IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2025, accepted 29 September 2025, date of publication 7 October 2025, date of current version 16 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3618851

RESEARCH ARTICLE

Validity of Feature Importance in Low-Performing Machine Learning for Tabular Biomedical Data

YOUNGRO LEE[1,2,3,4,5], GIACOMO BARUZZO[6], JEONGHWAN KIM[7], JONGMO SEO[1,2,3,4,5], (Member, IEEE), AND BARBARA DI CAMILLO[6,8,9]

[1]Department of Electrical and Computer Engineering, Seoul National University, Gwanak-gu, Seoul 08826, Republic of Korea
[2]Institute of Engineering Research, Seoul National University, Seoul 08826, Republic of Korea
[3]Inter-University Semiconductor Research Center (ISRC), Seoul National University, Seoul 08826, Republic of Korea
[4]Interdisciplinary Program of Medical Informatics, Seoul National University College of Medicine, Seoul 03080, South Korea
[5]Biomedical Research Institute, Seoul National University Hospital, Seoul 03080, Republic of Korea
[6]Department of Information Engineering, University of Padova, 35131 Padua, Italy
[7]School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA
[8]Department of Comparative Biomedicine and Food Science, University of Padova, 35020 Padua, Italy
[9]Padua Center for Network Medicine, University of Padova, 35131 Padua, Italy

Corresponding author: B. Di Camillo (barbara.dicamillo@unipd.it)

This work was supported by the Supporting Excellence in Early-stage research Development (SEED) Project ''tRajectoriEs of baCtErial NeTwoRks from hEalthy to disease state and back (RECENTRE)'' through the Department of Information Engineering of the University of Padova under Grant DI_C_BIRD2020_01 and in part by the Institute of Information Communications Technology Planning Evaluation (IITP) Global Data X Leader Human Resources Development (HRD) Program through Korean Government [Ministry of Science and ICT (MSIT)] under Grant IITP-2024-0044-1407.

ABSTRACT In tabular data analysis, high model accuracy is often regarded as a prerequisite for discussing feature importance. This assumption stems from the expectation that the validity of feature importance correlates with model performance. In this work, we challenge this prevailing belief by demonstrating that even low-performing models can provide reliable feature importance on biomedical datasets. We conduct experiments to observe how feature importance rankings change as model performance progressively degrades. Using three synthetic datasets and four real-world biomedical datasets, we compare feature rankings from the full datasets to those obtained after reducing either the number of samples (samples removal) or the number of features (features removal), using different feature stability indices. Our results reveal that, in both synthetic and real datasets, feature rankings remain stable during performance degradation caused by features removal. In contrast, sample removal introduces greater discrepancies in feature importance rankings as performance deteriorates more severely. By analyzing the distribution of feature importance values and theoretically examining the probability that the model fails to distinguish importance between features, we show that models can still reliably identify feature importance despite performance degradation due to features removal. We conclude that the validity of feature importance can be preserved even at suboptimal model performance levels, as long as the degradation stems from insufficient features rather than insufficient samples. This has a considerable impact on biomedical research, where feature importance analysis plays a pivotal role in clinical decision support and translational bioinformatics.

INDEX TERMS Validity of feature importance, machine learning, performance, lack of samples, lack of features.

I. INTRODUCTION

A. BACKGROUND

Recently, machine learning (ML) methods have outperformed conventional regression across various domains [1], [2], [3]. Consequently, feature importance analysis is increasingly used in research and industry to understand each feature's impact on model outcomes. This process is applicable across fields but is especially critical in biomedical applications where feature importance can be the primary goal. In bioinformatics, it helps identify biomarkers among extensive genetic data points [4], [5], [6], [7]. In medical

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

fields, important features can guide the understanding of the mechanism behind diseases by pointing out important clues [8], [9], [10]. Comparing important features with prior clinical knowledge can validate decision support systems before clinical use, screening possible biases inside datasets [11], [12], [13].

While there are various types of machine learning and interpretation methods [14], [15], [16], [17], the methodology for tabular datasets is more settled compared to the other types of datasets. Although many deep learning methods for tabular datasets have been suggested recently [18], [19], tree-ensemble methods, such as Random Forest, XGBoost (Extreme Gradient Boosting), and LGBM (Light Gradient Boosting Model), and multi-layer perceptron (MLP), are more commonly used, usually in combination with various feature selection methods [6], [20]. In many cases, these approaches generally show higher performance with a lower computational burden with respect to deep learning [3], [19], [21]. As regards feature importance, intrinsic methods like Gini impurity or model-agnostic methods like SHAP (Shapley Additive exPlanations) are commonly employed [22], [23].

B. PROPOSED ANALYSIS FRAMEWORK

Here we propose an experimental framework to investigate how the validity of feature importance changes with the performance of ML models on tabular biomedical datasets. We generated three synthetic binary classification datasets with varying levels of class imbalance and collected four real world biomedical datasets of varying sample size and number of features. As tabular datasets are represented as 2-dimensional matrices with samples and features, performance issues can arise either from insufficient number of samples or inadequate choice of the features, i.e. lack of predictive features. We thus compared ML results obtained using the full datasets with those obtained with reduced sample size (samples removal) or fewer features (features removal). To assess the change of feature ranks, we used stability indexes, Spearman's rank correlation (SRCC), Canberra distance (CD), Bray–Curtis distance [6], [20] and rank difference.

The results showed that performance degradation caused by samples removal aligns with the conventional belief, significantly impacting the rank of features. However, when performance degradation is caused by features removal, the impact on the rank of features is significantly lower (Section IV-A and Section IV-B). To further investigate the mechanism behind this pattern, we examined the distribution of feature importance under the two degradation strategies (Section IV-C) and theoretically analyzed the probability that the model can distinguish the relative importance between features under simplifying assumptions (Section V).

II. RELATED WORKS

While the ML based interpretation is expected to discover the complex interplay among features, the validity of feature importance in ML is not yet well established as in statistics. Discussions on the validity of feature importance in ML highlight issues such as:

- high correlations between features can mislead interpretations [24], [25]
- feature properties (categorical or continuous) and the number of categories can arise biases [26]
- low performance can flatten the distribution of feature importance due to biases [27]

Despite plenty of discussion regarding the validity of feature importance, the current application of ML and feature importance for tabular datasets in biomedical fields often aligns the ranking of feature importance without thoroughly analyzing its validity [23], [28]. Instead of mentioning the validity of feature importance, high model accuracy is generally considered a prerequisite for discussing feature importance, with the assumption that its validity is correlated with model performance. This belief tends to generalize the methodological pipeline, which hinders the spotlight on further discussion about the validity and limit the use of ML based interpretation in cases of insufficient performance for predictive utility.

However, to the best of our knowledge, no direct experimental studies have explored this relationship between the performance of ML and the validity of feature importance on biomedical data. Why has this relationship not been clarified? Tracking feature importance relative to classification performance requires a measure to define feature importance error. However, the correct order of feature importance is subjective and possibly nonexistent. Validity can be measured indirectly: when noise features are independent of a label, the ability to filter out these noise features without confusing them with important features indicates validity [6], [29]. To observe the patterns of feature degradation as model performance declines, it is necessary to design an experiment that validates the feature ranking beyond the model's ability to exclude noisy features.

In this work, we propose a framework to evaluate the relationship between feature importance and ML model performance in biomedical settings. We applied this framework to both simulated and real-world biomedical datasets, providing an initial assessment of this relationship. A better understanding of this relationship is crucial for medical practitioners, who rely on trustworthy feature importance interpretations for informed decision-making.

III. DATA AND METHODS

A. REAL AND SYNTHETIC DATASETS

To create a dataset with clear feature rankings, we generated three simulated datasets with $N = 20$ features and $M = 10{,}000$ samples each, where binary labels $\{0, 1\}$

TABLE 1. Characteristics of datasets.

	Synthetic Dataset			Real Dataset			
	Generated Dataset 1	Generated Dataset 2	Generated Dataset 3	Real Dataset 1	Real Dataset 2	Real Dataset 3	Real Dataset 4
Target	Linear combination > threshold?			Diabetic Retinopathy	Survival after Thoracic Surgery	Heart Abnormality	Metabolic Syndrome
# of samples	10000	10000	10000	1151	470	267	2009
# of positive samples (%)	5085 (50.8)	2501 (25.0)	7501 (75.0)	611 (53.1)	70 (14.9)	212 (79.4)	712 (35.4)
# of negative samples (%)	4915 (49.2)	7499 (75.0)	2499 (25.0)	540 (46.9)	400 (85.1)	55 (20.6)	1297 (64.6)
# of features	20	20	20	19	16	44	21
# of features after deleting corr. > 0.9 (%)	–	–	–	12 (63.2)	16 (100.0)	44 (100.0)	19 (90.5)
# of features after deleting corr. > 0.7 (%)	–	–	–	9 (47.4)	16 (100.0)	28 (63.6)	19 (90.5)
# of features after deleting corr. > 0.5 (%)	–	–	–	8 (42.1)	15 (93.8)	14 (31.8)	17 (81.0)

were obtained based on a linear combination of positive coefficients $A = \{a_1, a_2, \ldots, a_N\}$ (with $a_n \in \mathrm{N}\ \forall n$) and independent continuous features $X = \{x_1, x_2, \ldots, x_N\}$ ($x_n \in [0, 1]$)

Positive coefficients a_n represent feature importance for feature x_n. Here we set $\{a_1, a_2, \ldots, a_N\}$ equal to $\{1, 2, \ldots, N\}$.

Label y_m in sample m is defined as

$$y_m = \begin{cases} 1, & \text{if } \sum_{n=1}^{N} a_n x_{nm} > \theta \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

To achieve a similar ratio between labels 1 and 0, the threshold θ must be set to $\frac{N(N+1)}{4}$, which corresponds to half of the maximum value that $A \cdot X$ can assume (when all features are equal to 1). For Generated dataset 2 and 3, the threshold was determined to ensure the balance shown in Table 1.

Four different real biomedical datasets were also used (Table 1), which are all open-source:

1) A diabetic retinopathy dataset from Messidor image samples [30], which classified diabetic retinopathy and contains 19 features and 1151 images.
2) A survival after thoracic surgery dataset [31], in which 470 lung cancer patients are classified based on death within one year after thoracic surgery, with 16 clinical features included.
3) A heart abnormality dataset [32], in which 267 single-photon emission computed tomography images are classified into normal or abnormal based on 44 features.
4) A metabolic syndrome dataset [33], comprising 2009 samples and 21 clinical features.

Real data have much higher complexities than synthetic datasets, since there are many noise features not related to the predictive target, and features have complex interactions with each other. To study feature importance without the confounding effect of correlated features, we removed correlations between features, allowing experiments to focus on more independent features. In particular, when two features exceeded the correlation threshold of 0.5, the less important feature in the initial experiment (done with full samples and features) was deleted.

We selected a threshold of 0.5 as it provided a practical balance between two competing factors: setting the threshold too low led to the elimination of too many features, limiting interpretability and statistical power, whereas setting it too high failed to ensure that the remaining features were sufficiently independent. To assess robustness, we additionally conducted experiments using thresholds of 0.7 and 0.9, with results reported in the Appendix.

B. METHODS

The overall experimental design for both synthetic and real datasets is illustrated in Figure 1. The experiment begins by training the predictive model on the original dataset, being $M_0(= M)$ the number of samples, and $N_0(= N)$ the number of features. Either the number of samples was reduced while keeping the number of features constant, or the number of features was reduced while keeping the number of samples constant (Figure 1a).

As predictive model we used Random Forest algorithm [34], [35], which is a representative tree-ensemble model. Gini-impurity index was used to calculate the rank of features. Details on the use of Random Forest (how it was applied and why we used Random Forest among many machine learning algorithms [36]) are available in Appendix I. One hundred experiments were performed for

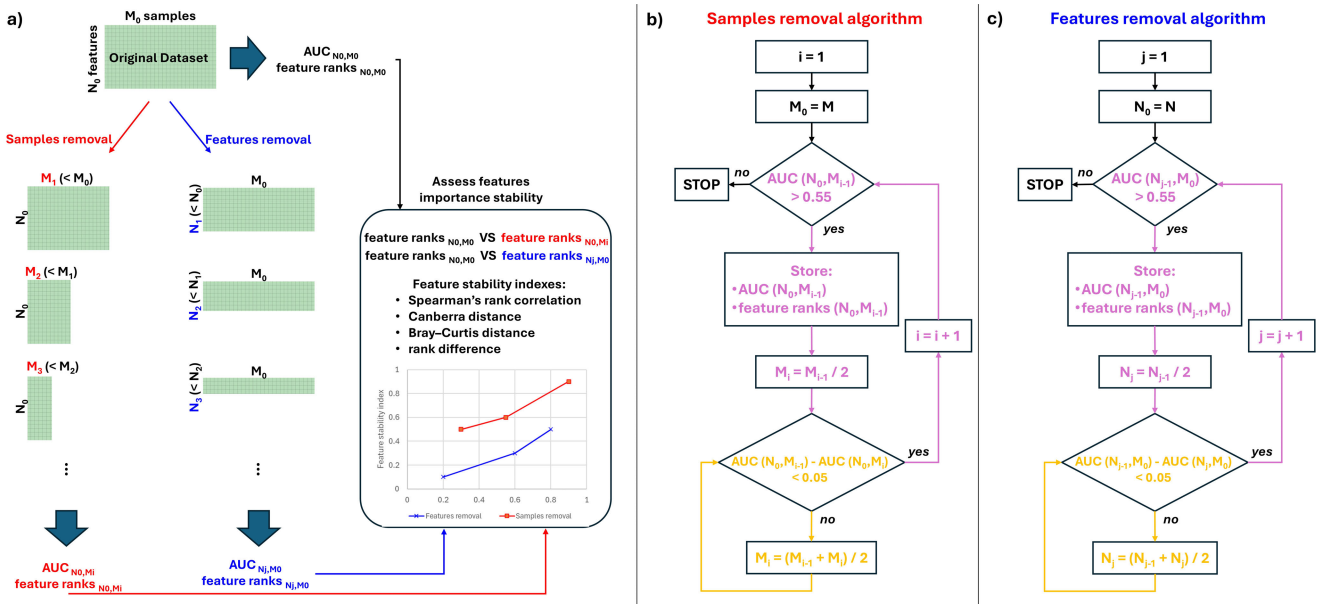


FIGURE 1. Graphical explanation for overall experiments a) Overall structure how the performance and stability indexes are obtained. The sample removal procedure (highlighted in red) iteratively reduces the number of samples, calculating the model's AUC and feature ranks at each step. Similarly the feature removal procedure (highlighted in blue) iteratively reduces the number of features, calculating the model's AUC and feature ranks at each step. Computed features ranks are compared with full dataset features rank using a variety of features stability indexes. b) Flowchart of the samples removal algorithm. The samples removal algorithm iteratively halves the number of samples, calculating the model's AUC and feature ranks at each step (highlighted in pink). If the AUC drops significantly between two consecutive iterations, the algorithm adjusts by retaining a larger subset of samples (highlighted in yellow). c) Flowchart of the features removal algorithms. The features removal algorithm iteratively halves the number of features, calculating the model's AUC and feature ranks at each step (highlighted in pink). If the AUC drops significantly between two consecutive iterations, the algorithm adjusts by retaining a larger subset of features (highlighted in yellow).

each dataset, with different random states to divide the samples into a training dataset (80%) and test dataset (20%) stratifying by class label. We averaged the classification performance and feature rankings across the 100 iterations. The Area Under the Curve (AUC), specifically under the Receiver Operating Characteristic (ROC) curve was used because it has the advantage of summarizing the overall performance in a single number [37], [38] without the need to fix a threshold to assign class labels and for assessing classification performance, which is otherwise mandatory for other performance indexes such as Matthews correlation coefficient (MCC) [39]. Further details are given in Appendix II.

The resulting average feature ranks were compared to those from the original dataset to calculate stability indexes; namely: rank difference, Spearman's rank correlation, Canberra distance, and Bray–Curtis distance [6], [20] (see Appendix III for details). All indices, except for SRCC, are distance metrics where smaller values indicate better stability. SRCC, on the other hand, is a correlation metric, with values closer to 1 indicating better stability.

The sample removal algorithm iteratively halves the number of samples, calculating the model's AUC and feature ranks at each step. This process continues until the AUC $>$ 0.55 and all classes are adequately represented in the test set in any bootstrap (highlighted in pink in the flowchart, Figure 1b). If the AUC drops significantly between two consecutive iterations (an AUC difference $>$ 0.05), indicating that halving the samples causes a sharp decline in performance, the algorithm adjusts by retaining a larger subset of samples. In this case, the number of retained samples is set to the average of the previous and current sample sizes. This adjustment continues until the drop in AUC stabilizes (highlighted in yellow in the flowchart, Figure 1b).

The features removal algorithm followed a similar approach to the samples removal algorithm, with features being sequentially removed based on their importance ranks from the initial experiment, starting with the most important (Figure 1c). Additional experiments were conducted by first removing the least important feature, which resulted in negligible changes to feature ranks at each step, and by removing features randomly, which led to highly variable results. Given these observations and the study's goal of evaluating the validity of feature importance in models affected by performance issues due to insufficient samples or poorly selected features (e.g., the absence of predictive features), the decision to first remove the most important features is both reasonable and aligned with our objectives.

All analyses were performed using Python (version 3.9.7). For building the predictive model algorithm and measuring stability indexes, the scikit-learn 0.24.2 library was used [35]. Tests were performed on a Windows 10 machine equipped with an AMD Ryzen 7 5800X 8-Core

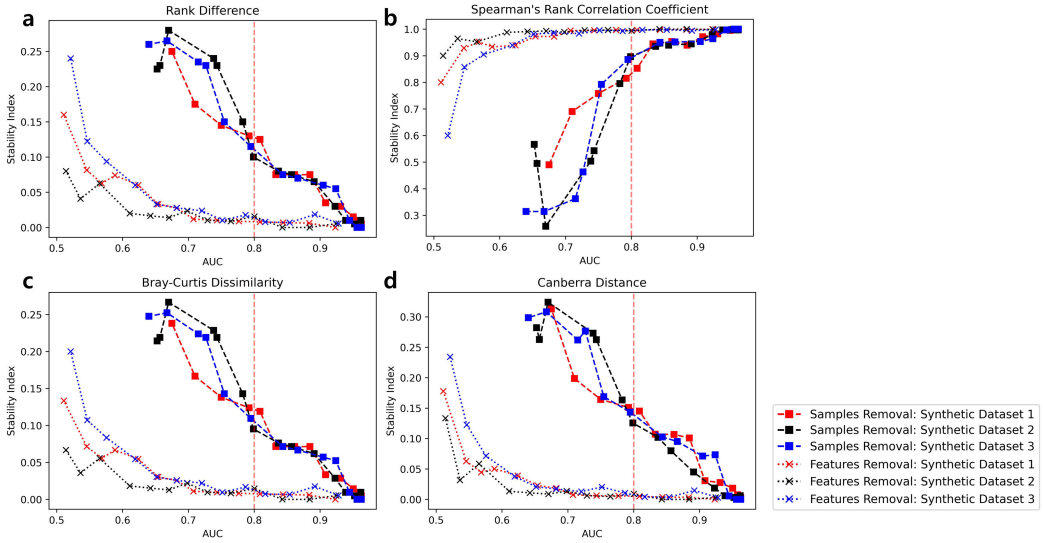


FIGURE 2. Comparison between features importance stability by samples removal and features removal on synthetic datasets. Red vertical line refers to x (AUC) = 0.8 to denote the point where the stability index starts to rapidly change in samples removal. a) Features importance stability measured with rank difference. b) Features importance stability measured with Spearman's Rank Correlation Coefficient. c) Features importance stability measured with Bray-Curtis dissimilarity. d) Features importance stability measured with Canberra distance. Samples removal and features removal results are reported with square and cross markers, respectively. Results from generated dataset 1, 2 and 3 are reported in red, black and blue, respectively.

Processor and 32 GB of RAM. The data and code are available with accompanying documentation at: https://github.com/youngrolee94/feature_importance_validity.

IV. RESULTS

A. PERFORMANCE DEGRADATION IN SYNTHETIC DATASETS

The results of the experiments in synthetic datasets are illustrated in Figure 2, where each panel depicts the relationship between AUC and different stability indices. The result demonstrates a consistent pattern across different synthetic datasets and stability indices. The stability of feature ranking drops consistently with performance degradation (from right to left in the x axis of the figures) with samples removal (decrease in SRCC and increase in the other distance indexes). While it diminishes slowly when AUC is roughly above 0.8, it drops rapidly when AUC falls below 0.8. Features removal, in comparison, does not noticeably change the overall feature rankings as the performance decreases, at least until the AUC is higher than 0.6. In Figure 2b, SRCC is higher than 0.9 (i.e. highly correlated), even when AUC was degraded to 0.6. Considering that predictive models with AUC around 0.6 to 0.7 are generally considered as poor predictors, we can say that the qualitative level of performance does not align with the stability of feature ranks during features removal.

B. PERFORMANCE DEGRADATION IN REAL DATASETS

Figure 3 illustrates the change of SRCC by performance degradation in real datasets. Generally, SRCC decreased when the performance was degraded by samples removal,

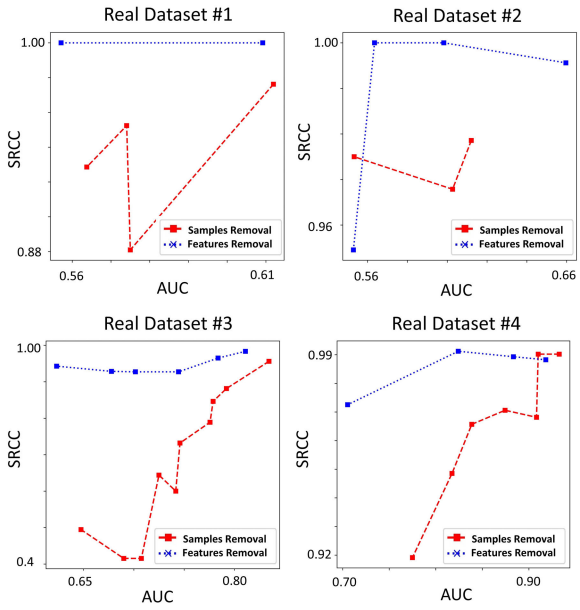


FIGURE 3. Comparison between features importance stability (measured as SRCC) by samples removal and features removal on real datasets. Samples removal and features removal results are reported with red square and blue cross markers, respectively.

as observed in synthetic datasets. However, in features removal, SRCC was higher than 0.9 in all analyzed AUC areas, which means that degraded models showed similar rank of features even when the performance was low. Particularly in real dataset 3 and 4, where the degradation was more compact compared to the other two datasets,

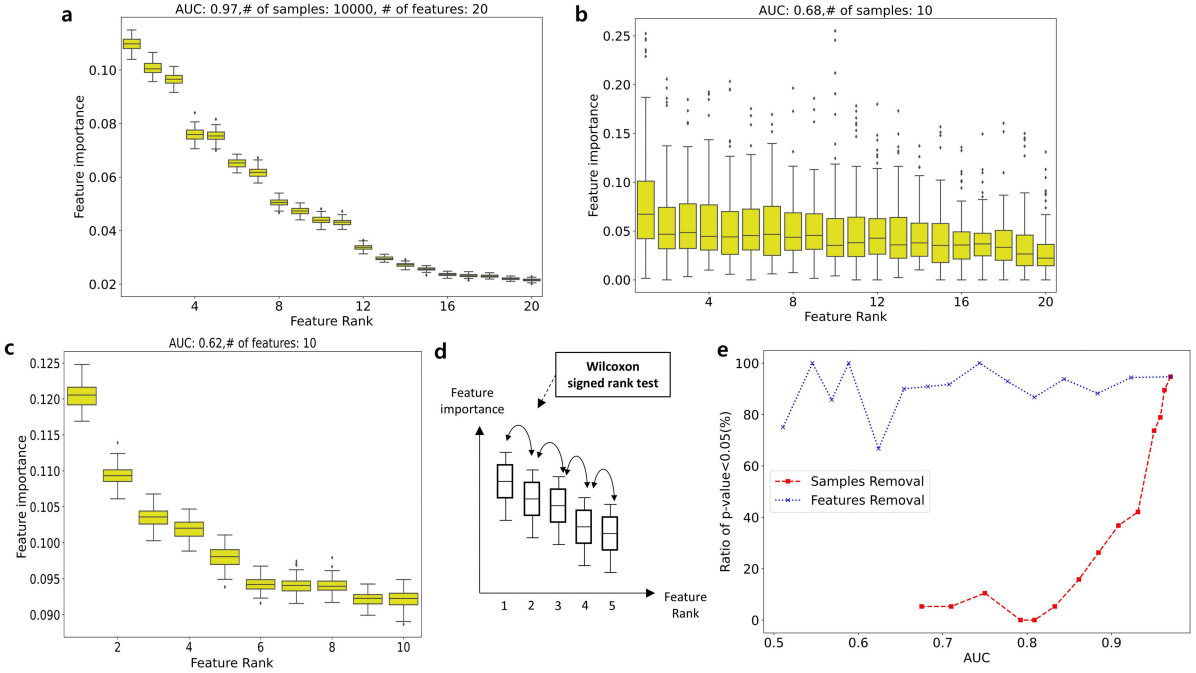


FIGURE 4. Feature importance distribution in the experiment on synthetic dataset 1. Feature importance of each feature for a) entire data samples and features, b) only with 10 samples (AUC=0.68), c) only with 10 features (AUC=0.62). Black dots show outliers from box plot. d) Schematic of calculating the statistical significance in feature importance distribution analysis. Feature importance values of adjacently ranked features are compared using statistical tests. e) Percentage of statistical tests on feature importance values on adjacent features ranks that shows a statistically significant difference (pvalue < 0.05). Features removal (blue) shows a higher percentage of statistically significant difference in feature importance on adjacent features ranks compared to data removal (red).

the rapid decline in SRCC caused by samples removal and the consistency under features removal were more prominently observed as in synthetic datasets. Similar pattern was observed in other stability indexes as illustrated in Appendix IV, Appendix Figure 1, 2 and 3 (rank difference, Canberra distance and Bray-Curtis dissimilarity, respectively).

As the experiments on the real datasets were conducted after filtering correlations, we compared the results with those obtained when correlations were not filtered or filtered using different thresholds, as shown in Appendix V, Appendix Figure 4, 5, 6 and 7 (real dataset 1, 2, 3 and 4, respectively). Across all datasets, feature rankings under features removal remained more consistent by performance degradation as more correlations were filtered. In contrast, performance degradation due to samples removal negatively impacted stability regardless of the correlation filtering threshold.

C. ANALYSIS OF FEATURE IMPORTANCE DISTRIBUTION

From Section IV-A and Section IV-B, we concluded that samples removal is the primary reason why performance degradation harms the validity of feature importance. To explore how this pattern happens, we compared how the feature importance values distributed differently by performance degradation algorithms, as illustrated in Figure 4. The box plot shows Gini-impurity values for each feature across 100 bootstraps, where box plots for each feature are aligned by their original rank by full dataset.

In the full dataset (Figure 4a), feature importance distributions exhibited few outliers, allowing adjacent box plots to be easily compared in terms of the average feature importance. With samples removal (Figure 4b), the distributions displayed numerous outliers, high variability, and no clear superiority by the ranks. In contrast, features removal (Figure 4c) produced distributions with lower variability and fewer outliers compared to samples removal, and a clear trend in terms of feature importance.

To statistically validate that the model in features removal distinguishes the rank of features better than that in samples removal at similar performance levels, we compared feature importance values between adjacent ranks using the Wilcoxon signed rank test, as illustrated in Figure 4d. In features removal, the average importance was clearly distinguished between most of adjacently ranked features with statistical significance (pvalue $<$ 0.05), regardless of performance (Figure 4e). However, in samples removal, the model failed to distinguish feature roles as the performance decreases (Figure 4e). The same analysis for real datasets is shown in Appendix VI (Appendix Figure 8). In every dataset, features removal shows a higher percentage of statistically significant difference in feature importance on adjacent features ranks compared to data removal.

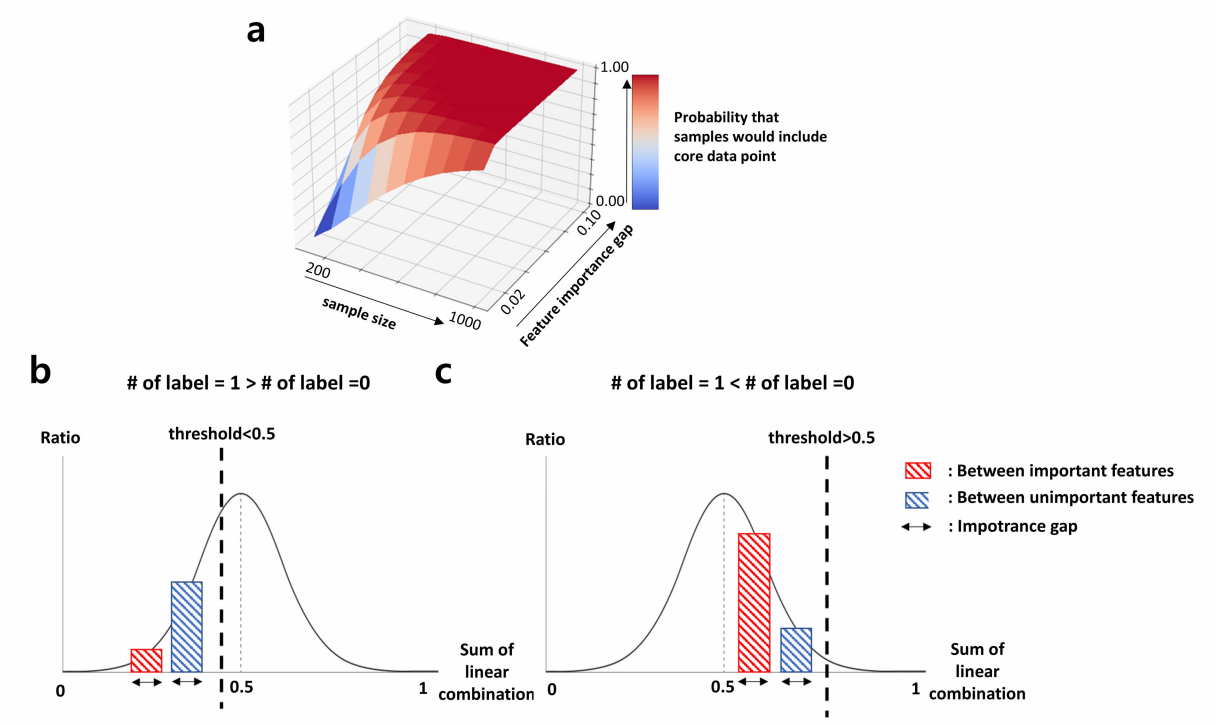


FIGURE 5. Analysis of the probability that a training dataset has sufficient samples to differentiate the importance of features. a) 3D plot to illustrate the probability described in Section V. Feature importance gap represents $a_j - a_i$ in Section V. b), c) Schematics of the probability distribution influenced by the class imbalance.

V. DISTINGUISHING FEATURE IMPORTANCE: A THEORETICAL PERSPECTIVE

To explore the theoretical mechanism behind the results presented above, we calculated the probability that a training dataset has sufficient samples to differentiate the importance of features.

We assume a population with M samples and N features where M and N are very large. We assume the same model used to generate synthetic data with independent features (1), except that the features are binary, i.e. $x_{n,m} \in \{0, 1\}$ ($x_{n,m}$ denotes the n-th feature value of the m-th sample). Each feature value $x_{m,n}$ is assumed to be distributed as a random variable X_n following Bernoulli distribution with $p = 0.5$.

Feature coefficients are here assumed to be proportional to $\{1, 2, \ldots, N-1, N\}$ scaled to satisfy

$$\sum_{n=1}^{N} a_n = 1 \tag{2}$$

We want to calculate the probability p that the training dataset includes information to differentiate the relative superiority of feature i-th with respect to feature j-th ($a_i < a_j$).

To be able to compare the importance of the j-th feature over the i-th feature, there should be at least two samples m_1 and m_2 in which the i-th feature and j-th feature values are $(x_{im_1}, x_{im_2}) = (1, 0)$ and $(x_{jm_1}, x_{jm_2}) = (0, 1)$. Let g_d denote the probability that such a pair exists when selecting d samples. This probability can be obtained by subtracting from 1 the probability that none of the d samples include both patterns. In other words, the selected d samples must then consist only of the pairs $(0, 0), (0, 1), (1, 1)$ or $(0, 0), (1, 0), (1, 1)$:

$$g_d = 1 - \left\{ \left(\tfrac{3}{4}\right)^d + \left(\tfrac{3}{4}\right)^d - \left(\tfrac{1}{2}\right)^d \right\} \tag{3}$$

Also, for the model to distinguish the importance of i-th feature and j-th feature, the model needs samples in which the importance gap between the i-th feature and j-th feature has a decisive impact on deciding labels. This happens when the sum of other features is between $(\theta - a_j, \theta - a_i)$ as the following:

$$\theta - a_j < \sum_{n=1}^{N} a_n x_{nm} - a_i x_{im} - a_j x_{jm} < \theta - a_i \tag{4}$$

In Appendix VII, we demonstrate that if N is sufficiently high, y_m is a realization of a variable Y that can be approximated by a Gaussian distribution with mean 0.5 and variance $1/3N$. Therefore, the probability p that Y is in the range $(\theta - a_j, \theta - a_i)$ is proportional to $a_j - a_i$ (though it may not be linearly proportional). Therefore, we can assume that the number of samples within the entire population satisfying 4 is, in average, equal to $M \cdot \frac{(\theta - a_i) - (\theta - a_j)}{1 - a_i - a_j} \approx M \cdot (a_j - a_i)$. Note that the denominator is approximated to 1 as the $a_i + a_j$, the sum of the feature importance of two features,

is assumed to be relatively very small compared to the entire sum of feature importance, which is equal to 1 (2).

Let k be the number of samples randomly selected as a training dataset from an initial pool of M samples ($k \ll M$) and assume d is the number of samples among the k samples in the training set that fulfills 3. To calculate the probability p that Y is in the range $(\theta - a_j, \theta - a_i)$, we can consider how many combinations include the information to distinguish the two features among the entire set of combinations containing k samples from the original M:

$$p = \frac{\sum_{d=2}^{\min(M(a_j-a_i),\,k)} g_d \binom{M-M\cdot(a_j-a_i)}{k-d}\binom{M\cdot(a_j-a_i)}{d}}{\binom{M}{k}} \tag{5}$$

The possible value for d is in range between 2 and $\min(M(a_j - a_i), k)$. The term $\binom{M\cdot(a_j-a_i)}{d}$ represents the combinations where picking d number of samples from the samples which fulfill 4. As d samples are picked from this group, the other $k - d$ samples do not satisfy 4, resulting in the term $\binom{M-M\cdot(a_j-a_i)}{k-d}$.

Because we can suppose a high value for the entire dataset size M, we can assume $M(a_j - a_i)$ to be bigger than k. Thus,

$$p = \frac{\sum_{d=2}^{k} g_d \binom{M-M\cdot(a_j-a_i)}{k-d}\binom{M\cdot(a_j-a_i)}{d}}{\binom{M}{k}} \tag{6}$$

To observe the pattern from the equation, we ran a simulation with $M = 1000$, $N = 20$. Figure 5a illustrates the distribution of probability p, at varying size of the training set and feature importance gap $(a_j - a_i)$. As the sample size decreases or the model attempts to distinguish features with similar importance values, the probability of including essential information declines. The contour of this probability, which is non-linear, aligns with the abrupt drop in stability observed during sample removal in the synthetic datasets (Figure 2).

Although the predictive performance may decline when features are removed, the variables in 6 remain unchanged. In other words, reducing the number of features does not alter the relative importance between the two features that remain in the dataset, ensuring their importance relationship stays stable.

There is a premise behind this section that either 1) the feature importance values of other features are well known to the model enabling it to focus the usage of the information that distinguishes the i-th feature and j-th feature or 2) there are at least two samples satisfying 4 but with all the same feature values except for i and j.

One may wonder whether it is easier to distinguish between important features or unimportant ones. As explained in Appendix VII, under specific assumptions, the probability that a sample satisfies 4 follows a normal distribution with a mean value of 0.5 and a variance inversely proportional to the total number of features. Figure 5b and Figure 5c illustrate the probability of selecting a sample that falls within the specified range. This probability can be calculated by integrating the probability density over the interval $(\theta - a_j, \theta - a_i)$. If the threshold θ is smaller than the mean value 0.5, the range is closer to the central as a_j is smaller, leading to a larger integrated area. This indicates that the probability is higher when comparing two unimportant features. Conversely, when θ is greater than 0.5, the probability becomes higher when two important features, except in cases where a_j is so large that $\theta - a_j$ becomes less than 0.5.

Real datasets often have more negative than positive samples. For example, in the case of medical datasets for certain disease, negative samples (zeros) are easier to obtain than positive samples (ones) [40], [41], [42]. And this balance between zero and one has an impact on threshold value: high concentration on negative cases (zeros) can be represented as a high threshold value. This high threshold value makes important features easier to compare than unimportant ones.

VI. DISCUSSION

This work provides an initial attempt to address key translational challenges in biomedical ML, ensuring that clinicians and biomedical researchers can correctly interpret features importance under real-world data constraints. By providing both theoretical insights and experimental evidence, this work shows that model performance and the validity of feature importance can be independent. In current biomedical data analysis practices, feature importance is often dismissed when models perform poorly, with researchers instead favoring statistical tests that cannot capture non-linear relationships. However, our study demonstrates that even low-performing models, limited by a lack of features, can still provide valid relative comparisons between features. Conversely, models with acceptable performance can produce unstable feature importance results when the number of samples is insufficient. While it is well-established that insufficient data negatively impacts machine learning [43], [44], [45], validity of feature importance and model performance are often mistakenly assumed to be closely linked.

This gap has also been indirectly acknowledged in the statistical literature, where foundational works have emphasized the importance of sufficient sample size to ensure reliable model estimation and inference. For example, Harrell's 10 events-per-variable (EPV) rule and Steyerberg's prognosis research strategies highlight how small datasets can undermine the validity of regression coefficients and model generalizability [46], [47], [48], [49]. These contributions, however, largely approach the issue from the perspective of data scarcity in regression models. In contrast, our study examines the problem through the lens of performance degradation in machine learning models. By framing validity not only in terms of insufficient data but also in terms of insufficient features, we show that low performance does not always invalidate feature importance rankings. Unlike linear regression models, tree-based methods such as random forests can capture non-linear feature contribution, making the distinction between sample-driven and

feature-driven degradation particularly salient. Our findings therefore extend prior statistical insights by demonstrating that degraded performance can still yield valid relative feature importance, provided the loss stems from a lack of features rather than a lack of samples. This offers a machine learning–specific perspective that is particularly relevant in biomedical domains, where predictive accuracy is often constrained by both limited datasets and limited sets of informative features.

It is worth noting that a 'bias phenomenon' has been reported in the literature, where linear classification models using gradient descent tend to overestimate the importance of weakly to moderately predictive features when training samples are insufficient [27]. In our experiments, we were able to observe the same phenomenon, i.e. weak regression predictors tend to overestimate the importance of weak features. In particular, in Section IV-C we observed that, when data size is insufficient, the gap between important and unimportant features can be underestimated.

Machine learning performance can plateau after reaching a certain sample size [43], [50]. Rajput et al. suggested evaluating sample size by examining effect sizes and performance. A possible interpretation of our results is that, if performance remains stable with small reductions but drops significantly with larger ones, low performance is likely due to a lack of features rather than insufficient data.

One limitation of our study is the handling of feature correlations, as discussed in Section III, Section IV-C and Appendix V. We filtered correlations to avoid interactions that could interfere with the primary goal of our framework, which is to compare samples removal and features removal. While we provided an approximate demonstration of the influence of correlations on our framework in Appendix V, further experiments could be conducted to explore this aspect in more detail. Although controlling interactions in real datasets is challenging, correlations or interdependencies could be intentionally introduced in synthetic datasets to better understand their impact.

In this study, all analyses were restricted to Random Forests with Gini impurity. This choice provided a clear and interpretable framework, but also represents a limitation, as the generalizability of the findings to other model classes and attribution methods remains to be established. In particular, methods such as boosting algorithms, neural networks, or SHAP values could exhibit different stability properties, although they ultimately share the same goal of quantifying the contribution of features to predictions. These approaches typically involve higher computational cost and variability, which was one reason we prioritized tree-based models in this first investigation. Moreover, we complemented the empirical experiments with theoretical analyses to ensure that the observed effects were not artifacts of the Random Forest–Gini framework but rather reflected natural consequences of data distribution. Although such simplified theoretical modeling cannot replace empirical validation, it provides reassurance that the findings are not model-specific. Nevertheless, future work should experimentally evaluate whether these patterns consistently appear across diverse model classes and interpretation methods, and investigate to what extent different algorithms may exhibit distinct tendencies in the stability of feature importance.

VII. CONCLUSION

To the best of the authors' knowledge, this study is the first to challenge the misconception that high model performance guarantees the validity of feature importance. By analysing how data limitations (e.g., lack of samples vs. lack of features) affect model performance and feature importance ranking, we provide a clearer understanding of ML constraints in biomedical informatics. Our framework and experiments show that even low-accuracy ML models can yield reliable feature importance rankings when performance loss stems from limited features rather than limited samples. In particular, our findings suggest that importance analyses may still highlight useful biomarkers or guide decision-support tools, even in studies with modest predictive performance.

REFERENCES

[1] M. Iwagami, R. Inokuchi, E. Kawakami, T. Yamada, A. Goto, T. Kuno, Y. Hashimoto, N. Michihata, T. Goto, T. Shinozaki, Y. Sun, Y. Taniguchi, J. Komiyama, K. Uda, T. Abe, and N. Tamiya, "Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study," *PLOS Digit. Health*, vol. 3, no. 8, Aug. 2024, Art. no. e0000578.

[2] F. Huang, X. Sun, A. Mei, Y. Wang, H. Ding, and T. Zhu, "LLM plus machine learning outperform expert rating to predict life satisfaction from self-statement text," *IEEE Trans. Computat. Social Syst.*, vol. 12, no. 3, pp. 1092–1099, Jun. 2025.

[3] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Inf. Fusion*, vol. 81, pp. 84–90, May 2022.

[4] X. Wang, Y. Xiao, X. Xu, L. Guo, Y. Yu, N. Li, and C. Xu, "Characteristics of fecal microbiota and machine learning strategy for fecal invasive biomarkers in pediatric inflammatory bowel disease," *Frontiers Cellular Infection Microbiol.*, vol. 11, Dec. 2021, Art. no. 711884.

[5] A. M. Thomas et al., "Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation," *Nature Med.*, vol. 25, no. 4, pp. 667–678, Apr. 2019.

[6] Y. Lee, M. Cappellato, and B. Di Camillo, "Machine learning–based feature selection to search stable microbial biomarkers: Application to inflammatory bowel disease," *GigaScience*, vol. 12, Dec. 2022, Art. no. 83.

[7] S. Aryal, A. Alimadadi, I. Manandhar, B. Joe, and X. Cheng, "Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease," *Hypertension*, vol. 76, no. 5, pp. 1555–1562, Nov. 2020.

[8] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Sci. Rep.*, vol. 13, no. 1, p. 22588, Dec. 2023.

[9] L. Kapsner, M. Feißt, A. Purbojo, H.-U. Prokosch, T. Ganslandt, S. Dittrich, J. Mang, and W. Wällisch, "Using machine learning and feature importance to identify risk factors for mortality in pediatric heart surgery," *Diagnostics*, vol. 14, no. 22, p. 2587, Nov. 2024.

[10] Y. Lee, J. Seo, and Y.-K. Kim, "AI-assisted diagnostic approach for the influenza-like illness in children: Decision support system for patients and clinicians," *Biomed. Eng. Lett.*, vol. 15, no. 2, pp. 327–336, Mar. 2025.

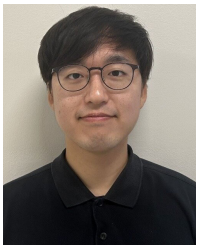
[11] S.-K. Hung, C.-C. Wu, A. Singh, J.-H. Li, C. Lee, E. H. Chou, A. Pekosz, R. Rothman, and K.-F. Chen, "Developing and validating clinical features-based machine learning algorithms to predict influenza infection in influenza-like illness patients," *Biomed. J.*, vol. 46, no. 5, Oct. 2023, Art. no. 100561.

[12] C. F. Tso, C. Lam, J. Calvert, and Q. Mao, "Machine learning early prediction of respiratory syncytial virus in pediatric hospitalized patients," *Frontiers Pediatrics*, vol. 10, Aug. 2022, Art. no. 886212.

[13] L. Zhang, Y. Wang, M. Niu, C. Wang, and Z. Wang, "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: The Henan rural cohort study," *Sci. Rep.*, vol. 10, no. 1, p. 4406, Mar. 2020.
[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[15] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 782–791.
[16] S. Hooker et al., "A benchmark for interpretability methods in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9737–9748.
[17] I. Covert, S. Lundberg, and S. Lee, "Understanding global feature contributions with additive importance measures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 17212–17223.
[18] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18932–18943.
[19] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, "SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training," 2021, *arXiv:2106.01342*.
[20] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1060–1073, 2019.
[21] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 507–520.
[22] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.
[23] Y. Lee, K. Kim, and J. Seo, "CLE-SH: Comprehensive literal explanation package for SHapley values by statistical validity," 2024, *arXiv:2409.12578*.
[24] L. Tolo̧si and T. Lengauer, "Classification with correlated features: Unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, Jul. 2011.
[25] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance," *Statist. Comput.*, vol. 31, no. 6, p. 82, Nov. 2021.
[26] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinf.*, vol. 8, no. 1, p. 25, Dec. 2007.
[27] K. Leino, E. Black, M. Fredrikson, S. Sen, and A. Datta, "Feature-wise bias amplification," 2018, *arXiv:1812.08999*.
[28] Y. Lee and J. Seo, "Suggestion of statistical validation on feature importance of machine learning," in *Proc. 45th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2023, pp. 1–4.
[29] S. Janitza, C. Strobl, and A.-L. Boulesteix, "An AUC-based permutation variable importance measure for random forests," *BMC Bioinf.*, vol. 14, no. 1, p. 119, Dec. 2013.
[30] B. Antal and A. Hajdu, "Diabetic retinopathy debrecen," UCI Mach. Learn. Repository, Univ. California, Irvine, CA, USA, Tech. Rep. C5XP4P, 2014, doi: 10.24432/C5XP4P.
[31] M. Lubicz, K. Pawelczyk, A. Rzechonek, and J. Kolodziej, "Thoracic surgery data," UCI Mach. Learn. Repository, Univ. California, Irvine, CA, USA, Tech. Rep. C5Z60N, 2014, doi: 10.24432/C5Z60N.
[32] K. L. Cios and L. Goodenday, "SPECTF heart," UCI Mach. Learn. Repository, Univ. California, Irvine, CA, USA, Tech. Rep. C5N015, 2001, doi: 10.24432/C5N015.
[33] A. Antony, "Metabolic syndrome," Kaggle, Inc., San Francisco, CA, USA, 2023. [Online]. Available: https://www.kaggle.com/datasets/antimoni/metabolic-syndrome
[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[35] F. Prodregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.
[36] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?" *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, Nov. 2018.
[37] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*. Hoboken, NJ, USA: Wiley, 2013.
[38] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *J. Thoracic Oncol.*, vol. 5, no. 9, pp. 1315–1316, Sep. 2010.
[39] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, Dec. 2020.
[40] L. Liu, X. Wu, S. Li, Y. Li, S. Tan, and Y. Bai, "Solving the class imbalance problem using ensemble algorithm: Application of screening for aortic dissection," *BMC Med. Informat. Decis. Making*, vol. 22, no. 1, p. 82, Dec. 2022.
[41] A. Jain, S. Ratnoo, and D. Kumar, "Addressing class imbalance problem in medical diagnosis: A genetic algorithm approach," in *Proc. Int. Conf. Inf., Commun., Instrum. Control (ICICIC)*, Aug. 2017, pp. 1–8.
[42] L. Gao, L. Zhang, C. Liu, and S. Wu, "Handling imbalanced medical image data: A deep-learning-based one-class classification approach," *Artif. Intell. Med.*, vol. 108, Aug. 2020, Art. no. 101935.
[43] D. Rajput, W.-J. Wang, and C.-C. Chen, "Evaluation of a decided sample size in machine learning applications," *BMC Bioinf.*, vol. 24, no. 1, p. 48, Feb. 2023.
[44] I. Balki, A. Amirabadi, J. Levman, A. L. Martel, Z. Emersic, B. Meden, A. Garcia-Pedrero, S. C. Ramirez, D. Kong, A. R. Moody, and P. N. Tyrrell, "Sample-size determination methodologies for machine learning in medical imaging research: A systematic review," *Can. Assoc. Radiologists J.*, vol. 70, no. 4, pp. 344–353, Nov. 2019.
[45] Z. Cui and G. Gong, "The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features," *NeuroImage*, vol. 178, pp. 622–637, Sep. 2018.
[46] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati, "Regression modelling strategies for improved prognostic prediction," *Statist. Med.*, vol. 3, no. 2, pp. 143–152, Apr. 1984.
[47] M. van Smeden, J. A. H. de Groot, K. G. M. Moons, G. S. Collins, D. G. Altman, M. J. C. Eijkemans, and J. B. Reitsma, "No rationale for 1 variable per 10 events criterion for binary logistic regression analysis," *BMC Med. Res. Methodol.*, vol. 16, no. 1, p. 163, Dec. 2016.
[48] E. W. Steyerberg, H. Uno, J. P. A. Ioannidis, B. van Calster, C. Ukaegbu, T. Dhingra, S. Syngal, and F. Kastrinos, "Poor performance of clinical prediction models: The harm of commonly applied methods," *J. Clin. Epidemiol.*, vol. 98, pp. 133–143, Jun. 2018.
[49] G. S. Collins, E. O. Ogundimu, and D. G. Altman, "Sample size considerations for the external validation of a multivariable prognostic model: A resampling study," *Statist. Med.*, vol. 35, no. 2, pp. 214–226, Jan. 2016.
[50] J. Faber and L. M. Fonseca, "How sample size influences research outcomes," *Dental Press J. Orthodontics*, vol. 19, no. 4, pp. 27–29, Aug. 2014.

YOUNGRO LEE received the Ph.D. degree in electrical and computer engineering from Seoul National University in February 2025. He is currently a Post-Doctoral Researcher at Seoul National University. His research centers on applying machine learning and explainable AI to diverse medical and biological domains, including microbiome biomarker discovery, influenza-like illness decision support systems, and knowledge discovery for metabolic health. He has collaborated closely with clinicians across multiple institutions to ensure the validity and interpretability of AI models. His broader interest lies in developing trustworthy and reproducible machine learning applications in healthcare and computational biology. He also conducted a research visit at the University of Padova in 2021 and has since continued collaborative research on computational biology and machine learning pipelines.

GIACOMO BARUZZO received the M.Sc. degree in computer engineering and the Ph.D. degree in information engineering from the University of Padova in 2013 and 2017, respectively. His Ph.D. and post-doctoral research focused on bioinformatics and high-performance computing, with a particular emphasis on computational methods for mining biological data obtained from high-throughput sequencing technologies. From 2017 to 2025, he was a Post-Doctoral Researcher at the Department of Information Engineering, University of Padova. In 2025, he was appointed as a tenure-track Assistant Professor of computer engineering at the University of Padova. His current research interests include the design and development of efficient algorithms, methods, and software for the preprocessing, analysis, and interpretation of large-scale biological and omics data.



JEONGHWAN KIM received the B.S. degree in electrical and computer engineering from Seoul National University, South Korea, and the M.S. degrees in electrical and computer engineering and mathematics from Georgia Institute of Technology, USA. He is currently pursuing the Ph.D. degree in robotics with Georgia Tech. His research interests include robot learning, physics-based animation, and machine learning-based control for legged robots.

JONGMO SEO (Member, IEEE) received the M.D., M.S., and Ph.D. degrees in biomedical engineering from Seoul National University College of Medicine in 1996, 2002, and 2005, respectively. He completed residency and fellowship training in ophthalmology at SNU Hospital.

He is currently a Professor with the Department of Electrical and Computer Engineering, SNU, and the Chair of the Interdisciplinary Program for Medical Informatics, Graduate School of Medicine. His research focuses on medical informatics, neural interfaces, and biomedical engineering, developing innovative technologies, such as artificial retinas and implantable biosensors for advanced healthcare solutions.

BARBARA DI CAMILLO received the M.Sc. degree in electronic engineering from the University of Padova in 2000 and the Ph.D. degree in bioengineering in 2004. Her Ph.D. thesis was on modeling dynamic gene expression profiles. She was a Visiting Scientist at the Mayo Clinic, USA, and the University of Technology of Graz, Austria, and held a post-doctoral position at the University of Padova. She has been a Faculty Member at the Department of Information Engineering, University of Padova, since 2006, serving as Assistant Professor from 2006 to 2017, an Associate Professor from 2017 to 2020, and a Full Professor since 2020. Her research focuses on the development and application of advanced modeling, data mining, and machine learning methods for high-throughput biological and clinical data. She has extensive expertise in predictive modeling, biomarker discovery, and biological network inference using approaches such as differential equations, Boolean models, and Bayesian networks.