
Journal of Informatics and Web Engineering

Vol. 5 No. 3 (October 2026)

eISSN: 2821-370X

Empirical Evaluation of Feature Selection Strategies for Interpretable and Robust Healthcare Machine Learning

Md. Abu Nishan¹, Choo-Yee Ting^{2*}, Chih-Yang Pee³, Helmi Zakariah⁴, Thanh-Trung Nguyen^{5}**

^{1,2,3}Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Selangor, Malaysia

⁴Hayat Technologies Sdn Bhd, Level 15, Centerpoint North, Lingkaran Syed Putra, Mid Valley City, 59200 Kuala Lumpur, Malaysia

⁵Faculty of Information Technology - Communications, Dong Nai Technology University, Dong Nai, Vietnam

* Corresponding author: (cyting@mmu.edu.my; ORCID: 0000-0001-5667-2816)

** Corresponding author: (nguyenthanhtrung@dntu.edu.vn; ORCID: 0009-0000-9948-1753)

Abstract - Feature Selection (FS) is a fundamental step in developing accurate, robust, and interpretable Machine Learning (ML) models, especially when working with high-dimensional structured datasets commonly found in healthcare, finance, and other data-intensive domains. Selecting the most relevant subset of features not only improves predictive performance but also enhances computational efficiency and model transparency. This paper presents a comprehensive and systematic study of widely adopted feature selection strategies, including filter, wrapper, embedded, model-agnostic methods, and dimensionality reduction techniques. Each category is examined through experimental evaluations across multiple classifiers and benchmark datasets to understand how different feature selection mechanisms affect model performance, interpretability, and stability. The results show that while filter methods, such as mutual information and chi-square, are computationally efficient and suitable for large-scale screening, wrapper and embedded methods achieve higher accuracy at the cost of increased computational demand. The model-agnostic approach, represented by Permutation Feature Importance (PFI), was employed to assess the relative contribution of each feature to model performance, thereby providing an interpretable and unbiased measure of feature relevance. Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), further assist in managing multicollinearity, improving class separability, and enhancing model generalization by transforming features into more discriminative lower-dimensional spaces. Moreover, the study captures the trade-offs between accuracy, speed, and explainability, offering practical recommendations for researchers and practitioners to select suitable methods based on their specific context and computational resources. The findings underscore the importance of adaptive, context-aware feature selection pipelines, particularly in healthcare analytics, where interpretable and trustworthy models can improve diagnostic confidence and support effective decision-making.

Keywords—Machine Learning, Feature Selection, Explainable AI, Data Analytics, Predictive Modelling.

Received: 14 July 2025; Accepted: 18 November 2025; Published: 16 October 2026

This is an open access article under the [CC BY-NC-ND 4.0](#) license.



1. INTRODUCTION

Machine Learning (ML) models are often used on big datasets with many features or variables. These datasets have much information, but not all features help the model predict well [1]. In fact, some features are not useful, redundant, or noisy, which can hurt how well the model works [2]. Irrelevant features can make the model learn noise instead of real patterns, so it will not work well on new data [3]. Also, redundant features can make the model too complex. This can cause long training times and more computing costs [4]. Lastly, using useless features can make the model hard to understand. This is a big problem in fields like healthcare, finance, and cybersecurity [5].

Feature Selection (FS) is an important step before using ML to solve these problems [6]. It means picking the most useful features and removing the irrelevant features that do not help the model perform well. This step has two big aims: (1) help the model perform better by removing noise and weak features, and (2) make the model easier to understand, so people can trust what it shows. In areas such as healthcare and finance, where mistakes can have serious consequences, interpretable models are crucial. People need to trust and understand that the model's decisions match expert knowledge and ethics [7].

The value of FS is not only to make better predictions. It is also important for making the model clear to users. In areas where the model's choices can affect lives, like hospitals or banks, it is very important that people can understand the model [8]. People who use the model, such as doctors, leaders, or patients, need to know how it makes its choices. They need to see what factors affect these choices. Without this, it is hard to trust the model. Sometimes, people might even ignore correct results if they cannot understand them [9].

In this study, our novelty does not lie in proposing a new FS method. Instead, we conduct a systematic and empirical comparison of multiple FS strategies on a large healthcare dataset, rather than proposing a new algorithm. We systematically assess how different methods affect model performance, interpretability, stability, and computational efficiency. We focus on many types of FS, like filter, wrapper, embedded, and model-agnostic methods. By evaluating them in a healthcare context, we provide insights that are useful for clinical decision-making and practical guidance for practitioners and researchers. We use many classifiers and ways to check the results of each method. This makes sure our results are strong and useful for many ML problems.

Our contributions are threefold: First, we provide a comprehensive and practical evaluation of various FS methods, revealing their strengths and limitations in healthcare datasets. Second, we investigate the trade-offs between predictive performance, model interpretability, stability, and computational cost, offering guidance for real-world deployment. Third, we emphasize that the novelty of our work lies in the comparative evaluation of FS methods on healthcare data, showing healthcare-specific insights and patterns that can help medical decision-making, clinical practice, and policy.

2. LITERATURE REVIEW

Previous studies have shown that feature selection (FS) is very important for improving the performance of machine learning (ML) models. However, few works have provided a comprehensive comparison that simultaneously focuses on both interpretability and stability. Recent studies indicate that combining FS with model explanation tools helps users better understand how models make decisions. Nevertheless, there is still a lack of systematic evaluation of multiple FS methods that balance predictive power and interpretability across diverse models and datasets. This study aims to fill that gap by conducting empirical tests on model performance together with assessments of explainability.

Table 1 summarizes recent studies on FS methods using ML algorithms, including their datasets/domains, methods, and main limitations. Khaire and Dhanalakshmi [10] provided a comprehensive review focusing on FS stability; however, the work is mostly theoretical and lacks sufficient testing on healthcare data. Wang et al. [11] compared Shapley Additive exPlanations (SHAP) with built-in feature importance on the Kaggle credit card fraud dataset, offering useful insights into model explainability. However, model-based importance methods often outperform SHAP on large datasets, making SHAP less practical in such cases. Theng and Bhoyar [12] conducted an extensive survey of filter, wrapper, and embedded methods over more than two decades, but the study remains primarily theoretical and does not adequately address noisy or imbalanced data, particularly in healthcare. Bouke et al. [13] proposed the BukaGini algorithm, which integrates stability into the Gini index for more robust FS; however, it was tested on only a few datasets, so its real-world applicability requires further validation. Wang et al. [14] demonstrated FS applications in perovskite materials discovery, but the domain-specific focus limits its generalization to other areas such as

healthcare. Zhang and Xiong [15] introduced SpoFe, a statistical self-supervised learning approach combining kernel PCA with sparse polynomial and knockoff techniques for high-dimensional tabular data. Although promising, its high computational cost and narrow focus may hinder broader adoption. Kodipalli et al. [16] developed an ensemble explainable AI approach using SHAP and statistical tests on a multi-modal ovarian cancer dataset. While valuable, the findings are domain-specific and may not generalize to other diseases. Lyu et al. [17] explored interpretable ML for FS in neural learning-to-rank tasks, providing useful insights into ranking models, though its applicability to other ML tasks remains uncertain. Moraffah et al. [18] proposed a causal FS framework for fair ML applications, but the survey lacks sufficient experimental validation across multiple domains. Jahin et al. [19] introduced KAN-DDQN, a human-in-the-loop reinforcement learning approach for instance-level FS, which performed well on image datasets such as MNIST and Fashion-MNIST. However, it requires substantial human feedback and is designed mainly for image data, making it difficult to generalize to other domains. Nagaraju [20] proposed an RL-based autoencoder for FS that may work effectively on various tabular datasets, though it has not yet been tested across different fields. Zhao et al. [21] developed a knockoff-based explainable AI method with False Discovery Rate (FDR) control for FS in sports data, but its applicability to other domains remains to be demonstrated. Shaik et al. [22] integrated attention mechanisms with SHAP in deep learning for FS in medical image and tabular data. Although potentially effective, it is still a new approach that requires further research to balance interpretability and performance. Barbieri et al. [23] evaluated FS techniques in terms of performance and stability using UCI and real-world datasets but omitted deep learning and reinforcement learning methods. Lamsaf et al. [24] provided a comprehensive review of causal ML approaches for FS; however, the lack of empirical validation limits the assessment of their practical utility.

Table 1. Summary of Recent Studies on FS Methods using ML Algorithms

Ref.	Authors & Year	Dataset / Domain	Method	Limitation
[10]	Khaire & Dhanalakshmi (2019)	General ML (review)	Review of FS techniques & instability	Theoretical; limited empirical validation
[11]	Wang et al. (2024)	Credit card fraud (Kaggle)	SHAP vs built-in importance	SHAP underperforms on large datasets
[12]	Theng & Bhoyar (2023)	Mixed ML datasets (20+ years)	Survey of filter, wrapper, embedded FS	Limited testing on noisy/imbalanced data
[13]	Bouke et al. (2023)	Standard ML datasets	BukaGini: Gini-index with stability	Narrow dataset scope; needs real-world validation
[14]	Wang et al. (2023)	Perovskite materials	ML-based FS for materials discovery	Domain-specific; not transferable
[15]	Zhang & Xiong (2025)	High-dimensional tabular data	SpoFe: PCA + sparse poly + knockoffs	High computation; specialized domain
[16]	Kodipalli et al. (2025)	Ovarian cancer (multi-modal)	Ensemble explainable AI (SHAP + stats)	Domain-limited; lacks cross-disease testing
[17]	Lyu et al. (2024)	Neural LTR datasets	Interpretable FS for ranking tasks	Task-specific; generalization unclear
[18]	Moraffah et al. (2024)	Responsible ML cases	Causal FS framework	Survey-based; lacks empirical testing
[19]	Jahin et al. (2026)	MNIST, Fashion-MNIST	KAN-DDQN: RL with human feedback	Image-focused; requires manual input
[20]	Nagaraju (2025)	Tabular datasets	RL-based autoencoder for FS	Preprint; cross-domain use untested
[21]	Zhao et al. (2025)	Sports analytics	Knockoff-based XAI with FDR control	Domain-specific; generalizability unproven
[22]	Shaik et al. (2025)	Medical imaging/tabular	Attention + SHAP in DL	Early-stage; trade-off accuracy vs explainability
[23]	Barbieri et al. (2024)	UCI + real-world datasets	FS performance and stability analysis	Omits deep learning and RL methods
[24]	Lamsaf et al. (2025)	Mixed domains	Causal ML FS survey	Broad but lacks detailed benchmarks

3. METHODOLOGY

Figure 1 shows the complete methodological framework used in this study. The process begins with collecting and combining data. After that, it goes through full data preprocessing. Then, FS is done, followed by evaluation of the selected features. Finally, the most important feature set is chosen.

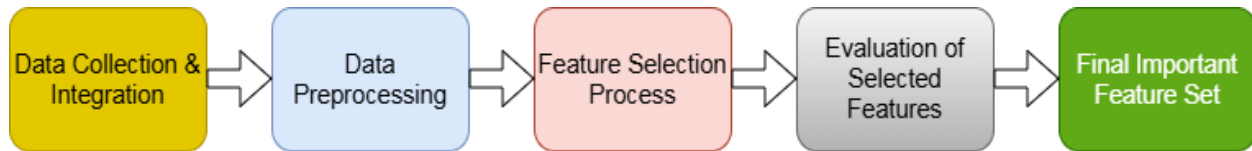


Figure 1. Research Framework

3.1 Data Collection and Integration

In this study, a private healthcare dataset was developed to identify individuals who may require medical attention due to potential health risks. The dataset was constructed by integrating two complementary parts collected from a private source.

- **Healthcare, Lifestyle, and Comorbidity Dataset:** This dataset had 19,536 records of participants and 67 features. It included basic demographic data (age, gender, race, marital status, household size, household income), history of illnesses and comorbidities (e.g., hypertension, diabetes, autoimmune or kidney disease, cancer, eye disease), vaccination details (COVID-19, HPV, child vaccines), and family history of chronic illness or cancer. Lifestyle and behaviour data were also collected, including smoking and e-cigarette use, alcohol use, physical activity (e.g., whether participant could reach 10,000 steps per week), sleep hours, diet habits, and healthcare use (medical_insurance, annual_checkups, medical_card_benefit, out-of-pocket_costs for GP and pharmacy). Other health factors included past sexually transmitted infections, allergies, injuries (especially eye injuries), snoring, menopause, medications, willingness to take mammograms, and eye check-ups. Height and weight were also measured.
- **Geo-demographic and Socio-environmental Dataset:** This dataset had 14,729 records and 472 features. It captured population and environmental indicators for the same regions. Features included population structure (total population, male and female numbers, women of reproductive age, elderly, youth, children) and many categorical attributes describing the local area and facilities. Examples include housing types (residential areas, serviced apartments, resorts, homestays), commercial places (markets, hypermarkets, shops), food outlets (restaurants, cafes, fast food, traditional food shops), education places (schools, colleges, libraries, tuition centres), healthcare services (clinics, pharmacies, care centres), entertainment venues, financial services, transport hubs, and public facilities (parks, worship places, community halls).

By merging the two datasets using participants' unique IDs, we obtained a final dataset of 14,713 participant records with 582 features (after column splitting of multi-value fields such as disease lists and one-hot encoding of categorical attributes). It showed both personal health and lifestyle details as well as the wider social and environmental context. The dataset was fully anonymized before analysis to meet privacy rules. Because of the large number of features, strong FS methods were used to remove repeated or correlated features, make data easier to understand, and improve model results [25]. The target variable was a binary flag "need attention" (1 = at risk/needs medical follow-up, 0 = no attention needed), with all 582 features initially used as possible predictors.

3.2 Data Preprocessing

A structured and systematic preprocessing workflow was implemented to improve data quality and ensure analytical reliability [26]. First, all feature names and text answers that were first in Malay were changed to English to make the dataset the same in language. Missing data were handled with special imputation ways that fit each feature's type and spread. Categorical variables were changed by one-hot encoding to make them work with ML methods. Features that were not useful, repeated, or had low value, like those with too much missing data or little change, were removed in a planned way. Also, complex or mixed fields and not relevant category variables were left out to reduce noise and avoid possible bias. Finally, all features were put on the same scale to make sure they were the same size, which is

very important for methods that are sensitive to big or small feature values. This careful preprocessing work gave a strong base for a clear and strong FS step.

3.3 Feature Selection Process

The main goal of FS is to choose the most useful variables. These variables help the model make good predictions. This makes the model work better and easier to understand. It also lowers the chance of overfitting. In this study, FS methods were divided into five main groups: filter methods, wrapper methods, embedded methods, model-agnostic methods, and methods for reducing dimensions. Different algorithms from each group were used and compared to find the best set of features.

3.3.1 Filter Methods

Filter methods look at each feature alone to check if it is useful for the target variable. They do not use any prediction model [27]. Their main advantages are fast speed and simple steps, making them ideal for first feature screening, especially in large healthcare datasets that have hundreds of variables. Filter methods give an early ranking of features. This allows researchers to remove features that are not useful or repeated before using more complex methods that need more computer power.

- **Mutual Information (MI):** MI checks both straight-line and non-straight-line links between each feature and the target. This is very useful in healthcare datasets, where the link between health signs and diseases is often not simple. For example, the effect of blood sugar on diabetes risk may change together with age or BMI. MI can catch these complex links, so important predictors are not missed.
- **Chi-squared Test (χ^2):** This test checks the link between category features and a category target. It is often used in health studies with surveys or population data. For example, χ^2 can show if lifestyle factors, such as smoking or diet groups, are clearly linked to heart disease.
- **Analysis of Variance (ANOVA):** ANOVA checks if number features are very different across target groups. In healthcare, it helps to see if body signs or lab results change between patient groups. For example, ANOVA can find which blood markers are not the same in people with and without high blood pressure, guiding which features are more important.
- **Correlation-Based Feature Selection:** This method looks at pair links between features to find repeats. Very similar features can confuse the model and make results unstable. By removing repeated features, correlation-based selection makes the model easier to explain and steadier. This is very important in health, where we must know clearly how each predictor works.

3.3.2 Wrapper Methods

Wrapper methods check groups of features by testing how they change the result of one prediction model [28]. Unlike filter methods, they also look at how features work together, giving a more correct choice but needing more computer use. Wrappers are very useful when the dataset has health variables that affect each other, since they can find effects that single filters may not see.

- **Recursive Feature Elimination (RFE):** RFE step by step removes the least useful features based on scores from a model, keeping only those that help most for accuracy. In healthcare, RFE can pick the most important biomarkers and remove noisy or repeated ones. Its step process keeps the right balance between simple model and good results.
- **Boruta:** Boruta uses Random Forest (RF) to compare real features with random “shadow” ones. Features that are clearly better than shadows are kept. Boruta works well for hard healthcare datasets with both number and category features, making sure only useful predictors are chosen.

- Sequential Feature Selection (SFS): These methods add (forward) or remove (backward) features step by step by checking model result. They can find features that work well together to improve prediction, which is very important in healthcare where many factors together affect patient outcome.

3.3.3 Embedded Methods

Embedded methods do FS while the model is being trained, using the model's own rules [29]. They mix speed and correctness by giving weight to features based on how much they help the model's result. Embedded methods lower the chance of overfitting, which is very important in health datasets where the number of samples is not always big.

- LASSO (L1 Regularization): LASSO makes small the values of less useful features until they are zero, so it also does FS. This is very helpful in healthcare datasets with many related variables, since only the strongest ones remain in the model.
- Ridge Regression (L2 Regularization): Ridge makes the values of features smaller but does not remove them. It reduces problems from high relation between variables while keeping their information. In clinical studies with related biomarkers, Ridge lets models use all useful features but controls unstable results.
- Tree-Based Feature Importance: Models like RF and Gradient Boosting Machines can give feature importance from the way they split. They can find non-straight-line links and complex relations, making them very good for health datasets with mixed data types like lab tests, age, and lifestyle.

3.3.4 Model-Agnostic Methods

Model-agnostic methods do not depend on the prediction model, so they can explain feature importance for any model [30]. These methods improve understanding, which is needed in healthcare where results must be explained to doctors and rule makers.

- Permutation Feature Importance: This method checks how much the model result becomes worse when a feature is shuffled. Features that make the biggest drop are the most important. It gives a simple and clear way to see feature use, letting researchers know the role of each variable without using model-specific values or trees.

3.3.5 Dimensionality Reduction Techniques

Dimensionality reduction changes the original feature space into fewer parts but keeps the main information. It is not direct FS, but it helps to reduce noise, make models simple, and show high-dimensional healthcare data.

- Principal Component Analysis (PCA): PCA makes new parts that are not related and keep the most variance. It is useful to reduce size in datasets with many related features like lab or gene data. But the new parts are mixes of old features, so they are harder to explain.
- Linear Discriminant Analysis (LDA): LDA is a supervised way that makes the groups more separate. It gives new parts that help for classification and show which features matter most for telling patient groups apart.

3.4 Evaluation of Selected Features

After FS, each chosen group was tested for stability, clarity, and value for prediction. Beyond normal measures like accuracy and AUC, this test also checked if feature importance was similar across many methods, giving trust in the strength and wide use of the chosen features. Stable features, often found as important, show strong value in health context, and may guide future study or treatment.

3.5 Final Important Feature Set

By combining results from filter, wrapper, embedded, model-agnostic, and dimensionality reduction methods, a final set of important features was made. This set keeps the balance between accuracy, clarity, and strong method base. The chosen features give a good base for future prediction, with both predictive value and clear interpretability that can support health decisions.

In this study, we used a fixed, stratified 80/20 train-test split with a random seed of 42 for all experiments. To keep the comparison fair and ensure performance changes resulted solely from the feature selection methods rather than model adjustments, all classifiers ran on default settings without separate validation sets. We skipped hyperparameter tuning and k-fold cross-validation entirely. Because our main objective was an equal evaluation of feature selection techniques under identical conditions, avoiding these extra validation steps kept the testing framework consistent across all models.

4. RESULTS AND DISCUSSIONS

In this study, many FS algorithms were used to find and rank the most important features from the dataset. The main goal was to know which features are always chosen as important and how different algorithms rank them. The FS methods in this study include MI (Figure 2), Chi-Squared Test (Figure 3), ANOVA (Figure 4), and Correlation-Based Selection (Figure 5). These belong to the filter group. For wrapper methods, we used RFE (Figure 6). We also used Boruta (Figure 7), SFS (Figure 8), and SBS (Figure 9). The embedded methods included LASSO (L1 Regularization) (Figure 10), Ridge Regression (L2 Regularization) (Figure 11), RF (Figure 12), and Gradient Boosting Machine (Figure 13). For reducing the number of dimensions, we used PCA (Figure 14) and LDA (Figure 15). Finally, the model-agnostic method of Permutation Feature Importance (Figure 16) was also checked. In all figures mentioned, higher scores indicate stronger non-linear relationships, while lower scores indicate weaker ones

4.1 Performance Comparison of Feature Selection Methods

This part shows the results of different FS methods used on the integrated healthcare dataset. The methods were grouped into filter, wrapper, and embedded types to find the most useful and easy-to-understand features for the health prediction task. By checking and ranking features in different ways, we wanted to make sure the results are strong and get better ideas about the links between features and the target variable.

Figure 2 displays the Mutual Information (MI) scores, evaluating the statistical dependence between each feature and the target variable. Capable of capturing both linear and non-linear relationships, higher MI scores reflect greater predictive relevance by quantifying the reduction in uncertainty about the target given a feature. Ranked by MI scores, the top 20 features highlight `test_blood` as the strongest predictor (≈ 0.02), followed by `age_18_to_30` (≈ 0.013) and `households_1_to_3` (≈ 0.012). While `test_blood` offers the highest individual predictive power, the remaining features spanning various demographic and lifestyle factors, provide complementary predictive value.

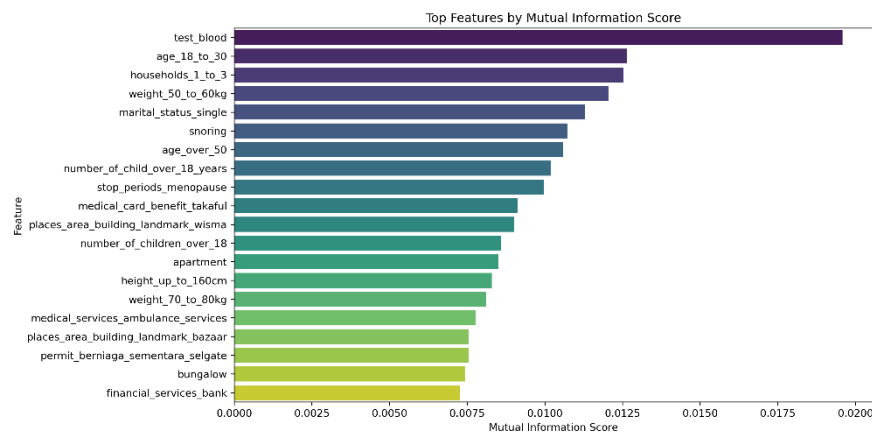


Figure 2. Top 20 Features Ranked by their MI Scores with the Target Variable

Figure 3 illustrates the Chi-Squared (χ^2) test results, which assess statistical independence between categorical features and the target variable. Higher χ^2 values indicate stronger statistical associations. Ranked by their χ^2 scores, the top 20 features highlight age_18_to_30 as the single most significant predictor ($\chi^2 > 300$), reflecting a robust link with the target outcome. Other top-ranked features include age_over_50, marital_status_single, number_of_child_over_18_years, and snoring. This distinct hierarchy underscores the pivotal role of demographic and behavioral traits in predicting the outcome and confirms their utility for model training.

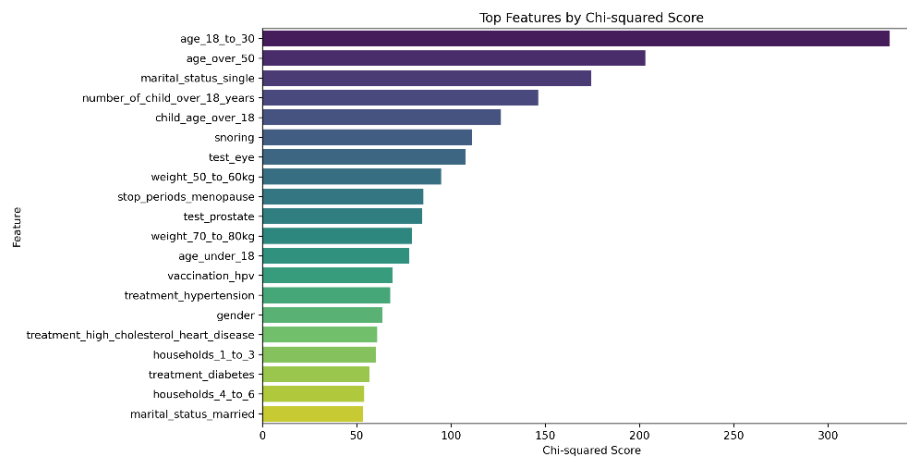


Figure 3. Top 20 Features Ranked by their Chi-Squared Scores with the Target Variable

Figure 4 illustrates the ANOVA F-test results, which evaluate whether continuous feature means differ significantly across target classes, with higher F-scores indicating stronger discriminative power. Displaying the top 20 features ranked by their F-statistics, the plot highlights test_blood as the single most influential feature ($F \approx 800$), more than double the score of the second-ranked predictor, age_18_to_30 ($F \approx 380$). Other notable features, including test_urine, age_over_50, and marital_status_single, also exhibit high F-scores, confirming their ability to distinguish between target classes. These findings underscore the dominant predictive power of test_blood while emphasizing the complementary importance of age and lifestyle indicators.

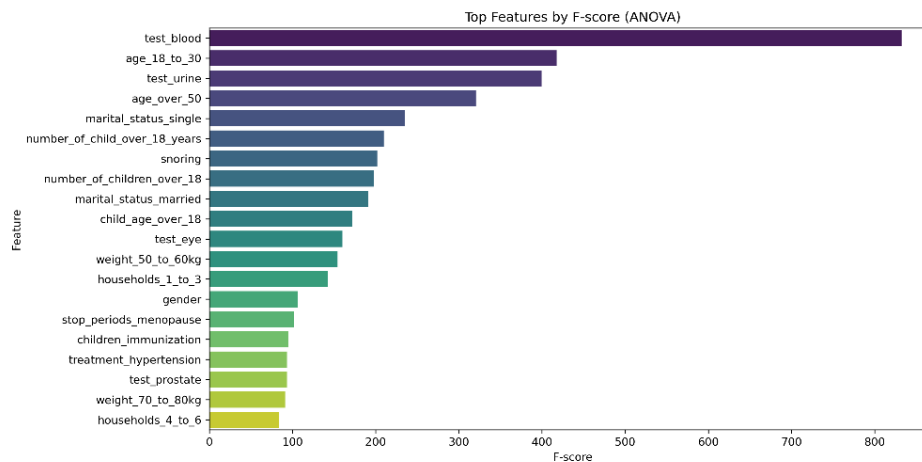


Figure 4. Top 20 Features Ranked by ANOVA F-scores

Figure 5 illustrates the Pearson correlation analysis results, evaluating linear relationships between numerical features and the target variable after pruning highly collinear pairs ($r > 0.85$) to mitigate multicollinearity. Displaying the top 20 remaining features ranked by their absolute correlation coefficients, the plot highlights test_blood as having the strongest linear relationship ($r \approx 0.21$), followed by age_18_to_30, test_urine, and age_over_50. This alignment with the trends observed across other filter methods underscores the consistent importance of test_blood and reinforces a core set of key demographic and clinical predictor.

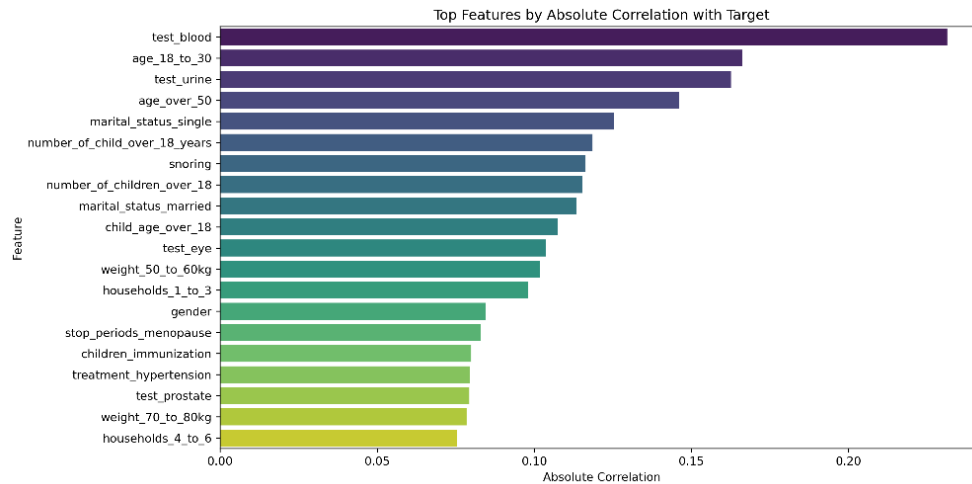


Figure 5. Top 20 Features Ranked by Their Relationship Strength with the Target Variable

Figure 6 illustrates Recursive Feature Elimination (RFE), a wrapper approach that iteratively removes the least important variables based on model coefficients until an optimal subset is retained. Displaying the top features ranked by their absolute logistic regression coefficients following RFE, the plot highlights `test_blood` as having the highest coefficient (≈ 0.5), affirming its dominant influence. Other prominent features include `age_over_50`, `age_18_to_30`, `snoring`, and `treatment_diabetes`. These findings underscore the robust predictive capacity of key demographic and health-related variables within an iterative elimination framework.

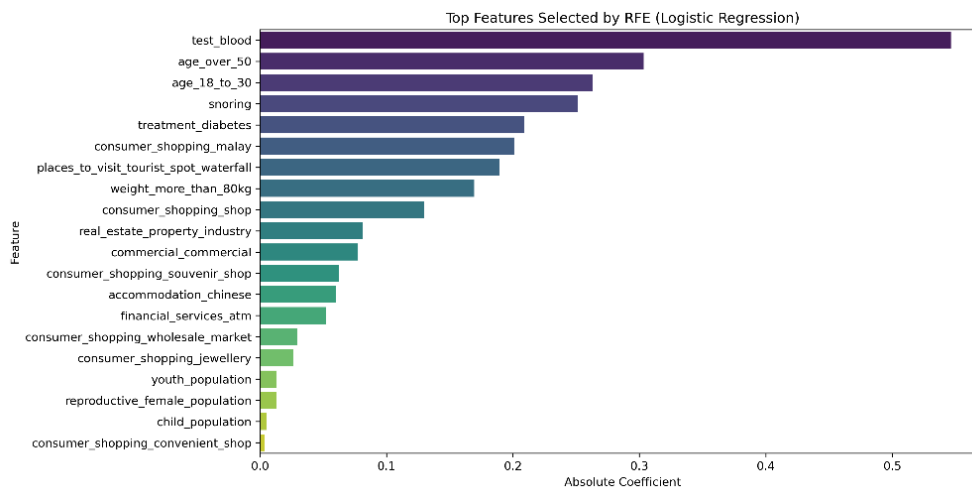


Figure 6. Top 20 Features Selected by RFE using Logistic Regression

Figure 7 illustrates the top 20 of the 21 features validated by Boruta, a robust Random Forest-based wrapper method that evaluates actual features against randomized shadow features to confirm statistical significance. Ranked by Random Forest importance scores, the leading predictors are `test_blood` (~ 0.31), `test_urine` (~ 0.13), `age_18_to_30` (~ 0.12), and `age_over_50` (~ 0.08). By successfully identifying both strongly and weakly relevant features, Boruta provides high confidence that the selected variables have a genuine impact on the target variable rather than reflecting random noise.

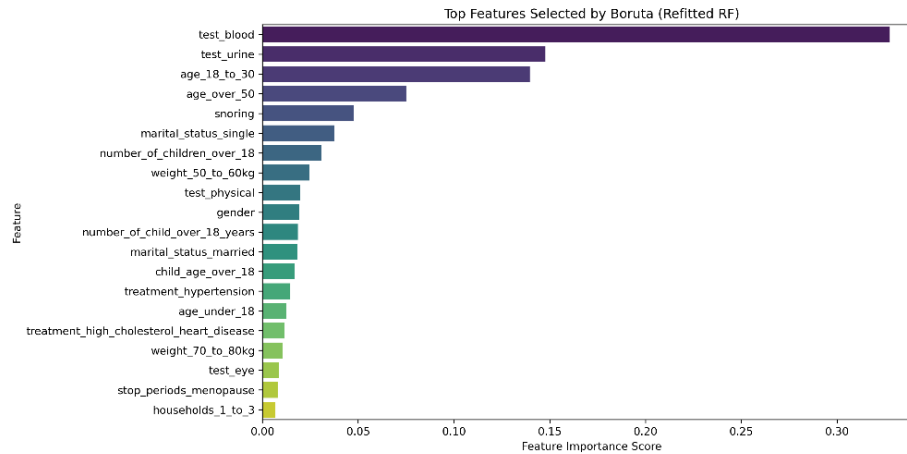


Figure 7. Top 20 Features Selected by the Boruta algorithm with a Refitted RF

Note: Boruta confirmed a total of 21 features; the figure displays the 20 highest-ranked ones for visual clarity.

Figure 8 illustrates the results of Sequential Forward Selection (SFS), a wrapper approach that iteratively constructs an optimal feature subset by adding the variable that provides the greatest marginal improvement at each step. Ranked by their absolute logistic regression coefficients, the top 20 features highlight test_blood as the primary predictor (≈ 2.0), followed by treatment_diabetes (≈ 1.0), age_18_to_30, and test_prostate. This distinct hierarchy underscores the dominant role of test_blood while confirming that metabolic and lifestyle indicators offer critical complementary value in predicting disease risk.

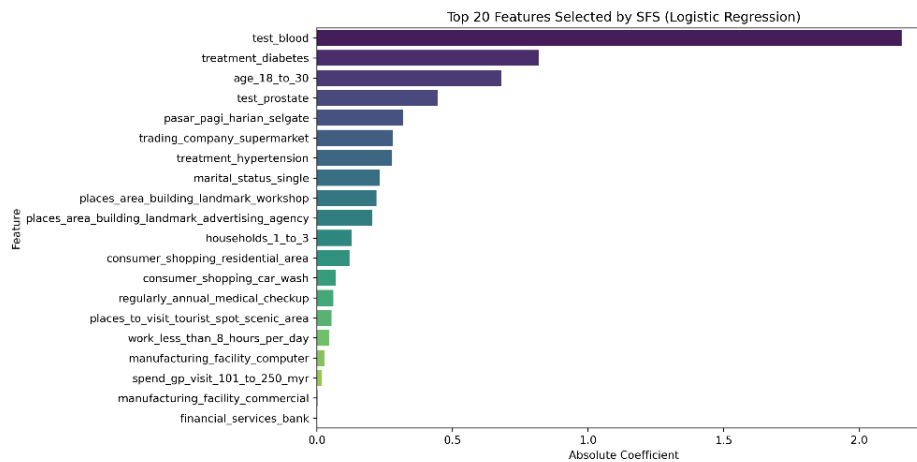


Figure 8. Top 20 Features Selected by SFS using Logistic Regression

Figure 9 illustrates Sequential Backward Selection (SBS), a wrapper approach that begins with the full feature set and iteratively eliminates the least impactful variable until the target subset size is reached. Ranked by their absolute logistic regression coefficients, the retained features highlight test_blood as the most critical predictor (≈ 0.4), followed closely by age_18_to_30 (≈ 0.3) and snoring (≈ 0.28). These results demonstrate which variables would cause the greatest performance loss if removed, reinforcing the essential role of test_blood alongside key demographic and lifestyle factors in health outcome prediction.

Figure 10 illustrates the LASSO regularization results, which integrate model fitting and feature selection by applying an L_1 penalty to shrink less relevant coefficients to zero, producing a sparse and interpretable model. Ranking the non-zero feature coefficients by their absolute magnitude, the plot identifies test_blood as the primary predictor (≈ 0.09), followed by age_18_to_30, test_urine, and snoring. This induced sparsity confirms test_blood's consistent importance across selection methods and isolates a refined subset of highly informative variables essential for both predictive accuracy and model interpretability.

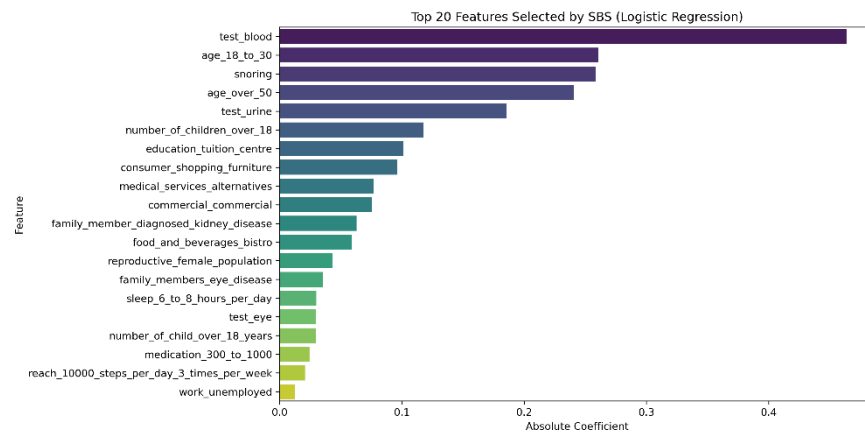


Figure 9. Top 20 Features Selected by SBS using Logistic Regression

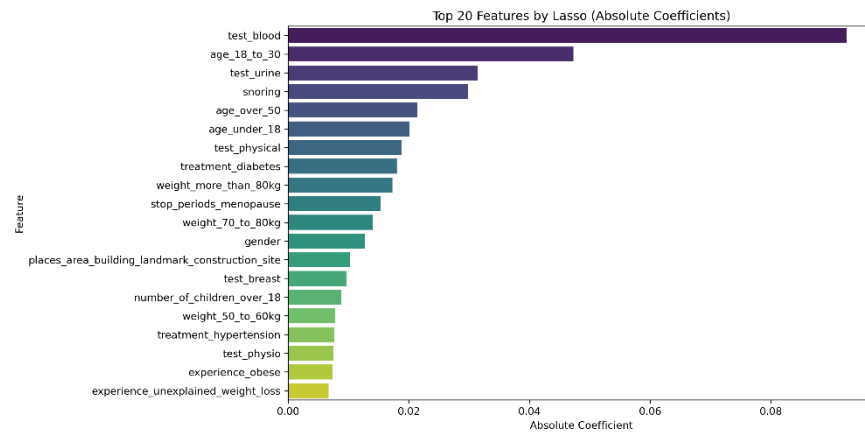


Figure 10. Top 20 Features Selected by Lasso Regularization, Ranked by Absolute Coefficient Values

Figure 11 illustrates the Ridge Regression results, which apply an L_2 penalty to shrink feature coefficients and mitigate multicollinearity without eliminating variables, preserving the complete feature space. Ranking the top 20 features by absolute coefficient magnitude, the plot reveals female_population as the leading predictor (≈ 0.3), followed by test_blood and several commercial activity variables (e.g., commercial_commercial and pasar_pagi_harian_selgate). These findings demonstrate that under L_2 regularization, demographic density and commercial factors hold substantial predictive power, complementing clinical variables and offering a broader, more nuanced view of the dataset's underlying structure.

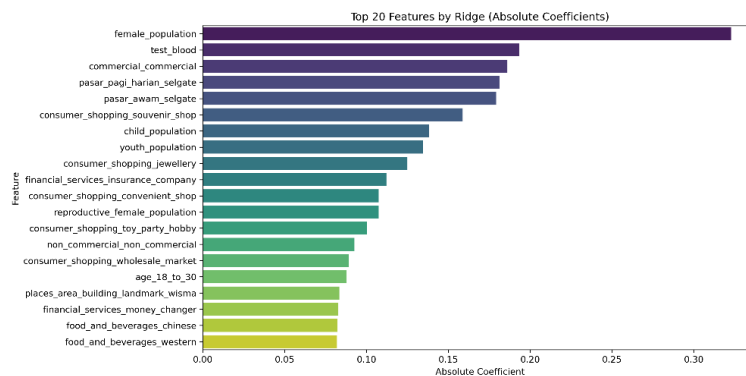


Figure 11. Top 20 Features Selected by Ridge Regression, Ranked by Absolute Coefficients

Figure 12 illustrates the Random Forest (RF) feature importance scores, which inherently quantify variable relevance based on mean decrease in impurity. Ranking the top 20 features, the plot highlights `test_blood` as the single most influential predictor (≈ 0.025), followed by `test_urine`, local commercial and spatial attributes such as `places_area_building_landmark_landmark` and `food_and_beverages_malay`. These results demonstrate the multifaceted nature of the target outcome, confirming `test_blood` as a primary clinical driver while uncovering the meaningful contribution of localized environmental context.

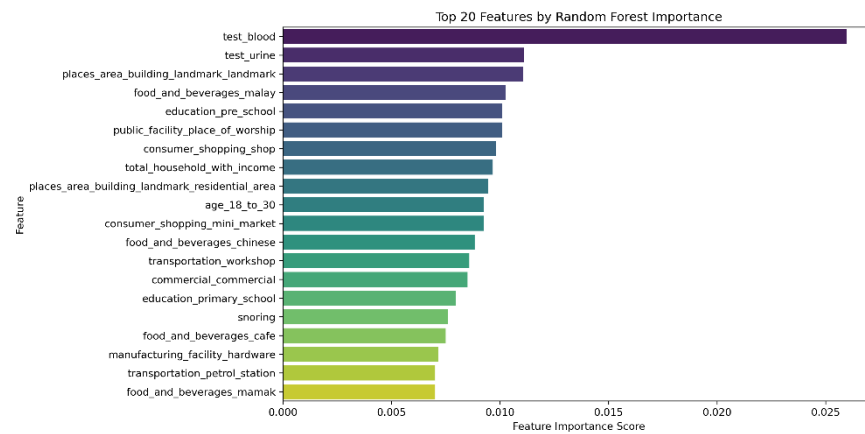


Figure 12. Top 20 Features Ranked by RF Importance Scores

Figure 13 illustrates the Gradient Boosting feature importance scores, which evaluate variable relevance by aggregating each feature's contribution to loss reduction across sequential trees. The plot highlights `test_blood` as having the highest importance score (≈ 0.26), more than double that of `age_18_to_30` (≈ 0.12), underscoring its dominant predictive role. Other top-ranked variables, including `test_urine` and `age_over_50`, closely mirror patterns observed across linear and ensemble methods. This consistent dominance across both boosting and bagging frameworks firmly establishes `test_blood` as a critical determinant of target predictions.

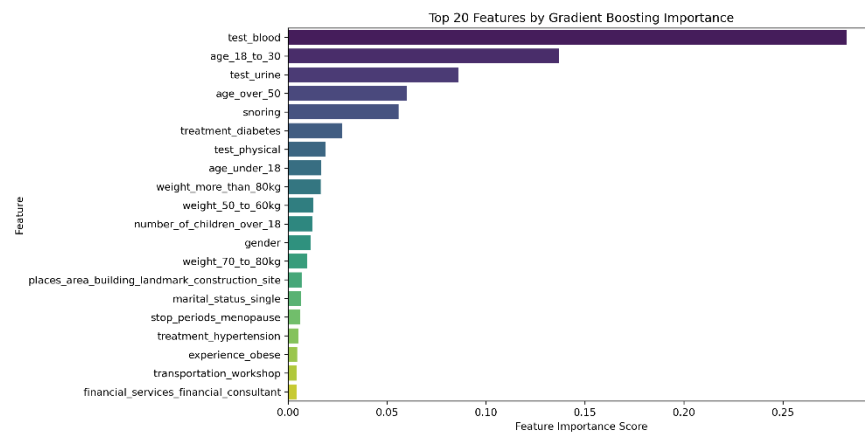


Figure 13. Top 20 Features Ranked by Gradient Boosting Importance

Figure 14 illustrates the Principal Component Analysis (PCA) results, which reduce dimensionality by projecting data onto orthogonal components that capture maximum variance. Ranking features by their absolute loadings on the first principal component (PC1), the plot identifies commercial and consumer behavior variables such as `consumer_shopping_pharmacy` and `consumer_shopping_shop` as the primary drivers of dataset variance. These findings indicate that commercial activities and consumer patterns account for the majority of overall data variability, offering complementary structural insight beyond pure predictive relevance to the target.

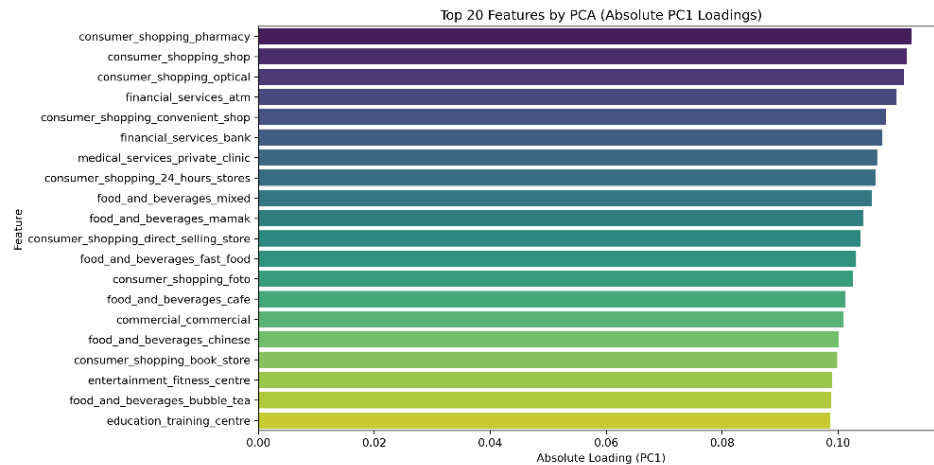


Figure 14. Top 20 Features Ranked by Absolute Loadings for the First Principal Component (PC1) from PCA

Figure 15 illustrates the Linear Discriminant Analysis (LDA) results, which construct linear combinations of features to maximize class separability. Ranking features by their absolute coefficients on the first linear discriminant, the plot identifies *female_population* (≈ 1.2) as the single most critical predictor, followed by key local commercial attributes including *pasar_pagi_harian_selgate* and *pasar_awam_selgate* and *test_blood*. These results highlight the primary importance of demographic density and targeted commercial factors in maximizing class separation, while once again confirming *test_blood*'s strong, consistent relevance across diverse feature selection paradigms.

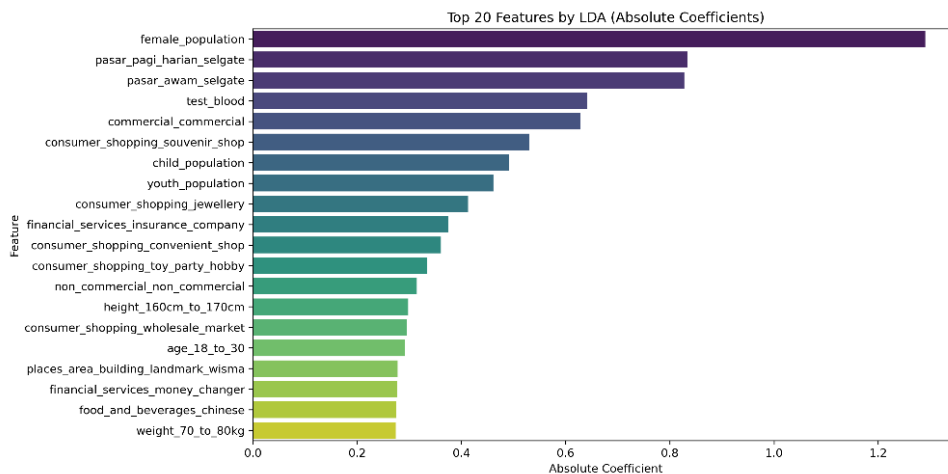


Figure 15. Top 20 Features Ranked by Absolute LDA Coefficients

Figure 16 illustrates the 20 most important features identified via permutation feature importance on the reduced dataset. The horizontal bars display mean importance scores, quantifying the decrease in model performance when a feature's values are randomly shuffled. Higher scores denote a greater impact on predictive capability, with *snoring*, *test_blood*, and *age_18_to_30* emerging as the three most influential variables in the model.

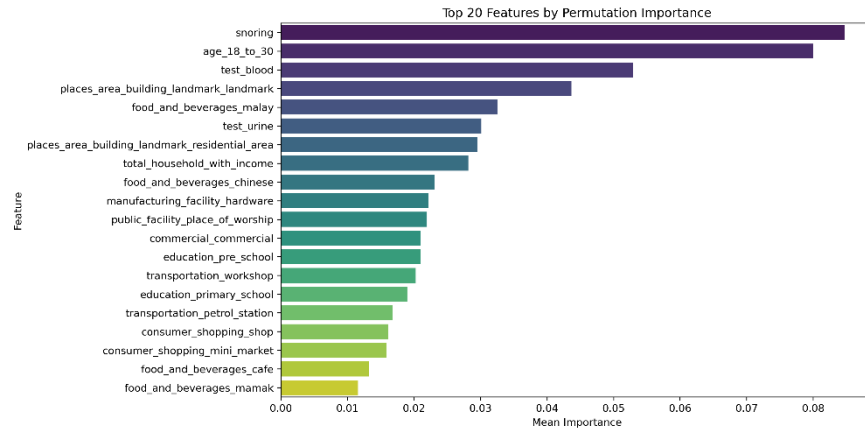


Figure 16. Top 20 Features Ranked by Permutation Importance

The top-ranked features using different FS methods are summarized in Table 2.

Table 2. Summary of FS Methods and Top-ranked Features

Method	Key Top-Ranked Features	Ref.
MI	test blood, age 18 to 30, households 1 to 3	Figure 2
Chi-Squared Test	age 18 to 30, age over 50, marital status single	Figure 3
ANOVA F-Test	test blood, age 18 to 30, test urine	Figure 4
Correlation Analysis	test blood, age 18 to 30, test urine	Figure 5
RFE	test blood, age over 50, snoring	Figure 6
Boruta	test blood, test urine, age 18 to 30	Figure 7
Sequential Forward Selection	test blood, treatment diabetes, age 18 to 30	Figure 8
SBS	test blood, age 18 to 30, snoring	Figure 9
LASSO (L ₁)	test blood, age 18 to 30, snoring	Figure 10
Ridge (L ₂)	female population, test blood, commercial commercial	Figure 11
RF	test blood, landmark-related, F&B malay	Figure 12
Gradient Boosting	test blood, age 18 to 30, test urine	Figure 13
PCA (PC1 loadings)	pharmacy, shop, optical, ATM	Figure 14
LDA	female population, pasar pagi, test blood	Figure 15
Permutation Feature Importance	snoring, test blood and age 18 to 30	Figure 16

The FS analysis used many methods. They were grouped into filter, wrapper, embedded, and dimensionality reduction ways. These were to find the most important predictors for the target. Filter methods like MI, Chi-Squared Test, ANOVA F-Test, and Correlation Analysis often picked features like test_blood, age_18_to_30, and test_urine as top parts. Wrapper methods included RFE, Boruta, Sequential Forward Selection, and SBS. These often chose test_blood, age_18_to_30, and extra features like snoring and treatment_diabetes. This showed they were important when using logistic regression and RF models. Embedded ways such as LASSO, Ridge, RF, and Gradient Boosting also showed test_blood and age_18_to_30 were important. They also picked other features like female_population and landmark-related parts. At last, dimensionality reduction ways like PCA and LDA showed special groups of features. They showed variables like pharmacy, shop, optical, ATM, female_population, and pasar_pagi. These may show bigger changes or group effects in the data. Overall, test_blood and age_18_to_30 emerged as consistently important across nearly all methods, underscoring their critical role in model performance and interpretability.

A comprehensive comparison of feature importance scores across all evaluated feature selection (FS) paradigms is summarized in the heatmap in Figure 17. This cross-method comparison provides clear consensus on which variables consistently demonstrate strong predictive power regardless of the selection technique. Following an extensive benchmark across filter, wrapper, embedded, and dimensionality-reduction approaches, Boruta was selected as the primary feature-selection framework for the final models. This decision rested on a downstream classifier comparison: a Random Forest classifier trained on the Boruta-selected subset of 21 features achieved competitive or superior

performance relative to identical Random Forest classifiers trained on the subsets produced by RFE, LASSO, PCA and other methods for the majority of evaluation metrics, while simultaneously delivering a compact, interpretable feature set.

4.1.1 BorutaPy Execution and Role in Feature Selection

The Boruta algorithm was executed over 100 iterations using a Random Forest classifier. Initially, all 582 features were categorized as tentative and benchmarked against randomized shadow features. During iterations 1–7, all features remained tentative while statistical significance was established. Iteration 8 marked a major convergence point, with 542 uninformative features rejected and 40 remaining tentative. From iteration 9 onward, features that consistently outperformed the maximum shadow-feature importance were progressively confirmed. The confirmed set stabilized at 21 features from iteration 9 through 100, while the tentative set shrank to 3 features, resulting in 556 total rejections.

This rigorous iterative process ensured that only statistically validated variables were retained. Unlike traditional minimal-optimal selection methods that isolate a small predictive subset, Boruta applies an all-relevant feature strategy, preserving weakly relevant yet potentially informative predictors. This property is valuable for high-dimensional healthcare data, where subtle feature interactions may contribute to model performance. Although computationally more intensive than pure filter methods, the trade-off is justified by the higher quality and interpretability of the resulting feature set.

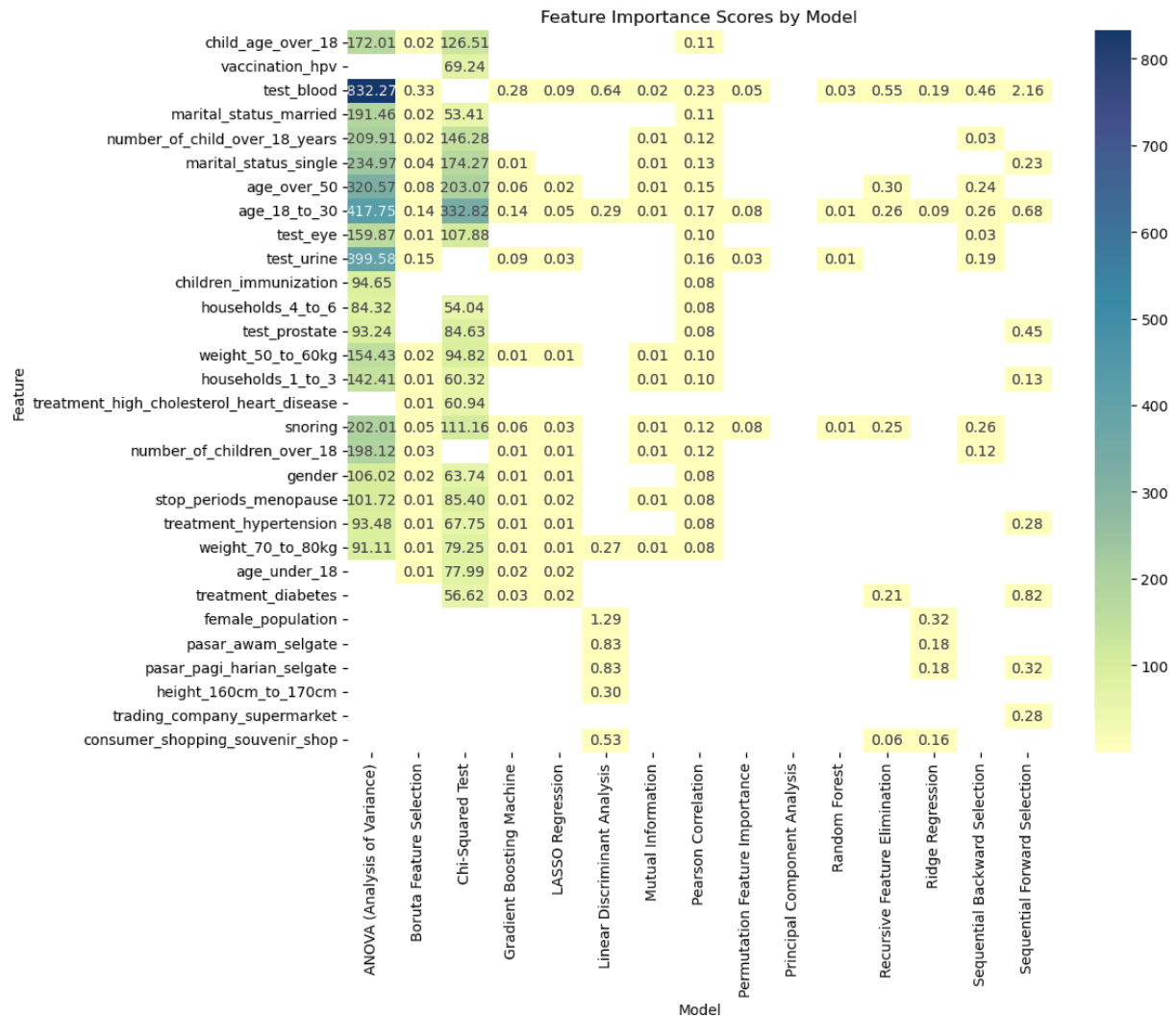


Figure 17. Heatmap Comparing Feature Importance Scores across Different Selection Methods

4.2 Interpretability

To ensure robust feature selection and model transparency, a multi-paradigm approach was adopted. Filter methods (Mutual Information, Chi-Squared, ANOVA, Correlation-Based Selection) identified algorithm-independent informative features. Wrapper methods (RFE, Boruta, Sequential Forward Selection, Sequential Backward Selection) evaluated feature subsets according to model performance. Embedded methods (LASSO, Ridge, Random Forest and Gradient Boosting importance) performed selection during model training. Dimensionality-reduction techniques (PCA, LDA) were also examined. Model-agnostic Permutation Feature Importance further quantified each feature's contribution across classifiers.

SHAP and LIME were not applied in the present study because the primary objective was comparative evaluation of feature-selection methods rather than post-hoc local or global explanation. Both techniques remain valuable for enhancing clinical interpretability and will be explored in future work. The combination of FS strategies employed here simultaneously identifies informative features and improves the transparency of the resulting models across different algorithms.

4.3 Robustness

Robustness of the FS methods was assessed by measuring the stability of selected features across multiple runs with different random seeds. Boruta and embedded techniques such as LASSO exhibited high stability, consistently returning nearly identical feature sets. Filter-based methods showed greater variability. Feature stability is critical in healthcare applications, where inconsistent selection can undermine prediction reliability, interpretability and clinical trust.

4.4 Performance Evaluation Metrics

The practical value of feature selection was quantified by training 12 machine-learning classifiers on two feature sets: (i) the full set of 582 features and (ii) the 21 features retained by Boruta. For clarity and space constraints in Table 3, the classifiers are abbreviated as follows: Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Gradient Boosting Machine (GBM), Gaussian Naïve Bayes (GNB), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Multilayer Perceptron Neural Network (MLP), Voting Classifier (Voting), Bagging Classifier (Bagging), and Stacking Classifier (Stacking). Specificity and the Matthews Correlation Coefficient (MCC) were also included. Results are reported in Table 3.

Table 3. Classifier Performance using All Features versus Boruta-selected Features (Before = all 582 features; After = 21 features selected by Boruta)

Classifiers	Accuracy		Precision		Recall		F1-score		Specificity		MCC	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
LR	0.783	0.792	0.798	0.799	0.961	0.975	0.872	0.878	0.183	0.175	0.233	0.266
SVM	0.771	0.771	0.771	0.771	1.000	1.000	0.871	0.871	0.000	0.000	0.000	0.000
DT	0.709	0.691	0.821	0.812	0.796	0.779	0.808	0.795	0.417	0.393	0.205	0.165
RF	0.781	0.748	0.789	0.816	0.978	0.869	0.873	0.842	0.117	0.341	0.197	0.230
GBM	0.801	0.797	0.802	0.803	0.984	0.976	0.884	0.881	0.183	0.196	0.310	0.296
GNB	0.540	0.729	0.823	0.846	0.514	0.792	0.633	0.818	0.629	0.513	0.120	0.287
KNN	0.724	0.744	0.775	0.792	0.905	0.905	0.835	0.845	0.116	0.200	0.029	0.136
LDA	0.784	0.790	0.811	0.802	0.939	0.966	0.870	0.876	0.261	0.197	0.273	0.265
MLP	0.781	0.785	0.792	0.803	0.970	0.956	0.872	0.873	0.144	0.208	0.208	0.252
Voting	0.697	0.763	0.826	0.829	0.770	0.873	0.797	0.850	0.453	0.392	0.208	0.285
Bagging	0.744	0.739	0.816	0.818	0.862	0.850	0.838	0.834	0.346	0.365	0.225	0.226
Stacking	0.788	0.793	0.798	0.801	0.972	0.974	0.876	0.879	0.171	0.184	0.250	0.275

When all 582 features were used, models typically achieved high recall at the expense of lower precision, specificity and overall balance, consistent with the presence of redundant or noisy variables. After Boruta selection, performance

changes were mixed as shown in Table 3. Several classifiers (Logistic Regression, Gaussian Naïve Bayes, K-Nearest Neighbors, LDA, MLP, Voting Classifier and Stacking Classifier) showed gains in Accuracy, F1-score or MCC. In contrast, Decision Tree, Random Forest, Gradient Boosting and Bagging Classifier exhibited modest declines on one or more metrics. Gradient Boosting, for example, recorded a slight reduction in F1-score ($0.884 \rightarrow 0.881$) and MCC ($0.310 \rightarrow 0.296$). Specificity improved for a number of models. SVM performance remained identical across all metrics.

Overall, the reduced feature set produced simpler, more stable models with competitive predictive performance while substantially mitigating the risk of overfitting and the curse of dimensionality—outcomes of particular practical importance in healthcare settings. These results demonstrate that an effective feature-selection process not only supports competitive prediction accuracy but also yields models that are simpler, more stable, and easier to deploy in clinical environments.

4.5 Computational Cost and Practical Trade-offs

To provide practical guidance, we quantified the computational cost of each feature-selection (FS) method under a controlled experimental protocol. All experiments were executed on a single HP EliteBook 840 G3 laptop equipped with an Intel® Core™ i5-6300U CPU @ 2.40 GHz (boost up to 2.50 GHz), 8.00 GB RAM (7.82 GB usable), Intel® HD Graphics 520, and a 704 GB storage drive, running Microsoft Windows 10 (64-bit). No GPU acceleration was used. The software environment consisted of Python 3.10.20 with the following library versions: numpy 1.26.4, scipy 1.13.1, pandas 2.2.2, matplotlib 3.10.9, scikit-learn 1.7.1, and Boruta 0.3. Timings were measured with Python's high-resolution time.perf_counter() and reported as wall-clock (elapsed) time for a complete run that selected the top 20 features. Each method was executed once under identical conditions after a warm-up pass; a fixed random seed of 42 was used wherever stochasticity exists, ensuring deterministic results. Default library settings were retained unless otherwise noted, and no additional parallelization beyond the libraries' internal defaults was applied.

Filter methods completed in well under two minutes: Mutual Information required approximately 1 min 39 s, Chi-Squared and ANOVA F-tests each finished in ~1 s, and Correlation Analysis took ~26 s. In contrast, wrapper methods were substantially more expensive because of their iterative model evaluations. Sequential Forward Selection and Sequential Backward Selection required ~50 min and ~55 min, respectively; Recursive Feature Elimination needed ~45 min. Boruta (Random-Forest-based) finished in ~3 min 24 s. Embedded methods occupied an intermediate range: LASSO (~19 s), Ridge (~3 s), Random-Forest feature importance (~18 s), and Gradient-Boosting feature importance (~30 s). Model-agnostic Permutation Feature Importance took ~20 min, while the dimensionality-reduction baselines PCA and LDA completed in ~1–2 s. Table 4 summarizes these wall-clock times together with brief methodological notes.

These timings highlight a clear accuracy–efficiency trade-off. Filter methods are well suited to large-scale screening or time-constrained pipelines, whereas wrapper and certain model-agnostic approaches, although frequently yielding more stable and predictive feature subsets, impose computational burdens that may be prohibitive under strict resource or latency constraints. Practitioners can therefore first apply fast filters for dimensionality reduction and subsequently refine the candidate set with a more expensive wrapper or embedded method when computational budget permits.

Taken together, the performance and computational results demonstrate that our study systematically compares filter, wrapper, embedded, and model-agnostic FS techniques on a large healthcare dataset, identifying clinically meaningful predictors of individuals requiring medical attention. Evaluation with multiple metrics (including specificity and Matthews correlation coefficient) yields healthcare-specific insights that directly support clinical decision-making— an aspect not addressed in prior work [10], [12]. Overall, some algorithms produce compact, highly focused feature sets while others retain larger collections that capture additional information. These differences underscore the necessity of aligning the chosen FS strategy with the study objective: interpretability, maximal predictive performance, or exhaustive retention of relevant variables.

Table 4. Computational Cost of Feature Selection Methods

Feature Selection Method	Computational Cost	Notes
MI (Filter)	~1 min 39 s	Quick statistical test for feature-target dependency
Chi-Squared Test (Filter)	~1 s	Simple test for categorical features
ANOVA F-Test (Filter)	~1 s	Fast test for numeric data
Correlation Analysis (Filter)	~26 s	Pearson/Spearman correlation for linear/non-linear relationships
RFE (Wrapper)	~45 min	Depends on estimator; iterative feature removal
Boruta (Wrapper)	~3 min 24 s	RF-based; moderate computational cost
Sequential Forward Selection (Wrapper)	~50 min	Iteratively adds features to improve model performance
SBS (Wrapper)	~55 min	Iteratively removes features based on model performance
LASSO (L1, Embedded)	~19 s	Linear model with L1 regularization for FS
Ridge (L2, Embedded)	~3 s	Linear model with L2 regularization; reduces overfitting
Random Forest Feature Importance (Embedded)	~18 s	Depends on number of trees; measures feature contribution
Gradient Boosting Feature Importance (Embedded)	~30 s	Depends on number of estimators/trees; model-based importance
PCA (Dimensionality Reduction)	~1 s	Eigen decomposition to reduce feature dimensions
LDA	~2 s	Supervised dimensionality reduction
Permutation Feature Importance	~20 min	Model-agnostic; measures impact of shuffling each feature

5. CONCLUSION

This study shows the key role of FS in building good, strong, and easy-to-understand ML models, especially in complex fields like healthcare. By step-by-step comparing many kinds of FS ways, such as filter, wrapper, embedded, model-agnostic, and dimensionality reduction methods, we showed that careful and case-based FS can improve model results, cuts overfitting, and makes models easier to explain.

Wrapper methods such as Boruta often produce competitive predictive performance while yielding compact and interpretable feature sets. We selected Boruta because it systematically validates features against shadow features, capturing both strong and weakly relevant predictors and providing high confidence in the final subset of 21 features. Although faster methods exist, Boruta's thorough validation process makes it a strong choice for identifying key features in high-dimensional healthcare data.

Even with these gains, problems still exist in picking the best feature groups for data with many dimensions and much noise. Future work should look at making mixed FS plans that combine good points in different ways and auto pipelines that can choose features in real time. These works will help push the use of ML in healthcare and other important areas, and lead to more trusted, clear, and general models. The mix of strong FS and expert knowledge will be very important to make sure ML is not only correct but also ethical and useful.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Md. Abu Nishan: Conceptualization, Data Curation, Methodology, Validation, Writing – Original Draft Preparation;
Choo-Yee Ting: Project Administration, Supervision, Writing – Review & Editing;
Chih-Yang Pee: Project Administration, Supervision, Writing – Review & Editing;
Helmi Zakariah: Project Administration, Supervision, Writing – Review & Editing;
Thanh-Trung Nguyen: Project Administration, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflicts of interest were disclosed.

ETHICS STATEMENTS

Our publication ethics follow the Committee of Publication Ethics (COPE) guidelines. <https://publicationethics.org/>

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

- [1] U. Stańczyk and G. Baron, “Does all or nothing always work best? In search of advantageous representation of attributes,” *Applied Sciences*, vol. 16, no. 6, p. 2679, Mar. 2026, doi: 10.3390/app16062679.
- [2] W. Cunha, A. Moreo, A. Esuli, F. Sebastiani, L. Rocha, and M. A. Gonçalves, “A Noise-Oriented and Redundancy-Aware instance selection framework,” *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–33, Nov. 2024, doi: 10.1145/3705000.
- [3] J. M. Kernbach and V. E. Staartjes, “Foundations of Machine Learning-Based Clinical Prediction Modeling: Part II—Generalization and Overfitting,” *Machine Learning in Clinical Neuroscience: Foundations and Applications*, pp. 15–21, Dec. 2021, doi: 10.1007/978-3-030-85292-4_3.
- [4] F. Xiang, C. Xing, Y. Ai, and Z. Wang, “An adaptive-weighted multi-level feature evidence and redundancy penalty method,” *Engineering Applications of Artificial Intelligence*, vol. 180, p. 115245, May 2026, doi: 10.1016/j.engappai.2026.115245.
- [5] Li, X., Xiong, H., Li, X., “Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond,” *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197–3234, Sep. 2022, doi: 10.1007/s10115-022-01756-8.
- [6] P. Dhal and C. Azad, “A comprehensive survey on feature selection in the various fields of machine learning,” *Applied Intelligence*, vol. 52, no. 4, pp. 4543–4581, Jul. 2021, doi: 10.1007/s10489-021-02550-9.

-
- [7] T. Miller, I. Durlík, A. Łobodzińska, L. Dorobczyński, and R. Jasionowski, “AI in Context: Harnessing Domain Knowledge for Smarter Machine Learning,” *Applied Sciences*, vol. 14, no. 24, p. 11612, Dec. 2024, doi: 10.3390/app142411612.
- [8] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, “Interpretable machine learning: Fundamental principles and 10 grand challenges,” *Statistics Surveys*, vol. 16, 1-85, Jan. 2022, doi: 10.1214/21-ss133.
- [9] C. Aliferis and G. Simon, “Overfitting, underfitting and general model Overconfidence and Under-Performance pitfalls and best practices in machine learning and AI,” in *Computers in health care*, 2024, pp. 477–524. doi: 10.1007/978-3-031-39355-6_10.
- [10] U. M. Khaire and R. Dhanalakshmi, “Stability of feature selection algorithm: A review,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 4, pp. 1060–1073, Jun. 2019, doi: 10.1016/j.jksuci.2019.06.012.
- [11] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, “Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods,” *Journal of Big Data*, vol. 11, no. 1, Mar. 2024, doi: 10.1186/s40537-024-00905-w.
- [12] D. Theng and K. K. Bhoyar, “Feature selection techniques for machine learning: a survey of more than two decades of research,” *Knowledge and Information Systems*, vol. 66, no. 3, pp. 1575–1637, Dec. 2023, doi: 10.1007/s10115-023-02010-5.
- [13] M. A. Bouke, A. Abdullah, J. Frnda, K. Cengiz, and B. Salah, “BukaGini: A Stability-Aware GiNI Index feature selection algorithm for robust model performance,” *IEEE Access*, vol. 11, pp. 59386–59396, Jan. 2023, doi: 10.1109/access.2023.3284975.
- [14] J. Wang, P. Xu, X. Ji, M. Li, and W. Lu, “Feature selection in machine learning for perovskite materials design and discovery,” *Materials*, vol. 16, no. 8, p. 3134, Apr. 2023, doi: 10.3390/ma16083134.
- [15] X. Zhang and H. Xiong, “Interpretable feature interaction via statistical self-supervised learning on tabular data,” *arXiv.org*, Mar. 23, 2025. <https://arxiv.org/abs/2503.18048>
- [16] A. Kodipalli, V. S. Devi, S. Guruvare, and T. Ismail, “Explainable AI-based feature importance analysis for ovarian cancer classification with ensemble methods,” *Frontiers in Public Health*, vol. 13, Mar. 2025, doi: 10.3389/fpubh.2025.1479095.
- [17] L. Lyu, N. Roy, H. Oosterhuis, and A. Anand, “Is interpretable machine learning effective at feature selection for Neural Learning-to-Rank?,” in *Lecture notes in computer science*, 2024, pp. 384–402. doi: 10.1007/978-3-031-56066-8_29.
- [18] R. Moraffah, P. Sheth, S. Vishnubhatla, and H. Liu, “Causal feature selection for responsible machine learning,” *arXiv.org*, Feb. 05, 2024. <https://arxiv.org/abs/2402.02696>
- [19] M. A. Jahin, M. F. Mridha, N. Dey, and Md. J. Hossen, “Human-in-the-Loop feature selection using interpretable Kolmogorov-Arnold Network-Based double Deep Q-Network,” *IEEE Open Journal of the Computer Society*, vol. 7, pp. 389–403, Jan. 2026, doi: 10.1109/ojcs.2026.3652986.
- [20] S. S. Nagaraju, “Automation and Feature Selection Enhancement with Reinforcement Learning (RL),” *arXiv.org*, Mar. 15, 2025. <https://arxiv.org/abs/2503.11991>
- [21] T. Zhao, J. Cabral, and G. Zhu, “A novel explainable artificial intelligence framework using knockoffs techniques with applications to sports analytics,” *Annals of Operations Research*, Apr. 2025, doi: 10.1007/s10479-025-06575-y.
- [22] T. Shaik, X. Tao, H. Xie, L. Li, N. Higgins, and J. D. Velásquez, “Towards transparent Deep Learning in Medicine: feature contribution and Attention Mechanism-Based Explainability,” *Human-Centric Intelligent Systems*, Jun. 2025, doi: 10.1007/s44230-025-00104-7.

- [23] M. C. Barbieri, B. I. Grisci, and M. Dorn, "Analysis and comparison of feature selection methods towards performance and stability," *Expert Systems with Applications*, vol. 249, p. 123667, Mar. 2024, doi: 10.1016/j.eswa.2024.123667.
- [24] A. Lamsaf, R. Carrilho, J. C. Neves, and H. Proença, "Causality, Machine Learning, and Feature Selection: a survey," *Sensors*, vol. 25, no. 8, p. 2373, Apr. 2025, doi: 10.3390/s25082373.
- [25] X. Cheng, "A comprehensive study of feature selection techniques in machine learning models," *Artificial Intelligence and Digital Technology*, vol. 1, no. 1, pp. 65–78, Nov. 2024, doi: 10.70088/xpf2b276.
- [26] A. N. Prasad, "Data quality and preprocessing," in *Apress eBooks*, 2024, pp. 109–223. doi: 10.1007/979-8-8688-1023-7_3.
- [27] A. Demircioğlu, "The effect of preprocessing filters on predictive performance in radiomics," *European Radiology Experimental*, vol. 6, no. 1, Sep. 2022, doi: 10.1186/s41747-022-00294-w.
- [28] J. Maldonado, M. C. Riff, and B. Neveu, "A review of recent approaches on wrapper feature selection for intrusion detection," *Expert Systems with Applications*, vol. 198, p. 116822, Mar. 2022, doi: 10.1016/j.eswa.2022.116822.
- [29] R. Jiao, B. H. Nguyen, B. Xue, and M. Zhang, "A survey on Evolutionary multiobjective feature selection in classification: Approaches, applications, and challenges," *IEEE Transactions on Evolutionary Computation*, vol. 28, no. 4, pp. 1156–1176, Jul. 2023, doi: 10.1109/tevc.2023.3292527.
- [30] Y. Wang, "A comparative analysis of model agnostic techniques for explainable artificial intelligence," *Research Reports on Computer Science*, Aug. 2024, doi: 10.37256/racs.3220244750.

BIOGRAPHIES OF AUTHORS

	Md. Abu Nishan is a data analyst, researcher, and IT consultant with over 10 years of experience, including expertise in banking MIS and data solutions. He is an Oracle Certified 11g Database Administrator (OCA, OCP- DBA). He obtained his bachelor's degree from the University of Calcutta, India. He holds an MSc in Computer Science & Engineering (MSCSE) from United International University (UIU), Bangladesh. He is pursuing a PhD in Information Technology (Data Science) at Multimedia University (MMU), Malaysia, focusing on healthcare risk prediction, especially Chronic Disease. He can be contacted at email: 1221404588@student.mmu.edu.my.
	Choo-Yee Ting is a Professor at the Faculty of Computing and Informatics and Deputy Director, Government Liaison & Industry Engagement, Multimedia University (MMU), Cyberjaya, Malaysia. He works on predictive analytics and Big Data and got his PhD (Information Technology), Specialization: Artificial Intelligence from Multimedia University in 2009. A Microsoft Research Fellow (2002), he has led national projects and won two Big Data prizes. He is also a certified trainer, consultant, and data scientist for Malaysia's COVID-19 vaccine programme. He can be contacted at email: cyting@mmu.edu.my.
	Chih-Yang Pee is an Assistant Professor at the Faculty of Computing and Informatics, Multimedia University (MMU), Cyberjaya, Malaysia. He works in statistical image analysis, pattern recognition, and moment invariants. He got his PhD in Mathematics from the University of Malaya in 2013. He has published journals like Pattern Recognition Letters and IEEE Access and leads a MOHE-funded project on discrete moment invariants. He also trains industry people in data science and Microsoft MCP. He can be contacted at email: cypee@mmu.edu.my.

	Helmi Zakariah is the Chief Executive Officer of Hayat Technologies Sdn Bhd. With a base in medicine and global health, his skill covers Big Data Analytics, Machine Learning, and global health policy. He has worked with government groups, WHO, UN, NGOs, and poor groups, which has grown his skill in advocacy, partnerships, and talks. Outside work, he likes to travel alone and read books on behavioral economics. He can be contacted at email: drhelmi@myhayat.my.
	Thanh-Trung Nguyen is a lecturer and Dean at the Faculty of Information Technology - Communications, Dong Nai Technology University, Dong Nai, Vietnam. He received his M.S. degrees (2008) and Ph.D. (2018) in Computer Science at University of Information Technology, Vietnam National University – HoChiMinh City, Vietnam. His research interests include Data Mining, Big Data, Blockchain, Face Recognition, Camera Cross-Tracking, IoT and AI-Robotics. He can be contacted at email: nguyenthanhtrung@dntu.edu.vn.