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 both used to improve the interpretability of medical machine learning, but they answer different questions. Feature selection asks which variables should be retained before training, whereas XAI asks which variables influenced an already trained model. This study compares six feature selection methods (ANOVA F-test, mutual information, RFE, L1 logistic regression, random forest importance, and extra trees importance) with three XAI views (SHAP values, permutation importance, and model-based random forest importance). Experiments were performed in Python on Breast Cancer Wisconsin Diagnostic, Pima Indians Diabetes, and Heart Disease Cleveland datasets. Logistic regression, random forest, SVM, and gradient boosting were evaluated by accuracy, precision, recall, F1-score, and ROC-AUC. Agreement between selected and explained top features was measured with Jaccard similarity and overlap percentage. The best full-feature ROC-AUC values were 0.998, 0.823, and 0.862 for the three datasets, respectively. Mean Jaccard agreement was moderate and dataset-dependent: 0.612 for Breast Cancer, 0.630 for Pima Diabetes, and 0.577 for Heart Disease. The results show that feature selection and XAI are complementary rather than interchangeable.

Keywords: feature selection; explainable artificial intelligence; SHAP;

permutation importance; medical machine learning; interpretable classification; Jaccard similarity

1. Introduction

Medical classification models must be accurate and interpretable, because clinical decisions require auditability and trust. Classical feature selection improves interpretability by reducing the input space before model training, but it does not necessarily explain the behavior of a specific fitted classifier. XAI methods, including SHAP and permutation importance, provide post-hoc evidence about which features affected model predictions.

This article studies whether classical feature selection and XAI identify similar medical variables. The main research question is: to what extent do selected top features coincide with features that a trained model considers important?

2. Related Work

Feature selection is usually divided into filter, wrapper, and embedded methods. ANOVA F-test and mutual information are filter methods; RFE is a wrapper method; L1 logistic regression and tree-based importance are embedded approaches. Recent reviews note that the usefulness of these methods in disease-risk prediction depends on sample size, feature correlation, and clinical objectives [3].

XAI research in healthcare emphasizes transparency, evaluation, and clinical integration [1, 2]. SHAP values, LIME, permutation importance, and model-based importance are common tools for interpreting predictions [4, 9, 10]. However, feature selection and XAI are often reported separately, while their agreement is less frequently quantified. This study addresses that gap by using Jaccard similarity and overlap percentage.

3. Materials and Methods

Three public medical classification datasets were used. Breast Cancer Wisconsin Diagnostic was loaded from scikit-learn; Pima Indians Diabetes and Heart Disease

Cleveland CSV files were included locally to keep the experiment reproducible in an offline environment. The Heart Disease target was binarized as absence or presence of disease.

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 outcome = 1	4
Heart Disease Cleveland	303	13	presence of heart disease	6

All experiments used a stratified 75/25 train/test split with random_state = 42. Missing values were imputed with the median fitted only on the training partition. In the Pima dataset, physiologically invalid zeros in Glucose, BloodPressure, SkinThickness, Insulin, and BMI were treated as missing. Logistic regression and SVM were trained inside scaling pipelines. Feature selection was fitted only on the training set to avoid data leakage.

The feature selection methods were ANOVA F-test, mutual information, RFE with logistic regression, L1-regularized logistic regression, random forest importance, and extra trees importance. The XAI comparison used mean absolute SHAP values, permutation importance measured by ROC-AUC decrease, and native random forest importance. For a fair comparison, the same top-k value was used for both selected and explained feature sets.

Model performance was assessed by accuracy, precision, recall, F1-score, and ROC-

AUC. For a selected feature set S and an explained top-feature set E, agreement was computed as Jaccard(S,E)=|S intersection E|/|S union E| and Overlap(S,E)=100 x |S intersection E|/k.

4. Results

Table 2 summarizes the best full-feature model and the best selected-feature configuration for each dataset. Feature selection preserved strong predictive performance while reducing the number of variables. In Breast Cancer and Pima Diabetes, selected-feature SVM configurations reached ROC-AUC values comparable to or slightly higher than the best full-feature models.

Table 2. Best full-feature and selected-feature results by ROC-AUC.

Dataset	Feature set	Model	FS method	N	Acc.	F1	ROC-AUC
Breast Cancer Wisconsin Diagnostic	All features	Random Forest	-	30	0.965	0.950	0.998
Breast Cancer Wisconsin Diagnostic	Selected features	SVM RBF	Random Forest Importance	10	0.972	0.962	0.999
Pima Indians Diabetes	All features	Logistic Regression	-	8	0.750	0.662	0.823
Pima Indians Diabetes	Selected features	SVM RBF	Random Forest Importance	4	0.760	0.693	0.824

Heart Disease Cleveland	All features	Random Forest	-	13	0.789	0.818	0.862
Heart Disease Cleveland	Selected features	SVM RBF	Random Forest Importance	6	0.789	0.814	0.876

Agreement between feature selection and XAI was moderate rather than complete. Breast Cancer reached the strongest average model-based agreement, Pima Diabetes showed stable overlap for Glucose and BMI, and Heart Disease produced the lowest mean Jaccard score, reflecting the smaller sample size and mixed clinical variables.

Table 3. Mean agreement between feature selection and XAI methods.

Dataset	Model J	Perm. J	SHAP J	Model %	Perm. %	SHAP %	Mean J
Breast Cancer Wisconsin Diagnostic	0.719	0.452	0.666	80.0	61.7	78.3	0.612
Pima Indians Diabetes	0.733	0.578	0.578	83.3	70.8	70.8	0.630
Heart Disease Cleveland	0.571	0.444	0.714	72.2	61.1	83.3	0.577

The most frequently selected and explained variables were clinically plausible. Breast Cancer outputs emphasized concavity, perimeter, radius, and area measurements. Pima Diabetes emphasized Glucose, BMI, Age, and DiabetesPedigreeFunction. Heart

Disease emphasized chest pain type (cp), ST depression (oldpeak), maximum heart rate (thalach), thalassemia category (thal), and major-vessel count (ca).

5. Discussion

The results support the central idea that feature selection and XAI are complementary. Feature selection is useful for building compact candidate models before training, while XAI audits the behavior of the trained model after fitting. High overlap between the two strengthens confidence that the model relies on robust and clinically meaningful features. Low overlap should motivate additional checks for correlated variables, unstable selection, model-specific artifacts, or possible leakage.

The Breast Cancer dataset produced high but not perfect agreement because many diagnostic measurements are strongly correlated. Pima Diabetes showed consistent agreement on Glucose and BMI, but competing variables such as Age, Insulin, Pregnancies, and DiabetesPedigreeFunction changed order across methods. Heart Disease had the lowest agreement, which is expected for a smaller dataset with mixed categorical and numeric variables.

6. Limitations

This experiment used public benchmark datasets rather than institution-level clinical cohorts, so the results demonstrate methodological behavior, not clinical readiness. Only one fixed train/test split was used; 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, and other model families may yield different explanation patterns. Finally, top-k agreement is easy to interpret but discards rank-order uncertainty.

7. Conclusion

This article compared six feature selection methods and three XAI techniques on three public medical classification datasets. The best full-feature ROC-AUC values were 0.998 for Breast Cancer, 0.823 for Pima Diabetes, and 0.862 for Heart Disease.

Selected-feature models remained competitive while using fewer variables. Mean Jaccard agreement between feature selection and XAI ranged from 0.577 to 0.630, showing moderate agreement rather than equivalence. Therefore, feature selection and XAI should be used together: feature selection reduces the model before training, and XAI verifies how the trained model actually uses the retained information.

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

[6] Kraev, E., Koseoglu, B., Traverso, L., & Topiwalla, M. (2024). shap-select: Lightweight Feature Selection Using SHAP Values and Regression. arXiv:2410.06815.

[7] Salih, A., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Menegaz, G., & Lekadir, K. (2023). A Perspective on Explainable Artificial Intelligence

Methods: SHAP and LIME. arXiv:2305.02012.

[8] Ponce-Bobadilla, A. V., et al. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clinical and Translational Science.

[9] scikit-learn developers. Feature selection. scikit-learn User Guide. https://scikit-learn.org/stable/modules/feature_selection.html

[10] scikit-learn developers. Permutation feature importance. scikit-learn User Guide. https://scikit-learn.org/stable/modules/permutation_importance.html

[11] UCI Machine Learning Repository. Breast Cancer Wisconsin (Diagnostic) dataset. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[12] UCI Machine Learning Repository. Heart Disease dataset. https://archive.ics.uci.edu/dataset/45/heart+disease

[13] Kaggle / UCI ML. Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

[14] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.