Gene Selection for Breast Cancer Classification Using T-Test Filtering and Wrapper-Based Optimization

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
DOI: https://doi.org/10.5281/zenodo.18112300

Abstract. Accurate classification of breast cancer using gene expression data is challenged by the high dimensionality and noise inherent in microarray datasets. This study proposes and evaluates a hybrid feature selection pipeline that integrates T-Test filtering with four wrapper-based optimization methods: Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Genetic Algorithms (GA), and Hill Climbing.

The T-Test was first used to reduce the initial feature space, after which each wrapper method was used to identify an optimal gene subset based on classifier performance. Experimental results show that SFS achieved the highest classification accuracy (98%) and AUC (0.99), while Hill Climbing offered the fastest execution time (1.68 seconds) but with a trade-off in specificity. Notably, the genes SFRP1 and CNTNAP3 were recurrently selected by different methods, suggesting their potential as key biomarkers. To further clarify these findings, we conducted an in-depth analysis of gene overlaps using Jaccard indices and explored the biological roles of selected genes, revealing SFRP1's tumor-suppressive function via Wnt pathway inhibition and CNTNAP3's emerging prognostic value in various cancers.

These findings emphasis the effectiveness of hybrid selection strategies and provide crucial insights into the trade-offs between classifier performance, computational cost, and gene stability in breast cancer classification. This extended analysis incorporates additional performance visualizations, detailed overlap metrics to offer a comprehensive view of hybrid feature selection's applicability in bioinformatics.

Keywords: Breast cancer classification, gene expression, feature selection, T-Test filtering, wrapper methods, Sequential Forward Selection, Genetic Algorithms, Hill Climbing, microarray data, bioinformatics

1 Introduction.

Cancer is one of the common diseases that affects women worldwide, it shares a large percentage of deaths related to cancer, it is also hard to diagnose and predict due to its partial variability. Recently, Microarrays are being used in cancer research because they facilitate the simultaneous measurement of thousands of gene expression

levels, which makes it possible to early diagnose the disease and afford personalized treatment. The analysis of gene expression faces many challenges one of the most important one is high dimensionality where the number of genes exceeds by far the number of samples, which may cause model overfitting, where the model performs well on training data, but it fails on unseen data.

Gene selection became an important step in the analysis pipeline, where we determine a small group of informative gene that can distinguish between cancerous and non-cancerous patients and exclude the irrelevant genes. Gene selection can lead to an efficient model, that have a low computational cost and more generalized.

Gene selections techniques are categorized into filter, wrapper, and embedded methods. In filter methods statistics is used to classify the genes based on their intrinsic characteristics independent of algorithm used, it requires a low computational, but sometimes fails to select genes which are useful when it is used with other genes. In wrapper methods, a specific learning algorithm is used to evaluate the quality of a gene subset. While it requires more computational, they can identify gene that give better results when its used with other genes and this often lead to better classification performance.

This study proposes a selection pipeline uses two-stages, it leverages the strengths of both filter and wrapper methods. In the first stage, we apply a T-Test, a robust statistical filter, to significantly reduce the initial high-dimensional feature space by selecting only the genes that are differentially expressed between breast cancer and normal samples. In the second stage, we employ four distinct wrapper-based optimization algorithms. Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Genetic Algorithms (GA), and Hill Climbing to search the reduced feature space for an optimal gene subset.

The primary objective of this research is to conduct a comparative analysis using four wrapper methods. To assess their effectiveness, we used a comprehensive set of performance metrics, which includes classification accuracy, sensitivity, specificity, AUC, computational cost, and the stability of the selected gene subsets. By systematically evaluating these trade-offs, the study aims to provide a valuable approach on selecting appropriate feature selection strategies for breast cancer classification.

The main contributions of this study are a systematic comparison revealing trade-offs between computational efficiency and classification performance among four wrapper methods, with SFS achieving the highest accuracy (98%). The study also identifies robust biomarker candidates, notably SFRP1 and CNTNAP3, recurrently selected across methods, highlighting their biological relevance for further investigation.

2 Literature Review

Modern DNA microarray and sequencing technologies measure thousands of gene expression levels simultaneously, yielding datasets with far more features than samples [1], [2]. While rich in information, these high-dimensional profiles pose challenges, limited sample sizes lead to overfitting and poor generalization of classifiers [1], [2].

For breast cancer, an especially heterogeneous disease, reducing dimensionality is crucial. Gene selection addresses this by identifying a small subset of genes that carry most of the signal for distinguishing tumor subtypes or detecting tumors early [1], [2]. Selected genes are not only used for more accurate classification but also yield biological insight into cancer mechanisms. Indeed, gene selection preserves interpretability by retaining actual gene variables which helps pinpoint biomarkers [2]. In contrast, generic feature-extraction loses this interpretability [2]. In summary, almost all modern studies on microarray-based breast cancer diagnosis begin with a feature selection step to mitigate the "curse of dimensionality" [1], [3].

Feature selection algorithms are broadly classified into filter, wrapper, embedded, and hybrid approaches [3]. Filter methods score each gene independently using statistical measures for example, t-tests, ANOVA, correlation, or information-theoretic scores – to rank features by relevance [3], [4]. Because filters ignore feature dependencies and do not involve any classifier, they run very fast and can handle thousands of genes [3], [4]. However, this independence means filters may miss combinatorial effects, a gene that is weakly predictive alone but highly informative in combination could be discarded [4]. Wrapper methods overcome this by evaluating subsets of genes directly with a learning algorithm: they "wrap" a classifier (e.g., SVM, k-NN) around the search process [3], [4]. Meta-heuristic search algorithms like genetic algorithms, particle swarm, binary bat, ant colony are often used to navigate the enormous space of subsets [3], [4]. Wrappers typically yield higher accuracy because the chosen genes are explicitly optimized for the specific model, but at great computational cost [3]. Embedded methods integrate feature selection into the model training itself. Finally, hybrid methods combine filter and wrapper steps [3], [4]. For example, a common strategy is to first apply a filter to eliminate a large fraction of irrelevant genes, then use a wrapper to finely search within the reduced set [3], [4].

Among filters, univariate statistical tests are popular for gene selection. For example, genes can be ranked by the p-value of a two-sample t-test comparing cancer vs. normal samples [5]. Alromema et al. [5] used a two-tailed unpaired t-test at significance level 5% to refine an initially small gene set for breast tumor prediction. In that study, t-tests identified genes whose mean expression differed significantly between classes, yielding a set of three top genes (MAPK1, APOBEC3B, ENAH) that together achieved high classification accuracy [5]. Other filter as fold-change, pearson correlation, information gain or chi-squared statistics play a similar role in gene selection. Filters are appreciated for removing obvious noise and reducing dimensionality very efficiently [1], [4]. However, as noted above, they assess each gene in isolation. Hashmi et al. [1] caution that filter-only selection "offers computational efficiency and the ability to reduce dimensionality, but accuracy results are limited." Thus, modern workflows rarely rely on a single filter stage alone but use it as a fast first step to remove irrelevant genes.

Wrapper-based search methods directly optimize classification performance. They can be seen as black-box optimization, a chosen learning classifier (e.g., SVM or random forest) provides a score for any candidate gene subset, and the wrapper algorithm seeks the subset with best score [3], [4]. Genetic algorithms (GA) have been especially dominant; for example, Almugren and Alshamlan [6] surveyed hybrid methods and found that GA-based wrappers are by far the most extensively used, achieving near perfect

accuracies with very few genes. Other popular wrappers include Particle Swarm Optimization (PSO), Bat Algorithm, Harmony Search, and newer nature inspired heuristics. Experimental studies often combine filters and wrappers in multi-phase schemes. Ghosh et al. [6] first aggregated the top-ranked genes from several filters into a candidate pool, then ran a GA to further optimize the subset. Likewise, Hameed et al. [7] applied Pearson correlation filtering before using a Binary PSO or GA wrapper to select final genes.

Hybrid (filter-wrapper) methods are now considered best practice in microarray gene selection [3], [4]. Typically, a multi-stage pipeline uses simple filters to quickly remove the bulk of irrelevant genes, and then uses an optimization-based wrapper on the remaining genes [4]. This two-phase design significantly reduces computational burden while retaining good accuracy. Hashmi et al. [1] note that "filter methods have been utilized individually or combined with the genetic algorithm or wrapper feature selection in order to improve cancer classification."

Because gene-selection is NP-hard [3], many researchers employ evolutionary and swarm algorithms to guide wrapper search. Differential Evolution (DE) is one such method that has shown promise in dimensionality reduction [3], [9]. For example, Zorarpacı and Özel [8] combined DE with the Artificial Bee Colony algorithm for feature selection, while Hancer et al. [21] applied DE with information-theoretic filters. These nature inspired searches, when used after an initial filter stage, can identify compact gene sets that maximize classifier accuracy [3], [6].

High-throughput expression profiling serves as a foundational tool for breast cancer research, facilitating the development of both subtype classifiers and prognostic indicators [10]. In recent work, many studies follow the hybrid FS paradigm. Alromema et al. [5] developed a sequential feature-selection framework for breast tumor diagnosis, they first applied the mRMR filter to reduce redundancy, then a t-test to remove insignificant genes, and finally metaheuristic search to select an optimal gene panel. This pipeline identified just three genes (MAPK1, APOBEC3B, ENAH) that yielded an XGBoost model with ~97% accuracy [5]. Similarly, Information Gain filtering is combined with a Grey Wolf optimizer to pick genes for breast cancer prediction [11].

Research on gene selection for breast cancer classification has converged on hybrid approaches as the most promising strategy. Foundationally, filters like t-tests or information measures quickly discard irrelevant genes [3], [4], while wrappers with metaheuristics fine-tune the selected set for optimal classifier performance [3], [4]. Numerous recent studies confirm that these two stages in combination yield superior accuracy [1], [5], [6]. Going forward, literature suggests continuing exploration of advanced hybrid and ensemble strategies, as well as integration with biological pathway information, to further improve robustness and interpretability of gene-based cancer diagnostics.

3 Methodology

This section details the dataset, preprocessing steps, two-stage feature selection pipeline, and performance evaluation metrics. The overall workflow is designed to be systematic and reproducible.

3.1 Dataset and Preprocessing

This study utilized a publicly available breast cancer gene expression dataset (BC-TCGA) [12] derived from The Cancer Genome Atlas (TCGA), a widely recognized repository for comprehensive genomic data [13]. BC-TCGA consists of 17,814 genes and 590 samples, including 61 normal samples and 529 breast cancer samples.
The dataset comprises high-dimensional gene expression profiles with binary class labels (Positive vs. Negative). Before analysis, the data underwent standard preprocessing

1. **Data Cleaning:** Removal of samples/genes with excessive missing values.
2. **Missing Value Handling:** Mean imputation for remaining missing entries, a robust technique for microarray data [14].
3. **Data Partitioning:** An 80:20 stratified split into training and testing sets to preserve class balance and ensure unbiased evaluation [15].
4. **Label Encoding:** Numerical encoding of class labels into binary format (1/0).

3.2 Two-Stage Feature Selection Pipeline

Our proposed methodology follows a hybrid approach, integrating a statistical filter with wrapper-based optimization methods.

Stage 1: T-Test Filtering

The first stage aims to reduce the high dimensionality of the dataset (17,814 genes) by eliminating genes that are unlikely to be informative. A two-sample independent T-Test was applied to each gene to test the null hypothesis that the mean expression levels are the same between the "cancerous" and "non-cancerous" groups. The p-value obtained for each gene represents the probability of observing the given difference in means or a more extreme one if the null hypothesis were true. The first one hundred Genes with p-values below a predefined significance threshold ($\alpha = 0.01$) were considered statistically significant and were retained for the next stage. This filtering step significantly reduces the computational burden for the subsequent wrapper methods by focusing their search on a much smaller, more relevant subset of genes [16], [17].

Stage 2: Wrapper-Based Optimization

The filtered subset of genes from the first stage was then fed into four different wrapper-based optimization algorithms. Each algorithm used a Support Vector Machine (SVM) [18] with a linear kernel as the core classifier to evaluate the performance of

different gene subsets. SVM was chosen for its robustness in high-dimensional spaces and its effectiveness in binary classification tasks. The objective for each wrapper method was to find a subset of 10 genes that maximized the classification accuracy, evaluated through a 10-fold cross-validation scheme.

The four wrapper methods implemented were:

1. The Sequential Forward Selection (SFS)algorithm was initialized with an empty set, it iteratively added one gene at a time from the T-Test filtered pool, in each step, every candidate gene was temporarily added to the current subset, the SVM was trained and evaluated, and the gene that provided the highest increase in cross-validated accuracy was permanently added to the subset, this process was repeated until 10 genes were selected.
2. The Sequential Backward Selection (SBS) algorithm started with the complete set of genes that passed the T-Test filter, it iteratively removes one gene at a time, in each step, every gene in the current subset was temporarily removed, the SVM was evaluated and the gene whose removal resulted in the smallest drop in accuracy was permanently eliminated, this continued until only 10 genes remained in the subset.
3. The Genetic Algorithm was used to evolve a population of candidate 10-gene subsets, it seeks to maximize the cross-validated accuracy of an associated SVM classifier. The algorithm utilized fitness-proportional selection, crossover, and mutation, where a random gene was swapped from a pre-filtered pool, to traverse the solution space and mitigate local optima. The final output was the highest-fitness subset identified throughout the evolutionary process.
4. The hill climbing algorithm is initialized with a random subset of 10 genes and then generates neighbor subsets by replacing one gene from the current subset with a gene from the outset. Each neighbor subset is evaluated based on the SVM's classification accuracy, if a neighbor achieves a higher accuracy it is adopted as the new current subset. The algorithm terminates when a local optimum is reached, which is when no neighbor subset can improve performance, to reduce the risk of convergence to a suboptimal solution, the process is repeated several times across random initialization and retaining the subset with the highest accuracy.

3.3 Performance Evaluation

A 10-fold cross-validation strategy [19] was employed to ensure a robust and unbiased evaluation of the feature subsets identified by each wrapper method. The performance of the final 10-gene subset from each algorithm was assessed using a comprehensive set of metrics:

- Classification Metrics:
  - **Accuracy:** The proportion of correctly classified samples.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

  - **Sensitivity (Recall):** The proportion of actual positive samples that are correctly identified.

$$Sensitivity = \frac{TP}{TP+FN} \quad (2)$$

- **Specificity:** The proportion of actual negative samples that are correctly identified.

$$Specificity = \frac{TN}{TN+FP} \quad (3)$$

- **Precision:** The proportion of predicted positive samples that are actually positive.

$$Precision = \frac{TP}{FP+TP} \quad (4)$$

- **F1-Score:** The harmonic mean of precision and sensitivity, providing a single score that balances both metrics.

$$F1 - Score = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \quad (5)$$

- **AUC:** The area under the Receiver Operating Characteristic curve, which plots sensitivity against (1 - specificity). AUC measures the overall ability of the model to discriminate between the positive and negative classes, with a value of 1.0 indicating a perfect classifier [20].

- **Computational Efficiency:** The total execution time in seconds required for each wrapper method to complete its search process was recorded. This provides a direct measure of the computational cost and practical feasibility of each algorithm.
- **Subset Stability (Feature Overlap):** The stability and similarity of the gene subsets identified by the different wrapper methods were assessed using the Jaccard Index. This metric quantifies the similarity between two finite sets as the size of their intersection divided by the size of their union:

$$J(A, B) = \frac{|A \cap B|}{|A \cap B|} \quad (6)$$

A value of one indicates that the two subsets are identical, while a value of zero indicates they have no genes in common. This analysis helps to understand how different search heuristics converge and whether they identify a common set of core biomarkers.

4 Results

This section presents the empirical results of our comparative analysis, evaluating the four wrapper methods SFS, SBS, GA, and Hill Climbing based on their computational efficiency, classification performance, and the characteristics of the selected gene subsets.

4.1 Computational Efficiency

The computational cost was measured as the total execution time, it varied dramatically among the four wrapper methods, highlighting the significant impact of the search strategy on the efficiency of each method. The results are summarized in Table 1.

Table 1. Execution Time of Wrapper Methods.

Method	Execution Time (s)
Hill Climbing	1.68
Genetic Algorithm	10.85
SFS	13.04
SBS	117.14
Hill Climbing	1.68

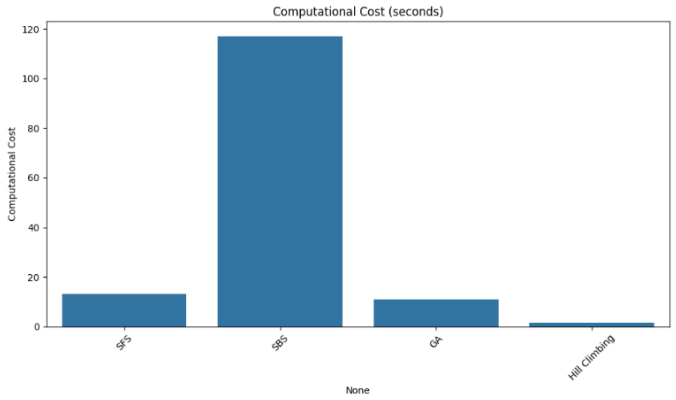

Fig. 1. Computational Cost of Wrapper Methods**.**

As detailed in Table 1 and Figure 1, Hill Climbing was the most efficient algorithm, it completed its search in 1.68 seconds due to its straightforward greedy heuristic. Genetic Algorithms and Sequential Forward Selection demonstrated moderate computational demands, as they finished in 10.85 and 13.04 seconds, respectively, in contrast, Sequential Backward Selection was substantially more costly, it required 117.14 seconds, approximately 70 times longer than Hill Climbing. This inefficiency stems from SBS's top-down approach, which initializes with the full feature set and requires numerous evaluations of a complex SVM classifier. Consequently, Sequential Backward Selection may be impractical for datasets with a large initial feature space.

4.2 Classification Performance

While computational efficiency is important, the ultimate goal of feature selection is to improve classification performance. We evaluated the 10-gene subset selected by each

method using a suite of standard classification metrics. The detailed performance results are presented in Table 2 and visualized in Figure 2.

Table 2. Classification Performance Metrics.

Method	Accuracy	Sensitivity	Specificity	Precision	F1-score	AUC
SFS	0.98	0.99	0.94	0.99	0.99	0.99
SBS	0.98	0.98	0.94	0.99	0.99	0.99
GA	0.98	0.98	0.94	0.99	0.99	0.99
Hill Climbing	0.97	0.98	0.88	0.98	0.98	0.99

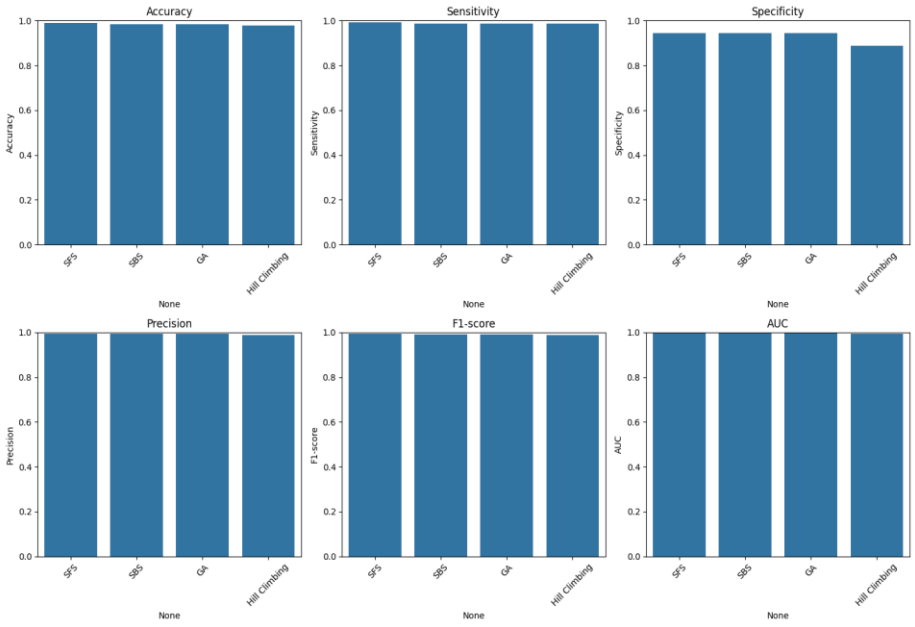

Fig. 2. Comparative Classification Performance.

A central observation is that all four methods achieved exceptionally high performance, with accuracy exceeding 97% and AUC exceeding 0.99. This indicates that the pre-filtered gene set contains a strong molecular signature for classification, and multiple subsets of genes can effectively capture it—a phenomenon akin to the Rashomon Effect in machine learning, where many different models can achieve similar, high predictive power.

Within this high-performance ceiling, SFS was the top-performing method, achieving the highest scores in accuracy (98%), sensitivity (99%), precision (99%), F1-score (99%), and AUC (0.99). SBS and GA delivered identical and excellent performance on most metrics, though the vastly different computational costs (117s for SBS vs. ~11s for GA) make GA a far more efficient option. Hill Climbing, despite being the fastest,

showed the lowest performance, particularly in specificity (88 %) and AUC (0.99). Its lower specificity indicates a greater tendency to misclassify true negative samples, highlighting a trade-off where its search speed is achieved at the expense of classification robustness, likely due to convergence to a local optimum.

4.3 Gene Selection Outcomes and Overlap

Each wrapper method produced a unique 10-gene subset (Table 3). The similarity between these subsets was quantified using pairwise Jaccard Indices (Figure 3), revealing distinct selection patterns.

Table 3. Gene Subsets Selected by Each Method

Method	Selected Genes
SFS	ITM2A, **SFRP1**, OSR1, HOXA4, UBE2E3, **CNTNAP3**, SPRY2, WDR51A, MMP11, TSLP
SBS	ITM2A, **SFRP1**, OSR1, HOXA4, UBE2E3, **CNTNAP3**, SPRY2, MATN2, GJB2, LAGE3
GA	**SFRP1**, LAGE3, PPAPDC1A, DACT2, FIGF, C6orf129, RACGAP1, IRS1, MID1, CCDC19
Hill Climbing	**CNTNAP3**, NVL, KIF4A, NFIB, PLA2G4A, NAP1L2, CDCA5, RHOJ, NTRK2, CDCA8

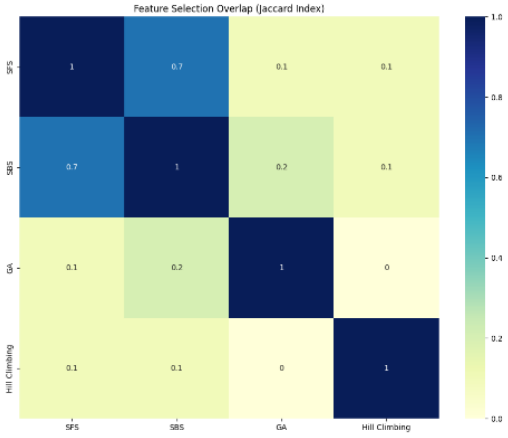

Fig. 3. Feature Selection Overlap (Jaccard Index).

The deterministic methods, SFS and SBS, demonstrated the highest consensus, with a Jaccard Index of 0.7, sharing 7 out of 10 genes (ITM2A, SFRP1, OSR1, HOXA4, UBE2E3, CNTNAP3, SPRY2), this convergence suggests a core set of robustly predictive genes. In contrast, the stochastic methods, GA and Hill Climbing, generated more divergent subsets, showing minimal overlap with each other (Jaccard Index = 0.0)

and with the deterministic pairs. This reflects their tendency to explore disparate regions of the feature space.

Notably, the recurrence of specific genes across methodologically distinct algorithms highlights their potential biological significance. SFRP1 (selected by SFS, SBS, and GA) and CNTNAP3 (selected by SFS, SBS, and Hill Climbing) emerged as particularly robust, marking them as prime candidate biomarkers for further validation.

5 Discussion

This study presents a comprehensive evaluation of wrapper-based feature selection methods within a hybrid pipeline for breast cancer classification. The analysis shows the critical trade-offs between computational efficiency, predictive accuracy, and feature subset stability, while also providing a rationale for the observed performance similarities among the methods.

5.1 Interpreting High performance with Divers Gene Sets

The convergence on high classification accuracy by methodologically distinct algorithms, despite their selection of non-identical gene subsets, is likely rooted in the functional redundancy of biological systems. Since genes operate in correlated pathways, multiple subsets can serve as surrogates for the same overarching cancer phenotype. This result exemplifies the Rashomon Effect, where many models explain the data equally well. The presence of a strong, correlated molecular signature in the dataset allowed each search algorithm to converge on a different but highly predictive solution. Therefore, the results indicate that for this task, there is no unique "best" subset, but rather a collection of genes from which numerous highly accurate classifiers can be built.

5.2 The Performance vs. Cost Trade-Off

The results show a clear trade-off between computational cost and classification performance among the evaluated methods, SBS achieved high accuracy but at a computational cost (117 seconds), which means rendering it impractical for high-dimensional data or scenarios requiring rapid iteration. Conversely, Hill Climbing was computationally efficient (under 2 seconds) but showed lower performance, particularly in specificity, indicating a susceptibility to local optima that compromises its reliability for critical clinical applications. In contrast, SFS emerged as the most effective method overall, it delivered the highest classification accuracy within a reasonable computational timeframe. Its greedy, bottom-up strategy proved highly effective for this task. GA presented a compelling alternative, matching SBS's accuracy but with a tenfold reduction in runtime, as its stochastic search facilitates a more global exploration of the feature space. For this specific problem, SFS represents the most reliable and effective choice, balancing top-tier performance with practical computational demands. However, GA

remains a robust alternative, particularly if the exploration of non-obvious feature interactions is a priority.

5.3 Stability and Diversity in Gene Selection

The composition of the final gene subsets offers significant insight into the behavior of the different search heuristics. The substantial overlap between SFS and SBS (Jaccard Index = 0.7) indicates the presence of a core set of highly informative genes within the filtered data, which deterministic greedy algorithms can reliably identify irrespective of their search direction. This consistency is a desirable property, boosting confidence in the biological relevance of the selected features.
In contrast, the stochastic methods, GA and Hill Climbing, generated markedly more diverse subsets. This diversity is not inherently a limitation; it reflects a capacity to explore disparate regions of the feature space, potentially uncovering novel biomarkers overlooked by more constrained greedy searches. However, such variability can also indicate instability, where different random initializations fail to converge on a consistent signature, complicating biological interpretation. This illustrates a fundamental trade-off in feature selection: choosing between the reliable, stable solution of a deterministic method and the potentially superior, yet less consistent, solution of a stochastic explorer.

5.4 Biological Significance of Selected Genes

A primary objective of gene selection in biomedical research is the identification of robust biomarkers for further investigation. The recurrent selection of SFRP1 and CNTNAP3 across multiple methodologically distinct algorithms provides evidence of their potential significance in breast cancer biology.

- ***SFRP1*** is a recognized modulator of the Wnt signaling pathway, a cascade frequently dysregulated in breast cancer. SFRP1 shows a context-dependent role, it functions as both a tumor suppressor and an oncogene. Its consistent identification by SFS, SBS, and GA in our analysis corroborates its critical role and establishes its expression level as a powerful discriminatory feature for tumor classification.
- While the direct role of *CNTNAP3* in breast cancer is less characterized, members of the contactin family are implicated in cell adhesion and signaling—processes fundamental to cancer progression and metastasis. Its selection by three distinct algorithms marks it as a high-priority candidate for subsequent experimental validation.

The identification of these robust gene candidates highlights the efficacy of integrating computational feature selection with domain-specific knowledge, and provides a strong, evidence-based foundation for subsequent experimental research.

6 Conclusion

This study conducted a systematic comparative analysis of four wrapper-based gene selection methods—SFS, SBS, GA, and Hill Climbing—integrated within a hybrid pipeline with T-Test filtering for breast cancer classification. Our findings provide clear insights into the trade-offs inherent in different feature selection strategies.

1. The hybrid approach, combining the efficiency of a statistical filter (T-Test) with the performance of a model-driven wrapper, is a highly effective strategy for gene selection in high-dimensional microarray data.
2. The similar high performance across methods, despite different selected genes, indicates a strong underlying signal in the data with significant feature redundancy. This means that for this specific task, the choice of feature selection method is less critical for final predictive performance than often assumed.
3. Sequential Forward Selection (SFS) emerged as the most effective method overall, delivering superior classification accuracy (98% accuracy, 0.99 AUC) with a moderate computational cost, making it an excellent choice for practical deployment.
4. A clear trade-off exists between computational speed and classification robustness. While Hill Climbing was the fastest algorithm, its performance was compromised. Conversely, SBS was highly accurate but computationally prohibitive. Genetic Algorithms (GA) provided a well-rounded balance of speed and accuracy, representing a strong alternative to SFS.
5. The study successfully identified a set of robust and potentially significant gene biomarkers, the recurrent selection of SFRP1 and CNTNAP3 across different algorithms strongly suggests their relevance to breast cancer pathology and requires further biological investigation.

Finally, this work emphasizes the importance of a careful and context-aware selection of feature selection algorithms, considering the specific goals of the analysis, whether they prioritize predictive accuracy, computational speed, or biomarker stability.

References

1. A. Hashmi, W. Ali, A. Abulfaraj, F. Binzagr, and E. Alkayal, “Enhancing Cancerous Gene Selection and Classification for High-Dimensional Microarray Data Using a Novel Hybrid Filter and Differential Evolutionary Feature Selection,” Cancers, vol. 16, no. 23, p. 3913, 2024. doi: 10.3390/cancers16233913.
2. F. Wang, A. M. Zain, Y. Ren, M. Bahari, A. A. Samah, Z. A. Shah, N. B. Yusup, R. A. Jalil, A. Mohamad, and N. F. Azmi, “Navigating the Microarray Landscape: A Comprehensive Review of Feature Selection Techniques and Their Applications,” Frontiers in Big Data, vol. 8, p. 1624507, 2025. doi: 10.3389/fdata.2025.1624507.
3. M. A. Hall and L. A. Smith, “Feature Subset Selection: A Correlation-Based Filter Approach,” in Proc. 4th Int. Conf. Neural Information Processing (ICONIP), Dunedin, New Zealand, 1997, pp. 855–858.

4. N. Almugren and H. M. Alshamlan, “A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification,” IEEE Access, vol. 7, pp. 78533–78548, 2019. doi: 10.1109/ACCESS.2019.2922961.
5. N. Alromema, A. H. Syed, and T. Khan, “A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data,” Diagnostics, vol. 13, no. 4, p. 708, 2023. doi: 10.3390/diagnostics13040708.
6. M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, S. Begum, and R. Sarkar, “Genetic Algorithm Based Cancerous Gene Identification from Microarray Data Using Ensemble of Filter Methods,” Med. Biol. Eng. Comput., vol. 57, no. 1, pp. 159–176, 2019. doi: 10.1007/s11517-018-1851-1.
7. S. S. Hameed, F. F. Muhammad, R. Hassan, and F. Saeed, “Gene Selection and Classification in Microarray Datasets Using a Hybrid Approach of PCC-BPSO/GA with Multi Classifiers,” J. Comput. Sci., vol. 14, pp. 868–880, 2018.
8. E. Zorarpacı and S. A. Özel, “A Hybrid Approach of Differential Evolution and Artificial Bee Colony for Feature Selection,” Expert Syst. Appl., vol. 62, pp. 91–103, 2016. doi: 10.1016/j.eswa.2016.06.004.
9. R. Storn and K. Price, “Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces,” J. Glob. Optim., vol. 11, no. 4, pp. 341–359, 1997. doi: 10.1023/A:1008202821328.
10. L. J. van ’t Veer, H. Dai, M. J. van de Vijver, et al., “Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer,” Nature, vol. 415, pp. 530–536, 2002. doi: 10.1038/415530a.
11. S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," Adv. Eng. Softw., vol. 69, pp. 46–61, 2014. doi: 10.1016/j.advengsoft.2013.12.007.
12. Xie, Haozhe; Li, Jie; Jatkoe, Tim; Hatzis, Christos (2017), “Gene Expression Profiles of Breast Cancer”, Mendeley Data, V1, doi: 10.17632/v3cc2p38hb.1
13. J. N. Weinstein et al., “The Cancer Genome Atlas Pan-Cancer Analysis Project,” *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.
14. S. Troyanskaya et al., “Missing value estimation methods for DNA microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
15. I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.
16. S. Dudoit, J. Fridlyand, and T. P. Speed, “Comparison of discrimination methods for the classification of tumors using gene expression data,” *J. Amer. Stat. Assoc.*, vol. 97, no. 457, pp. 77–87, 2002.
17. N. Alromema, A. H. Syed, and T. Khan, “A Hybrid Machine Learning Approach to Screen Optimal Predictors for the Classification of Primary Breast Tumors from Gene Expression Microarray Data ,” *Diagnostics*, vol. 13, no. 4, p. 708, 2023.
18. C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
19. R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in *Proc. IJCAI*, 1995, pp. 1137–1145.
20. J. A. Hanley and B. J. McNeil, “The meaning and use of the area under a receiver operating characteristic (ROC) curve,” *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
21. Hancer, Emrah. (2021). An improved evolutionary wrapper-filter feature selection approach with a new initialization scheme. Machine Learning. 113. 10.1007/s10994-021-05990-z.

انتقاء الجينات لتصنيف سرطان الثدي باستخدام الترشيح باختبار "ت" والتحسين القائم على التغليف

عبدالحميد الواعر[1] ، عبد العظيم دريدر[2]

[1] كلية تقنية المعلومات، جامعة طرابلس
[2] كلية العلاج الطبيعي، جامعة طرابلس
ab.elwaer@uot.edu.ly
DOI: https://doi.org/10.5281/zenodo.18112300

المستخلص. يواجه التصنيف الدقيق لسرطان الثدي باستخدام بيانات التعبير الجيني تحديات ناجمة عن الأبعاد العالية والتشويش في مجموعات بيانات المصفوفات الدقيقة. تقترح هذه الدراسة وتقيّم مساراً هجيناً لاختيار الميزات يدمج أسلوب الترشيح باستخدام "اختبار تي T-Test "مع أربع طرق تحسين قائمة على التغليف، وهي: الاختيار الأمامي المتسلسل، والاختيار الخلفي المتسلسل، والخوارزميات الجينية، وخوارزمية تسلق التل.

استُخدم "اختبار تي" أولاً لتقليل فضاء الميزات الأولي، وبعد ذلك تم استخدام كل طريقة من طرق التغليف لتحديد مجموعة فرعية جينية مثالية بناءً على أداء المصنف. أظهرت النتائج التجريبية أن طريقة حققت أعلى دقة تصنيف (98%) وأعلى مساحة تحت المنحنى بلغت 0.99، بينما وفرت خوارزمية تسلق التل أسرع وقت للتنفيذ (1.68 ثانية) ولكن بمقايضة في مستوى النوعية.

ومن الملاحظ أنه تم اختيار الجينين SFRP1 و CNTNAP3 بشكل متكرر بواسطة طرق مختلفة، مما يشير إلى إمكانية اعتبارهما مؤشرات حيوية رئيسية. ولمزيد من توضيح هذه النتائج، أجرينا تحليلاً معمقاً للتداخلات الجينية باستخدام مؤشرات جاكارد واستكشفنا الأدوار البيولوجية للجينات المختارة، مما كشف عن وظيفة جين SFRP1 الكابتة للأورام عبر تثبيط مسار Wnt، والقيمة الإنذارية الناشئة لجين CNTNAP3 في أنواع مختلفة من السرطان.

تؤكد هذه النتائج فعالية استراتيجيات الاختيار الهجينة وتوفر رؤى حاسمة حول المفاضلة بين أداء المصنف، والتكلفة الحسابية، واستقرار الجينات في تصنيف سرطان الثدي. يتضمن هذا التحليل الموسع تصورات مرئية إضافية للأداء ومقاييس تداخل مفصلة لتقديم رؤية شاملة لمدى قابلية تطبيق اختيار الميزات الهجين في مجال المعلوماتية الحيوية.