WORLD FORUM ON ACADEMIC INNOVATION AND DIGITAL LITERACY

COMPARING FEATURE SELECTION METHODS AND EXPLAINABLE AI TECHNIQUES FOR INTERPRETABLE MEDICAL CLASSIFICATION

G.Kh. Tashmatov

Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

E-mail: ganijon.tashmatov@gmail.com

Abstract

Feature selection and explainable artificial intelligence (XAI) are often used together in medical machine learning, but they answer different questions. Feature selection identifies variables to retain before training, whereas XAI explains which variables influenced a trained model. This study compares six feature selection methods - ANOVA F-test, mutual information, recursive feature elimination, L1 logistic regression, random forest importance, and extra trees importance - with SHAP values, permutation importance, and model-based importance. Experiments were conducted on three public medical classification datasets: Breast Cancer Wisconsin Diagnostic, Pima Indians Diabetes, and Heart Disease Cleveland. Logistic regression, random forest, SVM, and gradient boosting were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Agreement between selected and explained features was measured by Jaccard similarity and overlap percentage. The best full-feature ROC-AUC values were 0.998 for Breast Cancer, 0.823 for Pima Diabetes, and 0.862 for Heart Disease. Mean Jaccard agreement was moderate: 0.612, 0.630, and 0.577 respectively. The results show that feature selection and XAI are complementary tools for interpretable medical classification.

Keywords: feature selection; explainable artificial intelligence; SHAP; permutation importance; medical machine learning; interpretable classification.

1. Introduction

Medical machine learning increasingly requires predictive performance and transparent reasoning. In clinical decision support, a model with high accuracy but unclear decision logic is difficult to audit, validate, or trust. This is why recent work on explainable artificial intelligence emphasizes interpretability, especially in healthcare and other safety-critical domains [1, 2].

Feature selection is a classical way to improve interpretability. It reduces the number of variables before training and can make a model easier to inspect. However, feature selection does not necessarily explain the behavior of a final fitted model. XAI methods, such as SHAP values and permutation importance, operate after model training and estimate which features influenced predictions. Therefore, a variable may be statistically relevant but weakly used by a particular classifier, while another variable may become important through correlation or model-specific interactions.

The research question of this article is: to what extent do classical feature selection methods agree with XAI-based top features in medical classification? The main idea is that feature selection answers the question "which features should be kept before learning?", while XAI answers the question "which features influenced the trained model?".

2. Related Work

Feature selection methods are usually divided into filter, wrapper, and embedded approaches. Filter methods, such as ANOVA F-test and mutual information, rank features using statistical association. Wrapper methods, such as recursive feature

elimination, search for feature subsets using an estimator. Embedded methods, including L1-regularized logistic regression and tree-based importance, perform selection during model fitting. Reviews in disease-risk prediction show that the usefulness of these methods depends on the dataset, sample size, feature correlation, and clinical objective [3].

XAI methods provide another view of interpretability. SHAP values estimate feature contributions based on Shapley values, permutation importance measures the drop in model quality when a feature is shuffled, and model-based importance uses internal model statistics. Recent studies show growing interest in connecting SHAP-based explanations with feature selection [4, 5]. However, not all studies quantify how much feature selection and XAI agree. This work fills this gap by using Jaccard similarity and overlap percentage on medical classification tasks.

3. Materials and Methods

3.1. Datasets and preprocessing

Three public medical datasets were used. Breast Cancer Wisconsin Diagnostic was loaded through scikit-learn and corresponds to the UCI diagnostic breast cancer dataset. Pima Indians Diabetes and Heart Disease Cleveland CSV files were included locally to make the experiment reproducible offline. Missing values were imputed with the median fitted only on the training set. For Pima Diabetes, zero values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI were treated as missing because they are physiologically invalid in this context. All experiments used a stratified 75/25 train/test split with random_state = 42.

Table 1. Dataset summary.

Dataset	Instances	Features	Positive class	Top k

Breast Cancer Wisconsin Diagnostic	569	30	malignant	10
Pima Indians Diabetes	768	8	diabetes = 1	4
Heart Disease Cleveland	303	13	disease presence	6

3.2. Feature selection, models, and XAI

Six feature selection methods were compared: ANOVA F-test, mutual information, RFE with logistic regression, L1-regularized logistic regression, random forest importance, and extra trees importance. The same top-k value was used for feature selection and XAI ranking in each dataset to make comparison fair.

Four classifiers were evaluated: logistic regression, random forest, SVM with radial basis function kernel, and gradient boosting. Logistic regression and SVM were trained inside scaling pipelines. Feature selection was always fitted only on the training partition to avoid data leakage. XAI analysis was performed on a random forest model trained on the full feature space. Three explanation views were extracted: mean absolute SHAP values, permutation importance based on ROC-AUC decrease, and native random forest model-based feature importance.

Agreement was measured between the feature set selected by a method S and the top-k feature set explained by an XAI method E. Jaccard similarity was computed as |S intersection E| / |S union E|, while overlap percentage was computed as 100 x |S intersection E| / k.

4. Results

4.1. Classification performance

Table 2 summarizes the most important predictive results. Breast Cancer was the easiest dataset: random forest reached the highest full-feature ROC-AUC of 0.998 and SVM reached 0.993 accuracy. Pima Diabetes was more difficult, with logistic regression achieving the best full-feature ROC-AUC of 0.823. Heart Disease Cleveland reached the best full-feature ROC-AUC of 0.862 with random forest. Selected-feature models remained competitive; in each dataset, random forest-selected features with SVM produced the strongest compact configuration.

Table 2. Main performance results.

Dataset	**Full-feature result**	**Selected-feature result**	**Key features**
Breast Cancer	RF AUC=0.998; SVM Acc=0.993	RF+SVM AUC=0.999	concavity; perimeter
Pima Diabetes	LR AUC=0.823	RF+SVM AUC=0.824	Glucose; BMI; Age
Heart Disease	RF AUC=0.862	RF+SVM AUC=0.876	cp; oldpeak; thalach

4.2. Agreement between feature selection and XAI

Agreement was moderate rather than perfect. Table 3 shows the mean Jaccard and overlap values by dataset and XAI method. Breast Cancer showed strong agreement with model-based importance and SHAP, but weaker agreement with permutation importance. Heart Disease showed the strongest mean agreement with SHAP, while

Pima Diabetes showed high agreement with model-based importance. These results confirm that feature selection and XAI are related but not identical.

Table 3. Mean Jaccard similarity and overlap percentage.

Dataset	**Model-based Jaccard**	**Permutation Jaccard**	**SHAP Jaccard**	**Mean overlap range**
Breast Cancer	0.719	0.452	0.666	61.7-80.0%
Pima Diabetes	0.733	0.578	0.578	70.8-83.3%
Heart Disease	0.571	0.444	0.714	61.1-83.3%

4.3. Visual analysis

The following figures show representative visual outputs. To reduce the article length, only the most informative figures are kept in the text: ROC curves, SHAP summaries, and Jaccard heatmaps. The remaining figures can be stored as supplementary project files.



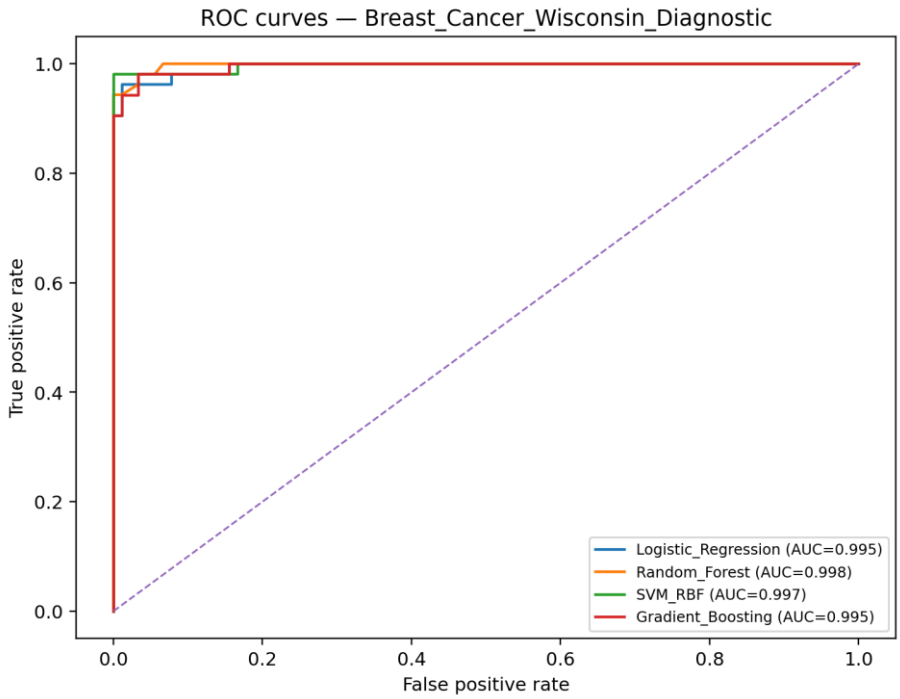


Figure 1. ROC curves for Breast Cancer Wisconsin Diagnostic.

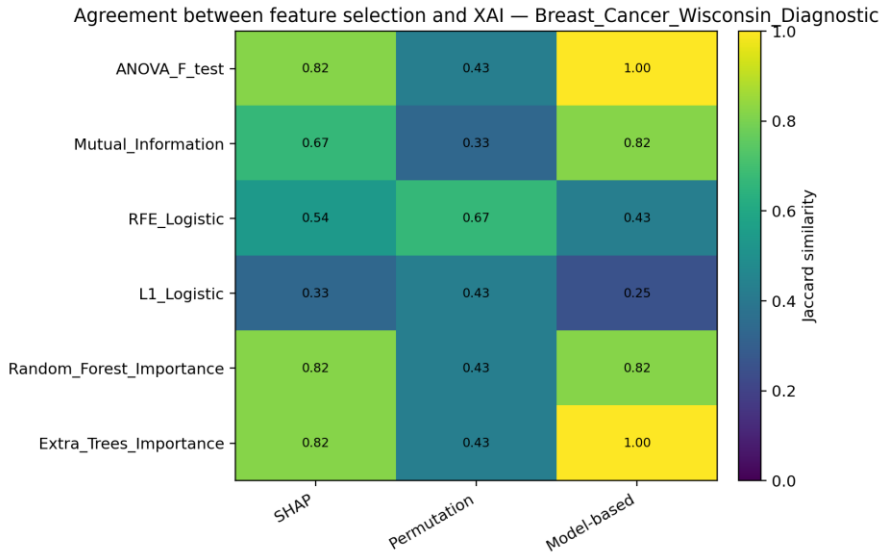


Figure 2. Jaccard similarity heatmap for Breast Cancer Wisconsin Diagnostic.



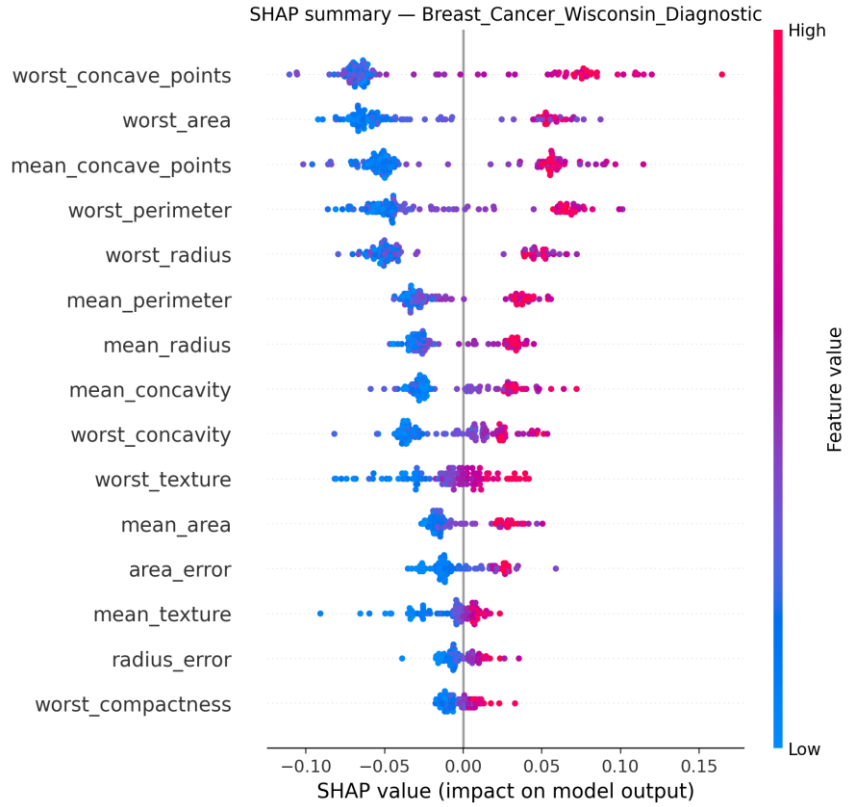


Figure 3. SHAP summary plot for Breast Cancer Wisconsin Diagnostic.

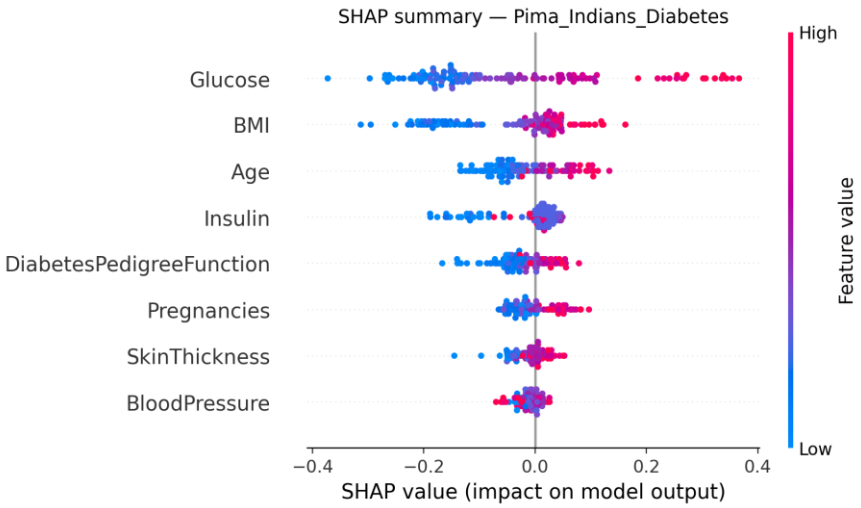


Figure 4. SHAP summary plot for Pima Indians Diabetes.

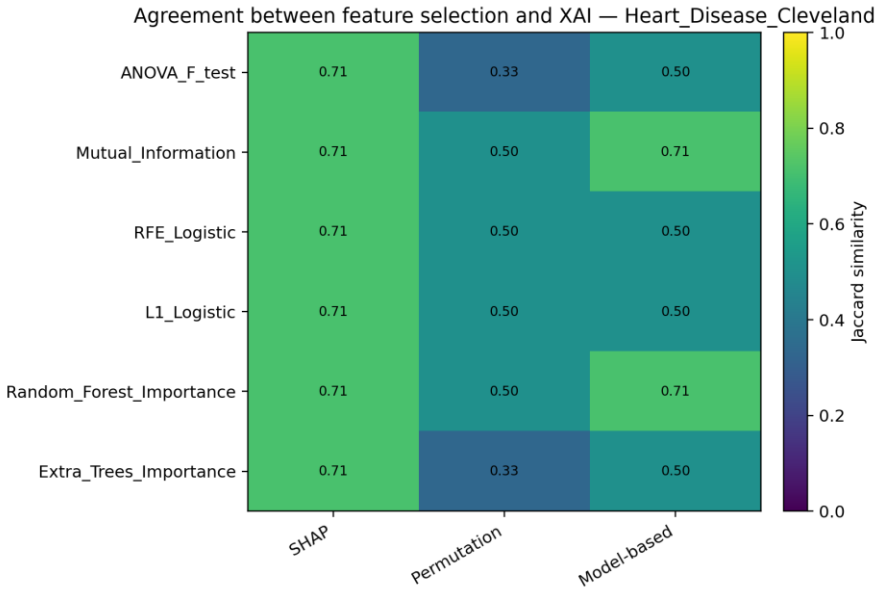


Figure 5. Jaccard similarity heatmap for Heart Disease Cleveland.

5. Discussion

The results support the central hypothesis that feature selection and XAI are complementary rather than interchangeable. Feature selection identifies compact candidate variable sets before training. XAI audits how a trained model used variables on the test distribution. Because these questions differ, the resulting feature lists can overlap without being identical.

The Breast Cancer dataset illustrates relatively high agreement. Morphological features related to concavity, perimeter, radius, and area appeared repeatedly across selectors and explanations. Pima Diabetes also showed a clear core of important variables, especially Glucose and BMI. Heart Disease Cleveland produced lower average agreement, probably because the dataset is smaller and contains mixed clinical variables. In practical medical ML, high overlap can strengthen confidence in a compact model, while low overlap should trigger checks for correlated variables, data leakage, instability, or model-specific artifacts.

6. Limitations

This study used public benchmark datasets rather than hospital-level clinical cohorts. Only one fixed train/test split was used, so repeated cross-validation and external validation would be needed for stronger statistical conclusions. SHAP, permutation importance, and model-based importance were computed for a random forest XAI base model; other trained models may yield different explanations. Finally, top-k agreement simplifies importance rankings into sets and does not fully represent uncertainty in importance estimates.

7. Conclusion

This article compared six feature selection methods and three XAI techniques on three public medical classification datasets. The experiment showed that feature selection can preserve strong predictive performance with fewer variables, but selected features do not always match the features emphasized by XAI. Mean Jaccard agreement ranged from 0.577 to 0.630 across datasets, indicating moderate agreement rather than equivalence. Therefore, feature selection and XAI should be used together: the first builds compact models before training, and the second audits model behavior after fitting.

References

[1] Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., et al. (2024). A review of Explainable Artificial Intelligence in healthcare. Computers and Electrical Engineering, 118, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

[2] Zhang, K., et al. (2026). A comprehensive review of explainable artificial intelligence in healthcare: methods, evaluation, and clinical integration. iScience.

https://doi.org/10.1016/j.isci.2026.115026

[3] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2, 927312. https://doi.org/10.3389/fbinf.2022.927312

[4] Dunn, J., Mingardi, L., & Zhuo, Y. D. (2021). Comparing interpretability and explainability for feature selection. arXiv:2105.05328.

[5] Zacharias, J., et al. (2022). Designing a feature selection method based on explainable artificial intelligence. Electronic Markets. https://doi.org/10.1007/s12525-022-00608-1