data obtained from the Yale University Open Data Access Project, which has an agreement with JANSSEN RESEARCH & DEVELOPMENT, L.L.C. The interpretation and reporting of research using this data are solely the responsibility of the authors and does not necessarily represent the official views of the Yale University Open Data Access Project or JANSSEN RESEARCH & DEVELOPMENT, L.L.C. Data and codes for reproducing the results in this paper are available at https://github.com/xalentis/Reproducibility.

References

[1] D.D. Martinelli, Generative machine learning for de novo drug discovery: A systematic review, Comput. Biol. Med. 145 (2022) 105403, http://dx.doi.org/10.1016/j.compbiomed.2022.105403.

[2] A. Karampuri, S. Kundur, S. Perugu, Exploratory drug discovery in breast cancer patients: A multimodal deep learning approach to identify novel drug candidates targeting RTK signaling, Comput. Biol. Med. 174 (2024) 108433, http://dx.doi.org/10.1016/j.compbiomed.2024.108433.

[3] S. Bhattacharjee, B. Saha, S. Saha, Symptom-based drug prediction of lifestyle-related chronic diseases using unsupervised machine learning techniques, Comput. Biol. Med. 174 (2024) 108413, http://dx.doi.org/10.1016/j.compbiomed.2024.108413.

[4] T. Zhu, X. Liu, J. Wang, R. Kou, Y. Hu, M. Yuan, C. Yuan, L. Luo, W. Zhang, Explainable machine-learning algorithms to differentiate bipolar disorder from major depressive disorder using self-reported symptoms, vital signs, and blood-based markers, Comput. Methods Programs Biomed. 240 (2023) 107723, http://dx.doi.org/10.1016/j.cmpb.2023.107723.

[5] L. Liu, Y. Li, N. Liu, J. Luo, J. Deng, W. Peng, Y. Bai, G. Zhang, G. Zhao, N. Yang, C. Li, X. Long, Establishment of machine learning-based tool for early detection of pulmonary embolism, Comput. Methods Programs Biomed. 244 (2024) 107977, http://dx.doi.org/10.1016/j.cmpb.2023.107977.

[6] M. Aslam, F. Rajbdad, S. Azmat, Z. Li, J.P. Boudreaux, R. Thiagarajan, S. Yao, J. Xu, A novel method for detection of pancreatic ductal adenocarcinoma using explainable machine learning, Comput. Methods Programs Biomed. 245 (2024) 108019, http://dx.doi.org/10.1016/j.cmpb.2024.108019.

[7] Y. Yuan, C. Shi, H. Zhao, Machine learning-enabled genome mining and bioactivity prediction of natural products, ACS Synth. Biol. 12 (9) (2023) 2650–2662, http://dx.doi.org/10.1021/acssynbio.3c00234.

[8] G. Magazzù, G. Zampieri, C. Angione, Clinical stratification improves the diagnostic accuracy of small omics datasets within machine learning and genome-scale metabolic modelling methods, Comput. Biol. Med. 151 (2022) 106244, http://dx.doi.org/10.1016/j.compbiomed.2022.106244.

[9] M. Yang, J. Ma, Machine learning methods for exploring sequence determinants of 3D genome organization, J. Mol. Biol. 434 (15) (2022) 167666, http://dx.doi.org/10.1016/j.jmb.2022.167666.

[10] M.B.A. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, M. Ghassemi, Reproducibility in machine learning for health research: Still a ways to go, Sci. Transl. Med. 13 (586) (2021) http://dx.doi.org/10.1126/scitranslmed.abb1655.

[11] L. Goetz, N. Seedat, R. Vandersluis, M. van der Schaar, Generalization—a key challenge for responsible AI in patient-facing clinical applications, Npj Digit. Med. 7 (1) (2024) http://dx.doi.org/10.1038/s41746-024-01127-3.

[12] O.E. Gundersen, S. Kjensmo, State of the art: Reproducibility in artificial intelligence, Proc. AAAI Conf. Artif. Intell. 32 (1) (2018) http://dx.doi.org/10.1609/aaai.v32i1.11503.

[13] J. Yang, A.A.S. Soltan, D.A. Clifton, Machine learning generalizability across healthcare settings: insights from multi-site COVID-19 screening, Npj Digit. Med. 5 (1) (2022) http://dx.doi.org/10.1038/s41746-022-00614-9.

[14] P. Ball, Is AI leading to a reproducibility crisis in science? Nature 624 (7990) (2023) 22–25, http://dx.doi.org/10.1038/d41586-023-03817-6.

[15] S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learning-based science, Patterns 4 (9) (2023) 100804, http://dx.doi.org/10.1016/j.patter.2023.100804.

[16] A. Ameli, L. Peña-Castillo, H. Usefi, Assessing the reproducibility of machine-learning-based biomarker discovery in Parkinson's disease, Comput. Biol. Med. 174 (2024) 108407, http://dx.doi.org/10.1016/j.compbiomed.2024.108407.

[17] R. Van Noorden, J.M. Perkel, AI and science: what 1,600 researchers think, Nature 621 (7980) (2023) 672–675, http://dx.doi.org/10.1038/d41586-023-02980-0.

[18] D. Gunning, D.W. Aha, DARPA's explainable artificial intelligence program, AI Mag. 40 (2) (2019) 44–58, http://dx.doi.org/10.1609/aimag.v40i2.2850.

[19] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, J. Qadir, Explainable, trustworthy, and ethical machine learning for healthcare: A survey, Comput. Biol. Med. 149 (2022) 106043, http://dx.doi.org/10.1016/j.compbiomed.2022.106043.

[20] M. Nagendran, P. Festor, M. Komorowski, A.C. Gordon, A.A. Faisal, Quantifying the impact of AI recommendations with explanations on prescription decision making, Npj Digit. Med. 6 (1) (2023) http://dx.doi.org/10.1038/s41746-023-00955-z.

[21] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, 2019, http://dx.doi.org/10.48550/ARXIV.1910.10045.

[22] P. Futoma, The lancet digital health JF - the lancet digital health, 2020, http://dx.doi.org/10.1016/S2589-7500(20)30186-2.

[23] A.M. Chekroud, M. Hawrilenko, H. Loho, J. Bondar, R. Gueorguieva, A. Hasan, J. Kambeitz, P.R. Corlett, N. Koutsouleris, H.M. Krumholz, J.H. Krystal, M. Paulus, Illusory generalizability of clinical prediction models, Science 383 (6679) (2024) 164–167, http://dx.doi.org/10.1126/science.adg8538.

[24] H. Ahmed, J. Lofstead, Managing randomness to enable reproducible machine learning, in: Proceedings of the 5th International Workshop on Practical Reproducible Evaluation of Computer Systems, in: P-RECS '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 15–20, http://dx.doi.org/10.1145/3526062.3536353.

[25] A. Mellinger, D. Justice, M. Connor, S. Gallagher, T. Brooks, The myth of machine learning non-reproducibility and randomness for acquisitions and testing, evaluation, verification, and validation, 2025, http://dx.doi.org/10.58012/g17y-gp09, (Accessed 27 February 2025).

[26] A. Chekroud, M. Hawrilenko, H. Loho, J. Bondar, R. Gueorguieva, A. Hasan, J. Kambeitz, P. Corlett, N. Koutsouleris, H. Krumholz, J. Krystal, M. Paulus, Code to accompany Illusory Generalizability of Clinical Prediction Models, Zenodo, 2023, http://dx.doi.org/10.5281/ZENODO.10086334.

[27] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, IEEE, 1995, pp. 278–282.

[28] K.P. Bennett, O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, Optim. Methods Softw. 1 (1) (1992) 23–34, http://dx.doi.org/10.1080/10556789208805504.

[29] J. Smith, J. Everhart, W. Dickson, W. Knowler, R. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, Orlando, 1988, pp. 261–265.

[30] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, Springer, New York, 2013.

[31] M. Ezekiel, Methods of Correlation Analysis, vol. 427, J. Wiley & Sons, 1930.

[32] T. Hothorn, A. Zeileis, Ipred: Improved predictors, 2023, URL https://CRAN.R-project.org/package=ipred.

[33] B. German, Glass Identification, UCI Machine Learning Repository, 1987, http://dx.doi.org/10.24432/C5WW2P.

[34] H. Wickham, Dataset: Diamonds, Zenodo, 2019, http://dx.doi.org/10.5281/ZENODO.3522106.

[35] A. Besga, M. Graña, D. Chyzhyk, Alzheimer's disease versus bipolar disorder versus health control MRI data and processed results, Front. Aging Neurosci. (2020) http://dx.doi.org/10.5281/ZENODO.3935636.

[36] R.C. Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2021, URL https://www.R-project.org/.

[37] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/a:1010933404324.

[38] A.L. Beam, A.K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, JAMA 323 (4) (2020) 305, http://dx.doi.org/10.1001/jama.2019.20866.

[39] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters, Proc. AAAI Conf. Artif. Intell. 32 (1) (2018) http://dx.doi.org/10.1609/aaai.v32i1.11694.

[40] R.D. Peng, Reproducible research in computational science, Science 334 (6060) (2011) 1226–1227, http://dx.doi.org/10.1126/science.1213847.

[41] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017, http://dx.doi.org/10.48550/ARXIV.1705.07874.

[42] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, http://dx.doi.org/10.1145/2939672.2939778.

[43] A. Besga, I. Gonzalez, E. Echeburua, A. Savio, B. Ayerdi, D. Chyzhyk, J.L.M. Madrigal, J.C. Leza, M. Graña, A.M. Gonzalez-Pinto, Discrimination between alzheimer's disease and late onset bipolar disorder using multivariate analysis, Front. Aging Neurosci. 7 (2015) http://dx.doi.org/10.3389/fnagi.2015.00231.

[44] J.H. Chen, S.M. Asch, Machine learning and prediction in medicine — Beyond the peak of inflated expectations, N. Engl. J. Med. 376 (26) (2017) 2507–2509, http://dx.doi.org/10.1056/nejmp1702071.

[45] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, G. Stiglic, Early detection of type 2 diabetes mellitus using machine learning-based prediction models, Sci. Rep. 10 (1) (2020) http://dx.doi.org/10.1038/s41598-020-68771-z.

[46] B.A. Goldstein, A.M. Navar, M.J. Pencina, J.P.A. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Informatics Assoc. 24 (1) (2016) 198–208, http://dx.doi.org/10.1093/jamia/ocw042.

[47] D. Bzdok, M. Krzywinski, N. Altman, Machine learning: a primer, Nature Methods 14 (12) (2017) 1119–1120, http://dx.doi.org/10.1038/nmeth.4526.

[48] Z. Obermeyer, E.J. Emanuel, Predicting the future — Big data, machine learning, and clinical medicine, N. Engl. J. Med. 375 (13) (2016) 1216–1219, http://dx.doi.org/10.1056/nejmp1606181.

[49] X. Bouthillier, C. Laurent, P. Vincent, Unreproducible research is reproducible, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 725–734, URL https://proceedings.mlr.press/v97/bouthillier19a.html.

[50] O. Ciobanu-Caraus, A. Aicher, J.M. Kernbach, L. Regli, C. Serra, V.E. Staartjes, A critical moment in machine learning in medicine: on reproducible and interpretable learning, Acta Neurochir. 166 (1) (2024) http://dx.doi.org/10.1007/s00701-024-05892-8.

[51] M.L. Wallace, L. Mentch, B.J. Wheeler, A.L. Tapia, M. Richards, S. Zhou, L. Yi, S. Redline, D.J. Buysse, Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction, BMC Med. Res. Methodol. 23 (1) (2023) http://dx.doi.org/10.1186/s12874-023-01965-x.

[52] G. Vos, K. Trinh, Z. Sarnyai, M. Rahimi Azghadi, Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices, J. Biomed. Inform. 148 (2023) 104556, http://dx.doi.org/10.1016/j.jbi.2023.104556.