Journal of Azzaytuna University (JAZU)
JAZU Homepage: https://azzujournal.azu.edu.ly
ISSN (Print):2523-1006, ISSN (Online): 3105-4420 DOI: 10.35778/jazu.i56.a648

Comparative Analysis of Filter Methods for Gene Selection

Abdelhamid Elwaer*[1] and Abdeladeem Dreder[2]
[1]Faculty of Information Technology, University of Tripoli, Libya
[2]Faculty of Physical Therapy, University of Tripoli, Libya
ab.elwaer@uot.edu.ly, abd.dreder@uot.edu.ly
Corresponding Author: (*)
Publishing Date: 31 December 2025

ABSTRACT: Gene expression data presents significant challenges due to their high dimensionality; effective gene selection methods are needed to obtain accurate analysis and biomarker discovery. In this paper, we conducted a comprehensive comparative study using nine filter-based gene selection techniques: Information Gain, Mutual Information, Correlation-based Feature Selection (CFS), Relief-F, T-Test, Wilcoxon, Chi2, Pearson correlation, and Gini index. A breast cancer microarray dataset was used to evaluate these methods based on their classification accuracy, computational efficiency, and stability of the selected gene subsets. Most methods achieve high predictive accuracy and perfect stability but differ in their computational costs. This study aims to provide practical insights for choosing appropriate filtering methods based on their balance performance and efficiency in analyzing gene expression.
Keywords: gene expression, feature selection, T-Test filtering, Information Gain, Wilcoxon, Chi2, Pearson correlation, Gini index, breast cancer microarray

الملخص: تطرح بيانات التعبير الجيني تحديات جسيمة نظراً لأبعادها العالية؛ مما يستلزم وجود طرق فعالة لاختيار الجينات لضمان دقة التحليل واكتشاف المؤشرات الحيوية. أجرينا في هذه الورقة البحثية دراسة مقارنة شاملة باستخدام تسع تقنيات لاختيار الجينات تعتمد على أسلوب الترشيح (Filter–based)، وهي: كسب المعلومات (Information Gain)، والمعلومات المتبادلة (Mutual Information)، واختيار الميزات القائم على الارتباط (CFS)، وخوارزمية (Relief–F)، واختبار (T–Test)، واختبار ويلكوكسون (Wilcoxon)، واختبار كاى تربيع (Chi2)، وارتباط بيرسون (Pearson)، ومعامل جيني (Gini index). وقد استخدمت مجموعة بيانات المصفوفات الدقيقة (Microarray) لسرطان الثدي لتقييم هذه الطرق بناء على دقة التصنيف، والكفاءة الحسابية، واستقرار مجموعات الجينات المختارة. أظهرت النتائج أن معظم الطرق تحقق دقة تنبؤية عالية واستقرارا تاما، إلا أنها تفاوتت في التكلفة الحسابية. تهدف هذه الدراسة إلى تقديم رؤى عملية لاختيار طرق الترشيح المناسبة بناءً على توازن الأداء والكفاءة في تحليل التعبير الجيني.
الكلمات المفتاحية :التعبير الجيني، اختيار الميزات، طرق الترشيح، اختبارT ، كسب المعلومات، اختبار ويلكوكسون، (Chi2)، ارتباط بيرسون، معامل جيني، المصفوفات الدقيقة لسرطان الثدي.

I. Introduction

The Microarray and RNA sequencing technologies are used in biological research to enable simultaneous measurement of the expression levels of thousands of genes. This proliferation of gene expression offers opportunities for understanding complex biological processes and disease mechanisms, and it can be used for developing novel diagnostic and prognostic biomarkers. In gene datasets, the number of genes exceeds the number of samples. This curse of dimensionality

can lead to increased noise, model overfitting, reduced interpretability, and higher computational costs, which may hinder the discovery of truly relevant biological insights [1, 2].

Gene selection has emerged as a crucial pre-processing step in bioinformatics, which addresses the curse of dimensionality in gene datasets. The primary goal of gene selection is to identify a minimal subset of genes that are most relevant to a particular biological question. it can classify disease subtypes or predict patient outcomes. By reducing the dimensionality of the data, gene selection enhances the efficiency and accuracy of subsequent machine learning analyses, improves the interpretability of the results by focusing on a smaller set of key genes, and contributes to the development of more robust and generalizable predictive models [1].

Gene selection methodologies can be broadly categorized into three main types: filter, wrapper, and embedded methods. Filter methods are the focus of this study; they are distinguished by their independence from the learning algorithm. They use statistical measures (e.g., T-test, Chi2, Pearson correlation) or information-theoretic criteria (e.g., Information Gain, Mutual Information) [3] to evaluate the intrinsic characteristics of individual genes or gene subsets. This model-agnostic approach renders filter methods computationally efficient and scalable, making them particularly well-suited for analyzing large gene expression datasets, where speed and simplicity are important. Their use in dimensionality reduction and preliminary feature ranking remains widely acknowledged, despite their inability to account for gene-gene interactions,

This paper presents a comparative analysis of nine filter-based gene selection methods: Information Gain, Mutual Information, Correlation-based Feature Selection (CFS), Relief-F, T-Test, Wilcoxon, Chi2, Pearson correlation, and Gini index. Our objective was to evaluate their performance in a breast cancer microarray dataset. Different metrics were used, such as classification accuracy, computational efficiency, and stability of the selected gene subsets. This study also aims to offer practical insights into the strengths and weaknesses of each method using a detailed empirical comparison. This can guide researchers in selecting the most proper gene selection strategy for their specific application, and balancing predictive performance with computational demands. The subsequent sections detail the materials and methods employed, present the experimental results, discuss the findings in the context of the existing literature, and conclude with the implications of our study.

II. Literature Review

A critical step in the analysis of high-dimensional gene expression data is gene selection; It aims to identify a minimal set of relevant genes that can effectively discriminate between different biological states, it can discriminate between disease and healthy phenotypes, and between different cancer stages. This process is essential for reducing dimensionality, improving model accuracy, enhancing computational efficiency, and providing biological insights into the underlying disease mechanisms [1]. Among the various gene selection approaches, filter methods

stand out because of their computational simplicity, independence from the chosen classification algorithm, and high interpretability.

Filter methods assess the relevance of individual genes or gene subsets based on the intrinsic properties of the data using statistical measures or heuristic algorithms. These methods rank or score genes based on their correlation with the target variable, variance, or ability to separate classes without involving a specific learning model. Then, a predefined number of top-ranked genes are selected for downstream analyses [2]. This model-agnostic nature allows filter methods to be computationally efficient, making them particularly suitable for high-throughput gene expression datasets that often contain tens of thousands of genes.

Several statistical and information-theoretic measures form the basis of common filtering methods:

- **Information Gain (IG)**: Information Gain measures the reduction in entropy or uncertainty about the class variable when a gene's value is known [3]. Genes with higher information gains are considered more discriminative. Because of its ability to capture nonlinear relationships, it has been widely applied in bioinformatics for feature selection.
- **Mutual Information (MI)**: Mutual Information quantifies the statistical dependency between two random variables, which are a gene's expression level and the class label [4]. Genes with higher MI values indicate a stronger statistical relationship with the disease state.
- **Chi-squared (Chi2)**: The dependency between a categorical gene feature and a categorical class label [5] is measured using the Chi-squared statistic. It assesses whether the observed frequencies of gene expression values across different classes deviate significantly from the expected frequencies, assuming independence. A larger Chi2 value shows greater dependency and thus higher relevance
- **T-test**: For gene expression data and binary class labels, the T-test uses the p-value to identify genes whose mean expression levels differ significantly between the two classes [6]. It considers genes with lower p-values more differentially expressed and thus more relevant to the study.
- **Wilcoxon Rank-Sum Test**: The Wilcoxon Rank-Sum test is A non-parametric alternative to the T-Test. It assesses whether two independent samples (e.g., gene expression in disease vs. healthy) come from the same distribution [7], it does not assume normality of the data , and it is robust to outliers. This makes it suitable for various gene expression distributions.
- **Pearson's Correlation Coefficient**: This method can be used to measure the linear relationship between a continuous gene expression value and class label [8]. Genes with high absolute Pearson correlation coefficients were selected as being highly associated with the outcome.
- **Gini Index**: The Gini index measures the impurity of a dataset. In gene selection, it can be adapted to evaluate how well a gene splits data into homogeneous classes, where a lower Gini index for a gene indicates better class separation [9].
- **Relief-F**: Relief-F is an instance-based filter method that assigns weights to genes based on their ability to distinguish between nearest neighbors from different classes [10]. It can identify relevant genes in the presence of strong dependencies among them.

- **Correlation-based Feature Selection (CFS)**: CFS is a heuristic filter algorithm that evaluates the importance of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them [11]. The core idea is to select subsets of genes that are highly correlated with the class but poorly correlated with one another.

While filter methods offer advantages in speed and scalability, their main limitation lies in their inability to capture interactions between genes, as they typically evaluate genes independently or in simple pairwise relationships [2]. Subsequent studies often combine filter methods with wrapper or embedded approaches to leverage their strengths while mitigating their weaknesses [12].

III. Materials and Methods

This section outlines the experimental design and methodology used to compare various filter methods for gene selection in breast cancer diagnosis. The process includes data preprocessing, gene ranking, and selection using nine filter methods. The selected gene subsets are used for subsequent classification evaluation.

Dataset

This study utilized a publicly available breast cancer gene expression dataset (BC-TCGA) from The Cancer Genome Atlas (TCGA), a widely recognized repository for comprehensive genomic data [12]. BC-TCGA consists of 17,814 genes and 590 samples, including 61 normal samples and 529 breast cancer samples. A 70:30 stratified split into training and testing sets to preserve class balance and ensure unbiased evaluation.

Filter Methods

A comprehensive set of nine widely-used filter methods, denoted as F, were selected for comparison: F={Information Gain (IG), Mutual Information (MI), Correlation-based Feature Selection (CFS), Relief-F, T-Test, Wilcoxon, Chi2, Pearson, Gini}

For each filter method $f \in F$, the following procedure was applied.

1. **Gene Scoring**: A score was computed for each gene in the dataset using method f.
    - **Information Gain (IG) and Mutual Information (MI)**: These methods involve computing entropy-related measures to quantify the amount of information each gene provides about the class labels.
    - **Correlation-based Feature Selection (CFS)**: This method calculates the correlation between each gene and the class labels and assesses inter-gene correlations to identify and remove redundant features.
    - **Statistical Tests (T-test, Wilcoxon, Chi2, Pearson)**: For T-test and Wilcoxon, p-values were derived to indicate the statistical significance of differences in gene expression between classes. Chi2 assessed the dependency between categorical gene features and the class labels. Pearson's correlation measured the linear relationship between gene expression and class labels.
    - **Relief-F**: This method assigns weights to genes based on their ability to differentiate between nearest-neighbor instances from different classes.

- **Gini Index**: This method evaluates how well a gene can separate the data into homogeneous classes based on impurity measures.

2. **Gene Ranking**: Genes were ranked in descending order based on their computed scores.

3. **Feature Subset Selection**: A subset of genes was selected by choosing those genes whose scores were above a preset threshold. Fifteen features were selected for consistency across all methods.

IV. Experimental Implementation

The entire experimental procedure, including data preprocessing, gene ranking and selection, and classification evaluation, was implemented using a Python program. The program was structured into three main phases:

1- Data Preprocessing: It was used for handling the raw microarray dataset to ensure that it is in a suitable format for analysis.
2- Gene Ranking and Selection: Each filter method was used to identify relevant gene subsets.
3- Classification Evaluation: Assessment of the performance of the selected gene subsets using a classifier.

Performance Metrics

To comprehensively assess the efficacy of each gene selection method, the following performance metrics were used:

1. Classification Accuracy: The proportion of correctly classified instances by the classifier using the selected gene subset.
2. Number of Selected Genes: The size of the feature subset selected by each method was fixed at 15 in this study.
3. Computational Cost: The time taken by each filter method to compute the scores and select features.
4. Gene Subset Stability: A measure of how consistent the selected gene subsets are across different runs or folds of the experiment.

These metrics collectively provide a complete view of the performance of each method, considering both the predictive power and practical implementation aspects.

V. Results

In this paper, nine filter-based gene selection methods were evaluated using a gene expression dataset comprising 413 training samples and 177 test samples. The performance of each method was assessed based on the cross-validation accuracy, test accuracy, computation time, and feature selection stability. All methods were configured to select 15 features each.

Overall Performance Summary

The overall performance of each filter method is presented in Table 1. All methods demonstrated perfect stability (1.0000), they indicated that they consistently selected the same set of 15 features across iterations of the evaluation process.

Table 1: Summary of Gene Filter Method Performance

Method	CV Accuracy (Mean ± Std. Dev.)	Test Accuracy	Computation Time (s)	Stability
Information Gain	0.9976±0.0073	0.9887	28.76	1.0000
Mutual Information	1.0000±0.0000	0.9774	43.67	1.0000
CFS	1.0000±0.0000	0.9944	283.74	1.0000
Relief-F	0.9927±0.0111	0.9831	67.38	1.0000
T-Test	1.0000±0.0000	0.9944	26.86	1.0000
Wilcoxon	1.0000±0.0000	0.9774	13.75	1.0000
Chi2	0.9976±0.0071	0.9831	0.76	1.0000
Pearson	1.0000±0.0000	0.9944	6.5	1.0000
Gini	0.9952±0.0096	0.9887	0.47	1.0000

Accuracy Assessment

Mutual Information, CFS, t-test, Pearson achieved a perfect mean CV accuracy of 1.0000±0.0000. CFS, T-Test, and Pearson demonstrated the highest test accuracy of 0.9944. Information Gain and Gini both achieved a test accuracy of 0.9887, whereas Relief-F, Chi2, Mutual Information, and Wilcoxon showed test accuracies of 0.9831 and 0.9774, respectively. Even the lowest CV accuracy observed (Relief-F at 0.9927±0.0111) indicates a very high level of predictive performance.

Computational Efficiency

Figure 1 shows the execution time of methods. Significant differences were observed in the computational times required for each method. Gini was the fastest, completing its execution in 0.474 seconds, closely followed by Chi2 at 0.759 seconds. Pearson and Wilcoxon also demonstrated high efficiency with times of 6.498 and 13.754 seconds. Information Gain and T-Test had moderate runtimes of 28.761 and 26.860 s. Mutual Information and Relief-F were slower, requiring 43.673 and 67.381 s. CFS incurred the longest computation time of 283.738 s.

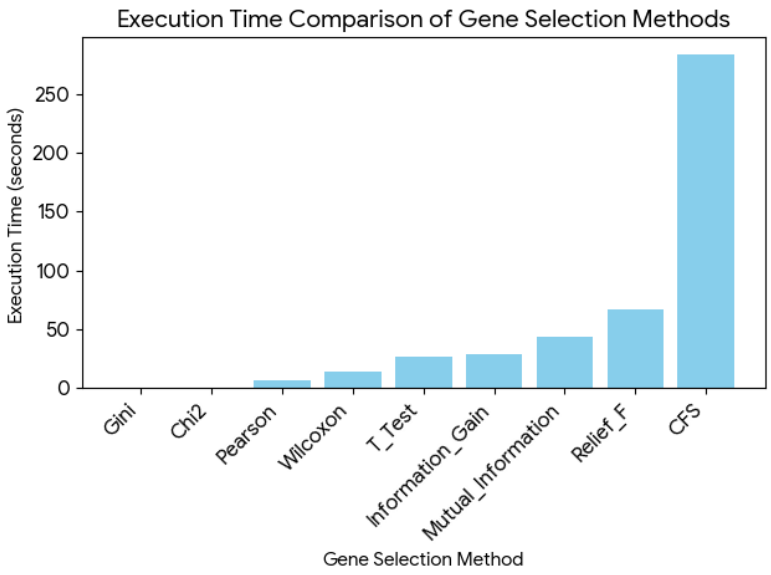

Figure 1. Execution time comparison of the Nine filter methods

Statistical Comparison

A Friedman test was conducted to determine whether there were statistically significant differences in the mean CV accuracies among the methods. The test yielded a statistic of 12.7407, with a corresponding P-value of 0.1211. Because the p-value (0.1211) was greater than the conventional significance level of 0.05, no statistically significant differences were detected among the methods based on their mean CV accuracies. Pairwise comparisons further confirmed that the differences in the mean CV accuracy between the methods were minimal.

Selected Features

Table 2 highlights the top 3 genes identified by each method, illustrating where the selection criteria align or diverge. There is a notable consensus among methods, which is based on statistical correlations and distributions. These distinct feature sets indicate that while the methods achieve similar predictive performances, they leverage different underlying characteristics of the genes for selection.

Table 2: Summary of Gene Filter Method Performance

Method	Top 1 Gene	Top 2 Gene	Top 3 Gene
Pearson	BTNL9	CPA1	ATOH8
T-Test	BTNL9	CPA1	ATOH8
CFS	ATOH8	NEK2	GPAM
Wilcoxon	KLHL29	CAV2	STARD9

Mutual Information	TPX2	NUF2	STARD9
Information Gain	SAMD14	CDC20	TRIM59
Chi2	LOC387911	PDE2A	CXCL2
Gini	ZFP106	DPP3	SHCBP1
Relief_F	GAS2	FREM1	CSN1S1

VII. Discussion

The objective of this study is to comprehensively evaluate and compare the performance of various filter-based gene selection methods using a gene expression dataset. The performance metrics considered were cross-validation (CV) accuracy, test accuracy, computation time, and stability, across a training set of 413 samples and a test set of 177 samples. All methods consistently selected 15 features and exhibited perfect stability (1.0000), indicating high consistency in feature selection across different runs.

Several methods have shown remarkable accuracy performance. Mutual Information, CFS, T-Test, Pearson correlation, and Wilcoxon all achieved a mean CV accuracy of 1.0000±0.0000. This implies that these methods are highly effective in identifying relevant genes that lead to perfect classification accuracy during cross-validation. Although their CV accuracies were identical, their test accuracies showed slight variations. CFS, T-Test, and Pearson correlation achieved the highest test accuracy of 0.9944, indicating superior generalization to the unseen data. Despite their perfect CV accuracy, MI and Wilcoxon yielded slightly lower test accuracies of 0.9774. This difference in test performance highlights the importance of evaluating methods on an independent test set to ascertain their real-world applicability beyond the training phase of the model.

Information Gain and Chi2 also performed strongly, both achieving a mean CV accuracy of 0.9976±0.0073 and 0.9976±0.0071 with test accuracies of 0.9887 and 0.9831. Gini, with a CV accuracy of 0.9952±0.0096 and a test accuracy of 0.9887, also showed robust performance. Relief-F had the lowest mean CV accuracy at 0.9927±0.0111 and a test accuracy of 0.9831, indicating a very high level of performance.

An important factor of filter methods is their computational efficiency. Chi2 emerged as the fastest method, finishing its calculations in an impressive 0.759 seconds. Gini also showed substantial efficiency, with a computation duration of 0.474 seconds. Pearson and Wilcoxon were relatively speedy as well, taking 6.498 and 13.754 seconds, respectively. Information Gain and T-Test experienced moderate computation times of around 28 seconds. Conversely, Mutual Information and Relief-F were significantly slower, with times of 43.673 and 67.381 seconds, respectively. CFS, while exhibiting excellent accuracy, came with the greatest computational cost, needing 283.738 seconds. This notable variation in computation time highlights the trade-off between performance and efficiency, which is a vital consideration for large-scale gene expression datasets.

The Friedman test produced a statistic of 12.7407 and a p-value of 0.1211, which showed that there were no statistically significant differences among the methods at the $p<0.05$ level based on the mean CV accuracy. This indicates that, although there are variances in performance, these variances are not substantial enough to be considered statistically significant across the assessed methods. Pairwise comparisons of the mean CV accuracy further corroborated this, revealing minimal differences between all method comparisons.

Even though the statistics do not indicate significance, practical implications can still be inferred. For instances where computational speed is critical, Chi2, Gini, Pearson, or Wilcoxon tests are favored due to their quick processing times without significantly sacrificing accuracy. In cases where achieving the highest predictive performance on unseen data is essential, CFS, T-Test, and Pearson correlation emerge as the top performers based on their accuracy. The consistent reliability across all methods is a particularly encouraging finding, suggesting that the gene sets selected are dependable and reproducible, independent of the chosen filtering technique.

The varied sets of top 15 features chosen by each method illustrate that, while their overall classification accuracy may be comparable, they reach this outcome by targeting different subsets of genes. For instance, Information Gain highlighted genes such as SAMD14, CDC20, and TRIM59, which are frequently linked to cell cycle regulation and proliferation. Mutual Information identified genes like TPX2, NUF2, and STARD9, known to be involved in mitotic processes. CFS and T-Test arrived at a similar group of genes, including ATOH8, NEK2, and CPA1, suggesting a common emphasis on features associated with cellular differentiation and enzymatic functions. These distinct sets of features imply that each filtering method may uncover different underlying biological connections within the dataset, resulting in similar predictive capabilities through varied mechanistic perspectives.

The feature selection analysis identified several robust biomarker candidates. The gene encoding Thymic Stromal Lymphopoietin (*TSLP*) was ranked among the top 15 features by six of fifteen methods employed: Mutual Information, Correlation-Based Feature Selection (CFS), T-Test, Wilcoxon rank-sum test, Chi-squared test, and Pearson correlation. Two additional genes, *CA4* and *MMP11*, were also consistently selected, each identified by five methods.

In summary, this comparative study shows that filter-based gene selection methods are highly efficient at identifying relevant genes for classification tasks, they show notable accuracy and stability. Although statistically significant differences in CV accuracy were not found among the methods, practical aspects regarding computational time and slight variations in test accuracy and selected feature sets provide direction for method selection. Researchers should consider both computational efficiency and minor improvements in test performance when determining the most suitable filter method for gene selection.

VIII. Conclusion

This study systematically compared nine filter-based gene selection methods for their efficacy in gene expression classification, focusing on predictive accuracy, computational efficiency, and feature selection stability. The findings demonstrate that all evaluated filter methods are highly effective in identifying relevant gene subsets, consistently achieving remarkable cross-validation and test accuracies, often exceeding 0.97. A notable observation was the perfect stability (1.0000) across all methods, highlighting their robustness and reliability in selecting consistent feature sets.

While several methods (Mutual Information, CFS, T-Test, Pearson, Wilcoxon) yielded perfect mean CV accuracies, the highest test accuracies (0.9944) were achieved by CFS, T-Test, and Pearson. This highlights the importance of evaluating performance on independent test sets to assess generalization capabilities. The Friedman test indicated no statistically significant differences in mean CV accuracies among the methods, suggesting that from a purely statistical standpoint, their predictive power is comparable.

However, significant differences were observed in computational efficiency. Gini and Chi2 emerged as the most computationally efficient methods, completing their tasks in under one second, making them highly suitable for large-scale genomic analyses where speed is critical. In contrast, CFS, despite its high accuracy, required considerably more time. The selection of distinct gene sets by each method, while yielding similar classification performance, suggests that different underlying biological signals might be prioritized by various filter criteria.

In conclusion, this research provides valuable insights into the strengths and trade-offs of different filter-based gene selection methods. For applications prioritizing rapid feature selection without compromising high accuracy, methods like Gini or Chi2 are highly recommended. Where the highest possible generalization performance is paramount, and computational time is less of a constraint, CFS, T-Test, or Pearson correlation prove to be excellent choices. The consistent stability across all methods reinforces their utility in providing reproducible gene signatures. Future research could focus on integrating these filter methods with wrapper or embedded techniques to potentially enhance performance or explore their applicability to diverse biological datasets and disease contexts.

References

1. H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Boston, MA: Kluwer Academic Publishers, 1998.
2. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
3. J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
4. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
5. G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. 11th Int. Conf. Mach. Learn.*, 1994, pp. 121–129.
6. D. G. Altman, *Practical Statistics for Medical Research*, 1st ed. London, U.K.: Chapman and Hall/CRC, 1990.
7. F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, 1945.

8. K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philos. Mag. Ser. 5*, vol. 50, no. 302, pp. 157–175, 1900.
9. L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
10. K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. 9th Natl. Conf. Artif. Intell.*, 1992, pp. 129–134.
11. M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1999.
12. N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Front. Bioinform.*, 2022.