Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

Deciphering The Genomic Clues: Advanced Feature Selection Strategies for Cancer Detection in Microarray Gene Expression Profiles

Ankan Bandyopadhyay[1], Ankita Banerjee[2], Dr. Abhishek Bandyopadhyay[3]
1,2 Computer Science & Engineering, Asansol Engineering College, West Bengal, India
3 Computer Science & Engineering (AI&ML), Asansol Engineering College, West Bengal, India

Abstract:- In the quest for early disease detection and efficient treatment, microarray gene data analysis emerges as a pivotal research domain. Public gene expression datasets, reflecting the complex activation profiles of thousands of genes in potential disease patients, present formidable challenges due to high-dimensional feature vectors. Identifying disease-associated genes becomes paramount. This research introduces a novel method fusing feature discretization and selection into a machine learning framework. Our experiments reveal exceptional accuracy, minimal false negatives, and substantial dimensionality reduction. The resultant gene subsets are interpretable by clinical experts, facilitating disease verification. Microarray technology, integral to genetic research, offers diverse applications in health, including disease prediction and cancer investigation. However, analyzing copious raw gene expression data encounters computational complexities. Our research encompasses feature selection methods, crucial for achieving robust cancer classification amidst high dimensions, small sample sizes applicable for both labelled and unlabeled data, and noise. The comprehensive taxonomy of these methods, open research inquiries, and potential inferences are meticulously explored, enriching the field of microarray-based cancer prediction.
Keywords*: Cancer, Classification, Dimensionality, Feature Selection, Microarray*

1. Introduction

Cancer, or malignant neoplasm, epitomizes a profound challenge, characterized by the tumultuous proliferation of cells and their invasive proclivity [1]. The World Health Organization (WHO) underscores the global scourge of cancer, with a staggering 14 million new cases documented in 2012, rendering it a paramount cause of morbidity and mortality. Globally, cancer stands responsible for almost one-sixth of all reported fatalities, claiming 8.8 million lives in 2015, securing its position as the second leading global cause of death [2]. As of 2023, there were an estimated 20 million new cases of cancer worldwide, resulting in 10 million deaths. In 2023, there were around 20 million new instances of cancer globally, leading to 10 million fatalities. It is expected that over the next two decades, the incidence of cancer will increase by approximately 60%, reaching nearly 30 million new cases by 2040. [105].

In the realm of cancer management, early diagnosis and treatment are pivotal to mitigating mortality rates. Venturing into medical data mining, a branch of data analysis, stands out as a promising method to carefully examine, convert, interpret, and present the vast collection of medical reports stored in databases. This enigmatic and challenging pursuit of medical data mining holds the responsibility not just for diagnosing and predicting

diseases but fundamentally impacts matters of life and death. Erroneous classifications or predictions can have catastrophic consequences for patients and their kin. It thus embodies an expert system leveraging machine learning to empower healthcare experts in the precise and expeditious diagnosis and prediction of maladies [3].

In this landscape, microarray data, specifically the microarray technology (MT), assumes paramount significance within cancer research. The imperative for early cancer detection, critical for treatment stratagem and survival, has made MT indispensable [4]. MT empowers biologists to unravel the orchestration of myriad genes in a single experiment, offering profound insights into cellular functionality. This wealth of information serves as a linchpin for diagnosing a plethora of diseases, including Alzheimer's, diabetes, and the multifaceted realm of cancer. Amidst this, gene expression data generated through MT emerge as the touchstone for cancer classification and prognosis. Yet, the genomic landscape is marred by high dimensionality, replete with superfluous, repetitive, and discordant genes that scarcely contribute to disease diagnosis. High-dimensionality intricacies, the delicate balance between gene abundance and sample scarcity, and the ubiquity of data redundancy collectively necessitate the deployment of potent dimensionality reduction strategies in the spectrum of medical data mining and machine learning, with a marked emphasis on cancer prediction [5][6].

Microarray datasets encapsulate the gene expression profiles of numerous genes under specific conditions, typically organized as a matrix. Each row signifies a gene, each column represents a distinct sample (e.g., cells or tissues at specific time points), and the matrix cells embody gene expression levels within these samples. These data serve the purpose of discerning gene expression patterns, especially in comparative analyses of different conditions, such as healthy versus diseased states. AI encompasses various algorithms, with machine learning prevailing, excelling in labeled or unlabeled data analysis [107]. Machine learning techniques have been pivotal in automating the utilization of microarray data, leading to the availability of numerous publicly accessible gene expression datasets [7]. These datasets are instrumental for constructing models capable of predicting disease presence based on an individual's gene expression data. Additionally, from a scientific standpoint, it's imperative to identify the most pertinent genes for disease classification and detection. However, these gene expression datasets exhibit high dimensionality, featuring a multitude of features, which presents challenges for human interpretation. Moreover, they often suffer from a scarcity of instances, significantly fewer than the number of genes/features, amplifying the complications associated with the "curse of dimensionality" [8][9].

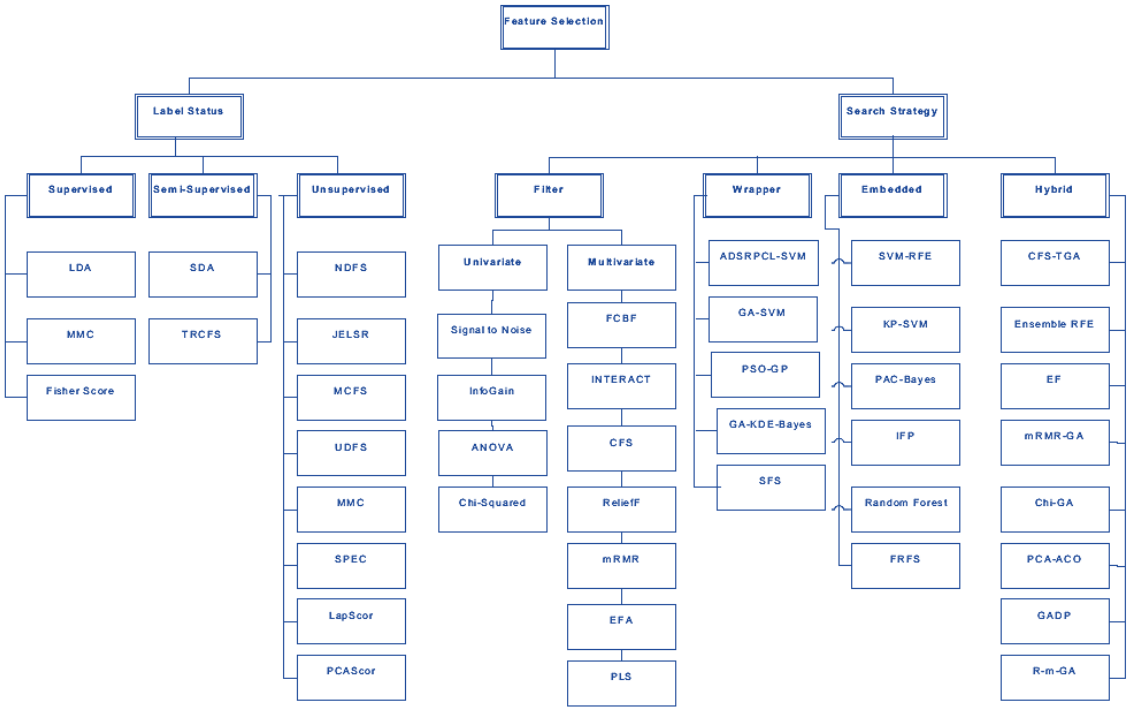

Fig.1. Categorization of the different types of feature selection approaches

Recent research endeavors have proposed machine learning pipelines that encompass Feature Discretization (FD) and Feature Selection (FS) stages for the analysis of microarray data [10][11][12]. The primary objective is to alleviate the curse of dimensionality by reducing data dimensionality and adopting discrete representations of numeric features. Furthermore, these pipelines conduct in-depth analysis on the selected feature subsets, with a focus on identifying the smallest yet informative subset of features for disease prediction. This endeavor is aimed at rendering the selected features interpretable by clinical experts [14].

In the quest for cancer classification accuracy, contemporary research pivots toward the embrace of computational intelligence algorithms, pinpointing the gems among the gene haystack that illuminate the path to precise diagnosis [13]. Notably, a plethora of studies underscores the substantial proportion of genes within DNA microarray datasets that bear no relevance to the arduous task of disease diagnosis. To grapple with the "curse of dimensionality," the mantle of feature selection, synonymous with gene selection, is donned, exquisitely curated to spotlight the genes that embellish disease diagnosis and the mantle of prediction [13][15][16][17][18].

The vast array of genes present in microarray expression data, combined with a restricted pool of patient samples, has brought about a significant shift in cancer prediction and identification. Utilizing this technological progress, accurate classification of cancer now depends on carefully choosing genes that are distinctly linked to each particular subtype of cancer, representing a substantial advancement in oncology research [106]. In a landscape marred by dimensionality challenges, dimensionality reduction assumes the character of a pivotal research epicenter spanning data mining, pattern recognition, machine learning, and statistics. It champions the noble pursuit of enhancing algorithmic classification accuracy by pruning the irrelevant and the superfluous from the microarray dataset, thereby optimizing predictive prowess. A multiplicity of dimensionality reduction approaches vies for primacy, their selection guided by the nature of the dataset and the peculiarities of the domain under scrutiny. Notably, feature selection methods traverse the terrain, parceled into four distinct genres: the filter, the wrapper, the embedded, and the hybrid [19].

The filter algorithms, honed to select features, dissect individual feature characteristics. In juxtaposition, the wrapper strategies enlist the might of machine learning algorithms and evolutionary paradigms to craft beguiling feature subsets. In the realm of large datasets, the filter's celerity renders it the favored choice, albeit occasionally at the cost of accuracy. In the wrapper universe, accuracy reigns supreme, albeit at the expense of computational complexity. However, the nub of the conundrum lies in the intricate relationship between the classifier and the feature interplay, as both the filter and the wrapper exhibit a degree of indifference to these mutual dependencies. Conversely, embedded methods, the unsung heroes, synchronize with specific learning algorithms, yielding the twin promise of classifier interaction and computational expediency [19].

Yet, the terrain remains uncharted, and the literature, though replete with an abundance of feature selection strategies, has hitherto lacked a comprehensive survey. Notably, while some surveys delve into specific feature selection paradigms or their broader medical implications, none have embarked on the panoramic voyage across the myriad feature selection methods, their taxonomy, the peculiar challenges in navigating microarray data, and the realm of microarray experimentation. The present endeavor unfurls its banner to unfurl this expansive panorama, drawing insights from over 150 academic narratives.

This research, with its resolute mission, aspires to illuminate the labyrinthine terrain of microarray cancer datasets, juxtaposing the vivid tapestry of feature selection methodologies. In so doing, it unveils the convoluted realm of microarray experiments and, with unswerving candor, it delineates the limits and outlines the shores of future research in this domain.

Fig.2. Flowchart of the procedures involved in gene microarray data extraction

2. Background

The introduction of Microarray technology has ushered in a transformative era in biological research, bestowing unprecedented insight into the intricate mechanisms governing genetic expression. This technological marvel enables the concurrent exploration of hundreds to thousands of gene activities, ushering in an era of unparalleled research possibilities. While its potential is undeniable, many biologists and interdisciplinary researchers grapple with the complexities of mining and harnessing the wealth of data it yields. Moreover, the results of Microarray experiments find their home in diverse and myriad databases, further complicating the landscape of data management.

The origins of Microarray technology can be traced to the latter part of the 1980s, [20] innovating DNA Microarrays, where about 4000 complementary DNA (cDNA) sequences were carefully organized on nitrocellulose [21]. Subsequently, it has emerged as a powerful tool, allowing biologists to scrutinize the expressions of hundreds of thousands of genes in a single sweep [22][23][24].

Microarray technology has expanded beyond the boundaries of biology and initiated a fresh domain of exploration covering bioinformatics, medical sciences, and machine learning [25][26]. DNA Microarrays, commonly known as DNA chips or biochips, contain an arrangement of tiny DNA spots meticulously positioned on firm surfaces. These frameworks act as portals into the manifestations of numerous genes at once, shedding light on the complexities of genotypes [23].

At the heart of this groundbreaking technology lies the cellular nucleus, housing the DNA that encodes the blueprint for future generations. DNA comprises both coding and non-coding components, with coding segments, known as genes, dictating the structure and function of pivotal proteins. Proteins, the workhorses of organisms, are synthesized in genes through a two-step process: the conversion of DNA into mRNA (transcription), followed by the translation of mRNA into proteins. The progression of molecular genetics tools, such as DNA Microarrays, has offered an advantageous perspective to observe the coordination of cellular functions and examine the simultaneous expressions of tens of thousands of genes.

Gene expression data, portrayed as the total transcribed mRNA within a genomic system, offer the gateway to comprehending how genotype metamorphoses into phenotype. A plethora of standardized approaches, including differential display, Microarray matching, RNA-seq sequencing, and Serial Analysis of Gene Expression (SAGE), aid in detecting differences in gene expression [28][29][30].

The cornerstone of Microarray experimentation lies in the hybridization reaction, which involves comparing the relative mRNA from a pair of tissue samples. This essential procedure occurs as RNA molecules or individual strands of DNA combine to form double-stranded complexes. The primary objective of gene expression Microarray experiments is twofold: firstly, to examine differential gene expression between groups through class comparison, and secondly, to forecast and explore classes for conducting classification studies [23].

This meticulous endeavor unfolds through four key phases. The preparation and labeling of samples inaugurate the journey, involving RNA extraction from specific tissues and subsequent labeling depending on the chosen technology. Hybridization, the second phase, marks the union of DNA or RNA probes with their complementary sequences in the hybridization array, forming Watson-Crick base pairs. Detection methods span optical, electrochemical, and mass-sensitive devices [23]. The washing stage ensues, eliminating non-specifically bound cRNA molecules from the microarray surface, thereby mitigating background and sensitivity effects. Finally, the image acquisition stage unveils the hybridized array's visual representation.

The datasets born from Microarray experiments manifest as extensive matrices (M x N), wherein rows represent samples and columns signify genes or features. The volume of Microarray data is substantial, where M corresponds to samples and N represents genes, with each cell housing a numeric value denoting gene expression in a sample [31].

In summary, the realm of Microarray technology has opened new frontiers in genomics research, unraveling the complexities of gene expression and bestowing invaluable data for a multitude of scientific endeavors.

Table I. Table shows the merits of each of the different Feature Selection Techniques

MODELS	MERITS
Filter models	**Univariate**
	Capable with high-dimensional data, rapid computation, algorithm-agnostic, and model-independent, ensuring versatile, efficient feature selection.
	Multivariate
	Addressing feature dependencies, agnostic to classifiers, offering superior computational efficiency compared to wrapper techniques, and maintaining model independence.
Wrapper models	**Deterministic**
	This methodology is characterized by its emphasis on simplicity, engagement with the classifier, representation of feature interdependencies, and computational efficacy, surpassing randomized techniques.
	Randomized
	This approach's robustness against local optima, active engagement with the classifier, representation of feature interdependencies, and improved classification efficacy highlights its strengths.
Embedded models	
	Effective interaction with classifiers, superior computational efficiency compared to wrappers, and the ability to model feature dependencies make this method exceptional.

3. Feature Selection Techniques and its types

Table II. Table shows the demerits of each of the different Feature Selection Techniques

MODELS	DEMERITS
Filter models	**Univariate**
	Neglects feature interactions, lacks dependency modeling, leading to subpar classification results without engagement with the classifier.
	Multivariate
	Significantly slower and less scalable than univariate methods, lacking interaction with the classifier and underperforming in classification tasks.
Wrapper models	**Deterministic**
	Wrapper methods, despite their effectiveness, present risks of overfitting and can be more susceptible to local optima due to their classifier-dependent, greedy search approach, making them challenging to navigate.
	Randomized
	Compared to deterministic algorithms, wrapper methods are computationally intensive and carry a higher risk of overfitting due to their classifier-dependent selection, making them less efficient for feature selection.
Embedded models	
	Selection techniques are dependent on the classifier used.

A. Search Strategies

The quest for optimal feature selection (FS) methods is akin to traversing a labyrinth, with multiple strategies and pathways to explore. These strategies are differentiated based on their searching approaches, leading us down three distinct routes: filter techniques, wrapper techniques, and embedded techniques.

Filter methods, akin to experienced scouts, commence the feature selection (FS) task prior to engaging in classification or clustering. In a dual-phase procedure, they first evaluate all characteristics and organize them according to predetermined criteria. The subsequent phase entails selecting the highest-ranking attributes. Filter techniques assess the significance of features by investigating the inherent properties of the data. Occasionally, a score determining feature relevance is computed, leading to the swift elimination of features with low scores. The crème de la crème of features, having earned high ranks, secures their position [32][33][34].

Some widely recognized filter methods are Receiver Operating Characteristics Analysis [38], Fuzzy Logic for identifying redundant features [42], ReliefF [35][36], Maximum-Minimum Correntropy Criterion [45], Mutual Information [46], Information Gain [36][39], Laplacian Score [40], F-Statistic [41], Consensus Independent

Component Analysis using gene expression for cancer classification [43], T-Test Feature Ranking for gene selection [44], and Signal-to-Noise Ratio [37].

Filter methods are impartial to specific machine learning algorithms, focusing on the general intrinsic properties of the data. This autonomy renders them computationally efficient with robust generalization capabilities.

In contrast, wrapper and embedded techniques require the involvement of machine-learning algorithms to execute FS. Wrappers, like discerning critics, employ a learning algorithm to evaluate candidate feature subsets. Their close interaction with the classifier contributes to their higher computational cost, although they often outperform filters. Wrappers scrutinize FS through the lens of the learning algorithm, such as the employment of Support Vector Machine-based Recursive Feature Elimination (RFE) [47] to identify important genes linked to cancer ailments.

Embedded techniques tread the middle path amidst filters and wrappers. Feature selection (FS) is seamlessly integrated into the training phase within the algorithm used for learning. During training, investigation into the occurrence of the best selection of features happens, such as the optimization of weights in a neural network. This integration results in embedded techniques being more computationally efficient compared to wrappers.

The landscape of FS strategies becomes even more intricate with the advent of hybrid approaches. Hybrid approaches fuse two or more FS algorithms of different search strategies sequentially, allowing for a judicious combination of computational efficiency and fine-tuning, much like a well-conducted orchestra. For example, a less computationally demanding filter may first prune the feature set, paving the way for a more complex and resource-intensive wrapper to carry out the final refinement.

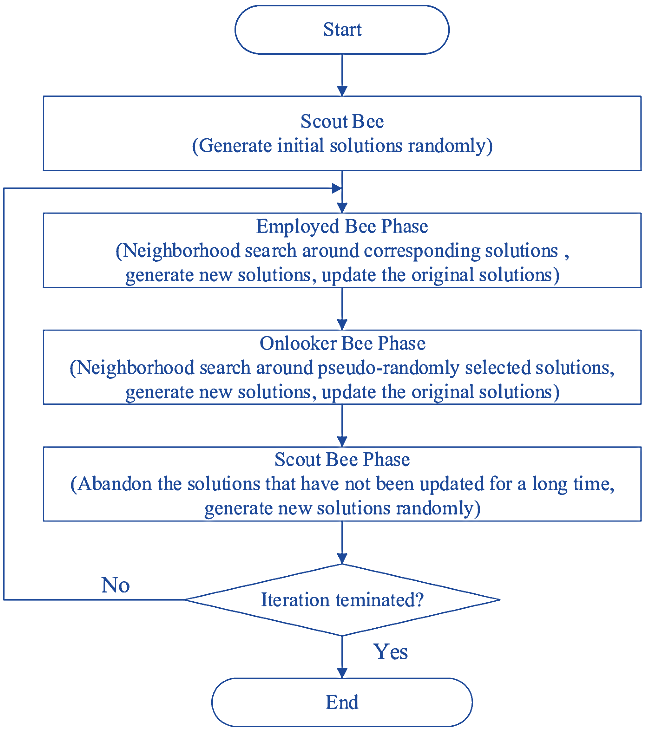

Fig.3. Artificial Bee Colony (ABC) Algorithm

B. Filter Methods

In the intricate realm of feature selection (FS) for predicting cancer using microarray gene profiles, the quest to uncover the most informative genes mirrors a harmonious symphony, with each FS technique playing a unique part in the composition. Among these, filter techniques stand as the initial movements in this orchestration, drawing inspiration solely from the intrinsic characteristics of the data.

Filter techniques, akin to soloists, rely on the inherent data characteristics, primarily statistical measures, to assess the relevance of individual genes or subsets concerning class labels. Various traditional filter methodologies, such as Fast Correlation-Based Filter (FCBF), ReliefF, Consistency-Based Filter, and Correlation Feature Selection (CFS), have been prominently featured in the realm of microarray data analysis. However, these virtuosos often focus solely on the individual features, neglecting the harmonious interplay among them, leading to comparatively lower accuracy in diagnostics compared to other FS methodologies.

Filter methods, nevertheless, offer their own virtues: they are computationally simple, easily scalable to high-dimensional datasets, and exhibit independence from specific classifiers. Furthermore, these techniques can be categorized into single variable analysis and multi-variable analysis. Univariate technique, which evaluate features one by one, often provide simplicity and speed. Notable examples include the signal-to-noise ratio. Conversely, multivariate methods, such as forward feature selection and base-pair selection, navigate the intricate interdependencies among features within subsets. Yet, they require more computational resources and must steer clear of overtraining.

Filter techniques have contributed significantly to various research endeavors, showcasing their diverse repertoire of methods. However, their chief limitation lies in their detachment from the learning algorithms, potentially resulting in suboptimal feature subsets and, in some cases, redundancy.

Distinguished Filter Techniques:

a. *Correlation-Based Feature Selection (CFS):* A multivariate filter algorithm conducts feature subset ranking based on correlation-driven heuristic evaluation [48]. This methodology prioritizes feature subsets that exhibit high correlation with the class label while ensuring minimal redundancy among the chosen features.

b. *The Fast Correlation-Based Filter (FCBF) technique:* This multivariate algorithm excels at evaluating both feature-class and feature-feature correlation. It utilizes Symmetrical Uncertainty (SU) to pick out features strongly associated with the class, employing heuristic methods to remove redundant features [49].

c. *The INTERACT algorithm:* Utilizing SU goodness measurement and introducing consistency contribution, it ranks features by their evaluated SU values. It also assesses the impact of excluding each feature on the consistency of the FS. Features that exceed a set consistency contribution threshold are selected [50].

d. *Information Gain:* A univariate filter approach that calculates mutual information for each class-attribute pair. Features are ranked based on their information gain values [51].

e. *ReliefF:* A prominent multivariate filter based on nearest neighbor concepts. It selects attributes that differentiate instances from different classes while maintaining consistent values within the same class [52]. ReliefF excels in multiclass scenarios and demonstrates resilience against missing and noisy data.

f. *The minimum Redundancy Maximum Relevance (mRMR) algorithm:* This methodology highlights features with high relevance to the target class and minimal redundancy, guided by mutual information criteria [39].

g. *Consistency-Based Filter:* A multivariate methodology that chooses subsets of features relying on their consistency with the class, incorporating an inconsistency criterion [53].

In this intricate symphony of feature selection, these filter techniques, each a maestro in its own right, contribute their distinct melodies to the composition, showcasing their rich diversity and unique capabilities in the pursuit of predicting cancer using microarray gene profiles.

C. Wrapper Methods

Wrapper techniques constitute the second category of feature selection methods, and they are known for using evolutionary approaches in their search strategies. These methods typically begin with a population of solutions, each representing a feature subset. Subsequently, a learning algorithm is used to assess the fitness of each subset, and an iterative process is employed to optimize the feature selection. In the literature, notable wrapper techniques cited include Genetic algorithm with SVM [58], Particle Swarm Optimization [59], Distance Sensitive Rival Penalized Competitive Learning – Support Vector Machine (ADSRPCL-SVM) [55], Artificial Bee Colony (ABC)

algorithm [56][57], Genetic Programming (utilized for predicting alternative mRNA splice variants) [60], and Ant Colony Optimization [54].

These techniques often outperform filter methods because they enable connections between results and indicators. In these methodologies, the feature subset search and model selection are integrated. The wrapper approach defines a search space for possible feature subsets, generating and evaluating many feature subsets. Each subset is evaluated using a specific classification algorithm tailored to a specific learning model. The exploration algorithm involves the classification method to investigate the complete space of feature subsets. However, as the number of features increases, the feature subset search space grows exponentially, which can become a significant challenge, especially for high-dimensional datasets. This is why wrapper techniques are relatively less common in the literature.

Wrapper methods were more prevalent during the initial period of Microarray data analysis. They engage in exploration within the genetic scope, wherein the merit of every gene subset is assessed by examining the precision attained using the designated learning algorithm. For instance, Inza, Sierra, Blanco, & Larrañaga (2002) utilized typical wrapper techniques such as floating selection, sequential forward and backward selection, and best-first search on three Microarray datasets [61]. However, the use of wrapper techniques in Microarray data analysis is less widespread in comparison to alternative feature selection methods.

Recent studies have explored new wrapper techniques. Sharma, Imoto, & Miyano (2011) presented an approach called Successive Feature Selection (SFS) [62]. This approach strives to surpass the constraints of singular ranking and forward selection methods. SFS divides the features into smaller segments and picks the most outstanding features from each segment, considering their accuracy in classification. It subsequently contrasts these outstanding features to determine the best set of features, leading to superior classification accuracy across numerous DNA Microarray datasets. In 2013, Wanderley, Gardeux, Natowicz, and de Pádua Braga presented the evolutionary wrapper technique known as Genetic Algorithm-Kernel Density Estimation (GA-KDE-Bayes), offering another illustrative instance in this domain [63]. This approach utilizes a Bayesian classification algorithm along with a non-parametric density estimation method. The authors clarified that non-parametric methods are apt for analyzing sparse and limited data, particularly observed in bioinformatics analysis, as they abstain from depending on pre-established assumptions about data structure and instead utilize the data itself for details. Their approach outperformed other methods on six Microarray datasets.

In summary, wrapper techniques excel in optimizing feature selection for specific learning algorithms and considering feature dependencies. However, they can be prone to overfitting, especially in scenarios with limited sample sizes, and are computationally expensive, particularly when dealing with a high number of features. The wrapper approach is less common in the literature compared to filter methods due to its computational demands.

D. Embedded Methods

Although filter methods are efficient in terms of computation, they lack engagement with the classifier, frequently resulting in less-than-optimal classifier performance compared to wrappers. In contrast, wrapper methods, though effective, are linked with significant computational expenses, especially when handling Microarray data. A middle ground between these two methods is identified in embedded techniques, which utilize the fundamental learning algorithm to evaluate feature condition. Embedded techniques aim to reduce computational time for evaluating various feature subsets by incorporating feature selection into the learning process. This integration serves as the primary purpose of embedded methods. An example of an embedded technique is Support Vector Machine based on Recursive Feature Elimination (SVM-RFE), introduced by Guyon, Weston, Barnhill, and Vapnik (2002) [47]. SVM-RFE was expressly crafted for the purpose of choosing genes in the classification of cancer. It repeatedly trains the SVM classifier with feature sets and eliminates the least significant features as indicated by the classifier.

A novel embedded technique is Kernel Penalized SVM (KP-SVM), proposed by Maldonado, Weber, and Basak (2011) [64]. KP-SVM identifies important features while constructing the classifier by penalizing individual features employed in the dual formulation of SVM. It improves the radial basis function (RBF) Kernel structure by eliminating features of minimal significance for the learning model. Test outcomes from standard Microarray

datasets and practical datasets indicated that KP-SVM surpassed other methods while utilizing fewer pertinent features.

In response to the problem of imbalanced data in specific Microarray datasets, Anaissi, Kennedy, and Goyal (2011) introduced an integrated technique relying on the random forest algorithm [65]. This method utilizes diverse strategies and algorithms to manage intricate gene expression data within Leukemia datasets. It seeks the optimal training error cost for distinct classes, addresses data imbalance, uses random forest for feature selection, and applies strategies to prevent overfitting. The results showed significant improvements in classification performance.

Another embedded technique, Probably Approximately Correct (PAC)-Bayes feature selection, was presented by Shah, Marchand, and Corbeil (2011) [66]. PAC-Bayes provides viable classification performance using fewer significant features.

Iterative Perturbation Method (IFP), introduced by Canul-Reich et al. (2012), is an embedded gene selector [67]. It detects the less important features using a method that eliminates features in reverse order, focusing on classifier performance when features are perturbed by noise. Characteristics are deemed important if their inclusion of random or extraneous data significantly alters the performance of the classifier. The IFP algorithm exhibited comparable or better overall accuracy within individual classes when contrasted with SVM-RFE across three of four datasets.

In another approach, Wang, Song, Xu, and Zhou (2013) [71] introduced the First Order Inductive Learner (FOIL) Rule-based feature subset selection algorithm (FRFS). FRFS initially generates FOIL classification rules using a modified propositional implementation of the FOIL algorithm. It merges subset features obtained from rule antecedents, eliminating redundant features while preserving interactive and informative ones. FRFS assesses the importance of features within the chosen subset using a fresh measurement termed Cover Ratio and removes unimportant features.

E. Label Methods

In the intricate realm of feature selection (FS), categorization based on label status provides a foundational framework for understanding the diverse array of methods at our disposal. These methods operate under the guidance of sample labels, where the presence of explicit information empowers the selection of pertinent features to discriminate between various sample classes. This categorical foundation leads us to a trichotomy of supervised, semi-supervised, and unsupervised FS methods, each carving its own path in the quest for feature optimization.

Within the realm of labeled data, supervised FS methods [68][69] reign supreme. Armed with the knowledge of explicit sample labels, these methods meticulously curate feature subsets that distinguish samples across diverse classes. The universe of label-aware FS methods traverses an expansive terrain, leveraging well-defined class boundaries to unveil the most informative features.

Intriguingly, semi-supervised FS methods tread a delicate balance between the known and the unknown. Here, a fraction of data is endowed with labels, while the rest linger in the realm of the unlabeled. Drawing inspiration from the realm of graph theory, many semi-supervised FS method sculpt feature selection landscapes guided by the intricate dance of similarity matrices and graph structures. This category thrives on harnessing the duality of labeled and unlabeled data to optimize feature selection.

On the contrary, unsupervised FS methods navigate the labyrinth of feature optimization without the beacon of labeled data. Resourceful and ingenious, these methods employ cunning strategies to sift through unlabeled datasets, illuminating the path toward the most relevant features. In a world devoid of labels, unsupervised FS techniques take center stage, proving that sometimes, the absence of guidance can spark the most innovative solutions [70].

In its most elemental form, FS revolves around the meticulous evaluation of individual features, each vying for a place of prominence in the grand tapestry of data. These features are appraised for their correlation with class labels, a fundamental tenet in the quest for the most informative feature subsets [70] Nevertheless, echoing the

wisdom of Hall (1999), we recognize that the most powerful feature subsets are those in which the features harmonize, avoiding the cacophony of strong inter-correlations.

F. Supervised Methods

In the intricate tapestry of feature selection (FS), supervised methods stand as sentinels of label-rich data, weaving a narrative that hinges upon the vital presence of explicit sample labels. These methods, underpinned by the bedrock of labeled data, embark on a quest to distill the essence of feature relevance, serving as indispensable tools for predictive analytics.

In the realm of supervised FS, a spectrum of approaches unfolds, each tethered to the rich tapestry of labeled data points. The conservative guardians of this domain, like the venerable Fisher Score, undertake the arduous task of individually ranking features, each in isolation, with little heed to feature interplay. In their pursuit of feature distinction, these methods pay homage to classical techniques.

Linear discriminant analysis (LDA), tracing its origins back to the work of Fisher in 1936, casts its discerning eye on the heart of feature selection. It aspires to craft feature subsets that maximize the delicate balance between 'between-class scatter and within-class scatter.' However, when faced with the limitations of small-scale data, LDA encounters significant obstacles [72]. To circumvent these limitations, ingenious solutions arise. Enter the maximum margin criterion (MMC), a harbinger of transformation, as demonstrated by Li, Jiang & Zhang (2004) [73]. It ingeniously transforms the calculation of the ratio between 'between-class scatter and within-class scatter' into a form based on subtraction, transcending the shackles of small sample size.

However, it's imperative to recognize that supervised methods, while potent, possess an insatiable appetite for appropriately labeled data. The efficacy of these methods hinges upon the richness of labels. When the wellspring of labeled data runs dry, as elegantly expounded by Luo et al. (2013), the performance of these methods wanes, a stark reminder of their data-dependent nature [74]. The fickle dance of feature selection thrives on the nuances of labeled data, emphasizing the symbiotic relationship between supervised methods and their labeled companions.

G. Unsupervised Methods

Within the labyrinthine realm of feature selection (FS), the enigmatic domain of unsupervised methods unveils itself as a daunting challenge. Devoid of the guiding beacon of labelled data, unsupervised FS grapples with the formidable task of navigating the intricate landscape of data without a roadmap. This challenge has propelled it into the spotlight as an intriguing frontier of exploration [70].

In this uncharted territory, the quest for informative features unfolds along diverse criteria, reflecting the ingenuity of FS researchers. A notable method entails choosing characteristics that maintain the inherent structure of the initial data manifold. Other avenue, often traversed, embarks on the quest to uncover cluster indicators through the lens of clustering algorithms, eventually converging with a supervised approach to problem-solving.

Unsupervised FS manifests itself in two distinct forms. The first approach intertwines the pursuit of cluster indicators with supervised feature selection. Exemplifying this fusion, Yang et al. (2011) fashioned a comprehensive framework, combining nonnegative spectral clustering and structural learning to distinguish characteristics from the data. [75].

The second approach follows a hierarchical trajectory. It initiates by seeking cluster indicators, then delves into the depths of feature selection, and culminates with iterative cycles of these stages as far as a specified condition is satisfied. In their 2011 work, Zhao et al. introduced Similarity Preserving Feature Selection (SPFS), a proficient approach adept at managing feature redundancy concerns [76]. This category encompasses various methods such as Joint Embedding Learning and Sparse Regression (JELSR) [79], Nonnegative Discriminative Feature Selection (NDFS), Unsupervised Discriminative Feature Selection (UDFS), Embedded Unsupervised Feature Selection (EUFS) [77], and multi-cluster feature selection (MCFS) [78].

Recent forays into this domain have witnessed a convergence of learning mechanisms and manifold structures, resulting in a fascinating blend of techniques. Prominent methodologies in this domain consist of Laplacian Score (LapScor) [40], PCA Score (PcaScor) [80], Minimum Redundancy Spectral Feature Selection (MRSF) [58],

MCFS [78], and Spectral Feature Selection (SPEC) [50]. These techniques embark on the intriguing journey of employing manifold structures, complemented by diverse metrics, to rank each feature's significance. The embellishment of limited restrictions in multi-output regression as witnessed in the MCFS and MRSF approaches [79], marks a pivotal stride in this expedition through the terra incognita of unsupervised feature selection.

H. Semi-Supervised Methods

In the intricate confluence of feature selection methods, semi-supervised approaches emerge as a bridge between the worlds of labeled and unlabeled data. Like skilled cartographers and researchers chart a course through the uncharted territory, demonstrating their prowess in selecting features from the vast expanses of unlabeled data, even when provided with only a limited compass of labeled samples.

At the heart of these semi-supervised techniques lies the notion that data samples predominantly reside on a low-dimensional manifold. This hypothesis paves the way for the application of graph Laplacian-based methods, exemplified by the likes of semi-supervised Discriminant Analysis (SDA) [81]. These approaches leverage the graph Laplacian matrix to harness the latent information harbored in unlabeled samples.

One notable protagonist in this semi-supervised saga is the trace ratio criterion for feature selection (TRCFS), a potent algorithm celebrated for its effectiveness in discerning informative features [82]. However, it is crucial to highlight that the semi-supervised facet of feature selection, while potent, is not without its challenges. Its computational demands, particularly when confronted with vast datasets, can be a formidable hurdle to overcome. The cost of computation time associated with the intricate graph structures poses a formidable challenge [83]. This characteristic reinforces the notion that while semi-supervised methods possess the ability to navigate the interface of labeled and unlabeled data, their effectiveness in the realm of immense datasets is a frontier that remains to be fully conquered.

4. Hybrid Methods

In the dynamic realm of microarray gene profiling, the focus has transitioned from conventional feature selection (FS) methods mentioned earlier to an increasing surge of hybrid and ensemble approaches. These avant-garde strategies harmonize the strengths of diverse FS methods, orchestrating a symphony of choosing characteristics in datasets with numerous dimensions, particularly in the realm of Microarray datasets.

Hybrid techniques, a contemporary paradigm, marry the virtues of filter and wrapper approaches. Their genius lies in their ability to balance feature reduction's time complexity and the quest for an optimal feature subset. The core philosophy of hybrid methods is to employ a filter approach initially, eradicating irrelevant features and diminishing the dimensionality of the original dataset. Subsequently, a wrapper technique takes the stage, seeking the finest features within the curated pool, thereby accelerating the feature selection process. Advocates of this approach posit that the filter's threshold for feature ranking can be set low, minimizing the risk of discarding valuable predictors.

In recent overtures, researchers have introduced their own virtuoso performances. In 2007, Wang et al. executed a paired-step Microarray data analysis, merging T-test and Class Separability (CS) grading to pick genes and a refined classifier for categorization [84]. They utilized a K-nearest neighbors (KNN) algorithm to manage absent values and contrasted the outcomes against both KNN and Support Vector Machine (SVM) classifiers. Rangasamy (2009) developed a framework that prioritized genes using conventional statistical techniques and two distinct machine learning algorithms for diverse datasets [85]. Martín-Merino & De Las Rivas (2009) presented Kernel Alignment KNN for categorizing cancer using gene expression profiles, surpassing traditional KNN [86]. Revathy & Amalraj (2011) merged SVM with an enrichment score for categorizing cancer in Microarray data [87].

Ghorai, Mukherjee, Sengupta & Dutta (2010) crafted a hybrid computer-aided diagnosis (CAD) framework based on filter and wrapper techniques, employing Minimum Redundancy Maximum Relevance (MRMR) ranking for feature selection [88]. El Akadi, Amine, El Ouardighi & Aboutajdine (2011) introduced a two-stage approach

with MRMR for gene filtering and GA for gene subset generation, combined with Naïve Bayes (NB) and support vector machine (SVM) classifiers [89].

Rajeswari & Reena (2011) harnessed Support Vector Machine (SVM) and Fuzzy Neural Network (FNN) for tumor cell identification, demonstrating superior diagnostic accuracy compared to conventional methods [90]. Sahu & Mishra (2012) innovatively used Signal-to-Noise Ratio (SNR) score, coupled with Particle Swarm Optimization (PSO), to reduce dimensionality before employing K-nearest neighbor (KNN), Probabilistic Neural Network (PNN), and Support Vector Machine (SVM) as classifiers [91].

Swathi, Babu, Sendhilkumar & Bhukya (2012) devised an Adaptive Resonance Theory (ART1) a network framework for identifying breast cancer, yielding outstanding outcomes for datasets in unsupervised machine learning. Dev, Dash, Dash & Swain (2012) ventured into the world of three classifiers: Backpropagation network (BPN), Functional Link Artificial Neural Network (FLANN), and PSO-FLANN, showcasing PSO-FLANN's remarkable classification prowess [93].

Abeer, Basma, El-Sayed & Abdel-Badeeh (2013) examined differentially expressed genes (DEGs) in Microarray data, using t-tests and KNN to assess their influence on learning accuracy [94]. Shreem, Abdullah & Nazri (2014) [95] introduced the hybridization of the Harmony Search Algorithm (HSA) and Markov Blanket (MB) as HSA-MB, performing gene selection in classification tasks. Abeer & Basma (2014) masterfully combined gene ranking methods with KNN and SVM classifiers, achieving remarkable results [96].

Alshamlan, Badr & Alohali (2015) introduced the Genetic Bee Colony (GBC) algorithm, showcasing its high classifier accuracy with a compact gene subset [97]. Doreswamy & UmmeSalma in 2016 orchestrated a Binary Bat Algorithm (BBA)-based Feedforward Neural Network (FNN) hybrid model, achieving excellent accuracy for breast cancer classification [3]. Alomari et al. (2017) harmonized Minimum Redundancy Maximum Relevancy (MRMR) with the Bat Algorithm (BA) and a support vector machine (SVM), demonstrating the robustness of their approach on diverse Microarray datasets [13].

In conclusion, the stage is set for hybrid feature selection approaches, a symphony of innovation, where the union of filter and wrapper techniques yields both elegance and efficiency in the realm of Microarray data analysis. Researchers continue to explore new melodies, seeking the perfect harmony between feature reduction and optimal subset selection.

5. Units Some more works in the fields of Feature Selection

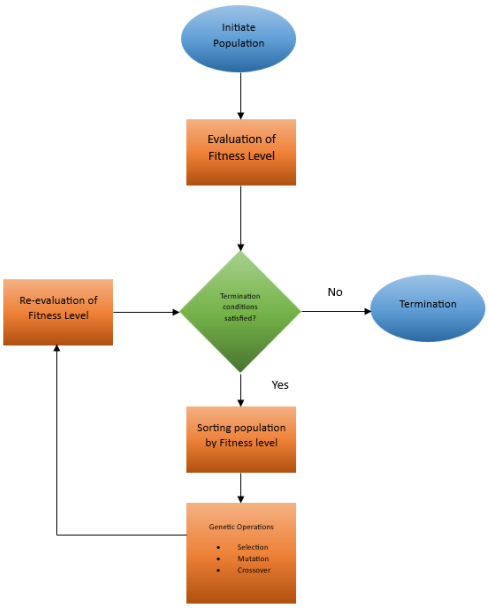

Fig.4. Diagrammatic representation of the basic Genetic Algorithm

Within the captivating realm of hybrid feature selection methodologies, an ensemble of brilliant duets and harmonious ensembles have graced the stage, each offering a unique blend of techniques and approaches. Here, we spotlight these noteworthy compositions:

a. *Chi-Squared statistics with Genetic Algorithm (GA):* This duo combines the statistical power of chi-squared analysis with the evolutionary prowess of genetic algorithms to create a powerful hybrid [98].

b. *Information gain with a novel memetic algorithm:* Information gain, a staple in feature selection, finds a partner in a novel memetic algorithm, resulting in a fresh take on feature subset optimization [99].

c. *A novel similarity scheme with Artificial Bee Colony (ABC):* In this innovative ensemble, a novel similarity-based approach joins forces with the collective intelligence of artificial bee colonies [100].

d. *MRMR with GA:* The Minimum Redundancy Maximum Relevance (MRMR) method pairs with genetic algorithms to create an efficient and effective feature selection partnership [89].

e. *Discrete Wavelet Transform (DWT) and modified genetic algorithm:* This modern ensemble leverages the signal processing capabilities of Discrete Wavelet Transform (DWT) alongside the adaptability of a modified genetic algorithm [101].

f. *Entropy and Signal-to-Noise Ratio (EnSNR):* In this innovative composition, the information theory concept of entropy joins forces with the Signal-to-Noise Ratio (SNR) to create a feature selection synergy [102].

g. *Large Margin Hybrid Algorithm for Feature Selection (LMFS):* A novel feature selection technique, the Large Margin Hybrid Algorithm for Feature Selection (LMFS), is introduced. LMFS employs a distance-based evaluation function and weighted bootstrapping to identify candidate feature subsets. These subsets are then assessed using a specific classifier and cross-validation to select the final features. LMFS was validated on six vibrational spectroscopic datasets with three classifiers, demonstrating its capacity to mitigate overfitting. It outperformed filter and wrapper methods, yielding features with superior classification performance and interpretability. LMFS also effectively managed computational time with varying classifier complexity, showing a preference for distance-based classifiers in the final feature subset selection [104].

h. *Hybrid of multiple filters and GA wrapper-based approach (MF-GARF):* A masterful blend of multiple filtering techniques and the power of a genetic algorithm wrapper delivers a compelling feature selection strategy [103].

These hybrid symphonies represent the cutting edge of feature selection, each offering a unique and harmonious fusion of methodologies to tackle the challenges of high-dimensional data analysis.

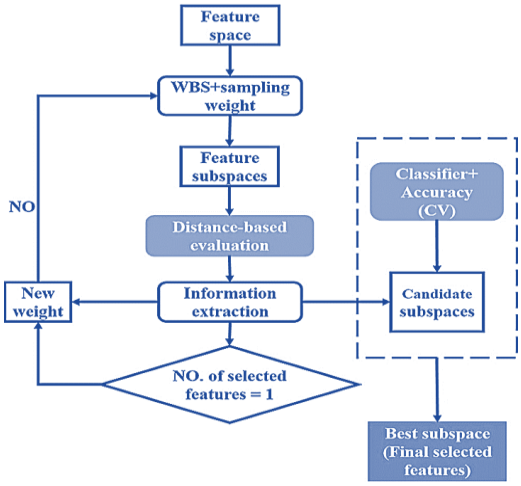

Fig.5. Large Margin Hybrid Algorithm for Feature Selection (LMFS)

6. Conclusion

In the culmination of our expedition through the intricate realm of feature selection for cancer prediction via Microarray gene profiles, a striking tapestry of progress unfolds. Beyond a mere reduction in feature dimensions, we immerse in the profound impact of this vital pre-processing tool. Its merits extend to taming the computational colossus, amplifying classification precision, and an array of consequential facets. The advent of DNA Microarray data has ushered in formidable machine learning challenges. The juxtaposition of numerous features with scarce samples poses a formidable foe. Researchers grapple not only with the enigma of countless features and meager samples but also contend with the intricacies of class imbalances, shifts between testing and training datasets, and the elusive presence of outliers.

In the face of these multifaceted trials, a procession of novel techniques unfurls year by year. Their goal transcends mere enhancements in classification precision, reaching towards the unraveling of the intricate nexus connecting gene expression to diseases. They serve as guiding lights, steering biologists on their quest for comprehension. In this odyssey, three techniques have commanded the spotlight: filter, wrapper, and embedded approaches. Filter methods, renowned for their computational efficiency, have been the stalwarts of choice. Meanwhile, the resource-intensive wrapper and embedded methods have been navigated with finesse. Yet, the crescendo of research reverberates with the harmonious cadence of hybrid feature selection methods. These hybrids, celebrated for their tenacious gene selection and the consequential augmentation of cancer classification accuracy, are harbingers of promise. Moreover, a groundswell of endeavors to amalgamate heterogeneous data sources, interweaving genomic, proteomic, clinical data, and beyond, signals a shift towards a holistic panorama of cancer classification.

This zenith signifies not just an evolution but a revolution in our quest to fathom and predict cancer via Microarray gene profiles. The symphony of feature selection serenades a promising era where the genome's secrets and its profound connection to disease are unveiled, note by note. These scholarly works collectively underscore the significance of gene expression analysis in biomedical and disease-related research. They delve into various techniques and methodologies, underscoring the importance of feature selection, dimensionality reduction, and machine learning methods in addressing the challenges posed by high-dimensional data. Cancer detection emerges as a prevalent application, with machine learning at the forefront. The advent of deep learning in identifying gene patterns associated with different cancer types signals the evolving landscape of cancer classification. Dimensionality reduction methods are crucially essential when addressing datasets with numerous dimensions, especially in the context of gene expression microarray data. The taxonomy of dimension reduction methods underscores the importance of benchmarking various approaches to determine their suitability for different data types. In the realm of survival data analysis, the benchmarking of 14 filter methods for feature selection in high-dimensional gene expression data reveals the standout performance of variance and carss filters, emphasizing the need for effective feature selection in biomedical applications.

Hybrid feature selection methods gain prominence in biomedical data processing. A novel three-stage gene selection approach, amalgamating various techniques for handling high-dimensional data efficiently, aims to enhance classifier accuracy while reducing computational complexity. The study on malaria detection exemplifies the application of machine learning to single-cell transcriptomics, showcasing the potential in biomarker and drug target identification. Cancer classification and the use of microarray data remain central themes. The anticipation of continued use and fusion of the Whale Optimization Algorithm (WOA) with genetic algorithms for gene selection suggests promising prospects. Lastly, the introduction of a novel feature selection method, WSNR, validated on benchmark problems, demonstrates its effectiveness, particularly in high-dimensional settings. Collectively, these investigations enrich the dynamic landscape of biomedical data analysis by providing valuable perspectives on navigating the complexities of high-dimensional data. They underscore the significance of dimensionality reduction, feature selection and machine learning in effectively confronting these challenges.

Refrences

[1]. American Cancer Society. (2017). Cancer facts and figures 2017 Atlanta.
[2]. WHO. (2018). Cancer facts sheet. WHO http://www.who.int/mediacentre/factsheets/fs297/ en/.

[3]. Doreswamy, & UmmeSalma, M. (2016). A binary bat inspired algorithm for the classification of breast cancer data. International Journal on Soft Computing Intelligence and Applications, 5(2/3), 1–21.
[4]. Selvaraj, S., & Natarajan, J. (2011). Microarray data analysis and mining tools. Bioinformation, 6 (3), 95.
[5]. Veerabhadrappa, & Rangarajan, L. (2010). Bi-level dimensionality reduction methods us- ing feature selection and feature extraction. International Journal of Computer Applica- tions, 4 (2), 33–38.
[6]. Bennet, J. , Ganaprakasam, C. , & Kumar, N. (2015). A hybrid approach for gene selection and classification using support vector machine. International Arab Journal of Information Technology, 12.
[7]. Alonso-Betanzos, A.; Bolón-Canedo, V.; Morán-Fernández, L.; Sánchez-Marono, N. A Review of Microarray Datasets: Where to Find Them and Specific Characteristics. Methods Mol. Biol. 2019, 1986, 65–85. _4. [CrossRef] [PubMed].
[8]. Bishop, C. Neural Networks for Pattern Recognition; Oxford University: Oxford, UK, 1995.
[9]. Hughes, G. On the mean accuracy of statistical pattern recognizers. IEEE Trans. Inf. Theory 1968, 14, 55–63. [CrossRef].
[10]. Garcia, S.; Luengo, J.; Saez, J.; Lopez, V.; Herrera, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Trans. Knowl. Data Eng. 2013, 25, 734–750. [CrossRef].
[11]. Duda, R.; Hart, P.; Stork, D. Pattern Classification, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2001.
[12]. Escolano, F.; Suau, P.; Bonev, B. Information Theory in Computer Vision and Pattern Recognition; Springer: Berlin/Heidelberg, Germany, 2009.
[13]. Alomari, O. A., et al. (2017). Mrmr ba: A hybrid gene selection algorithm for cancer classification. Journal of Theoretical and Applied Information Technology, 95 (12), 15.
[14]. Nogueira, A., Ferreira, A., & Figueiredo, M. (2022). A Machine Learning Pipeline for Cancer Detection on Microarray Data: The Role of Feature Discretization and Feature Selection.
[15]. Saqib, P., Qamar, U., Khan, R. A., & Aslam, A. (2020). MF-GARF: Hybridizing multiple filters and GA wrapper for feature selection of microarray cancer datasets. In Proceedings of the 2020 twenty-second international conference on advanced communication technology (ICACT) (pp. 517–524).
[16]. Saeid, M. M., Nossair, Z. B., & Saleh, M. A. (2020). A microarray cancer classification technique based on discrete wavelet transform for data reduction and genetic algorithm for feature selection. In Proceedings of the 2020 forth international conference on trends in electronics and informatics (ICOEI) (48184) (pp. 857–861).
[17]. Hambali, M., Saheed, Y., Oladele, T., & Gbolagade, M. (2019). ADABOOST ensemble algorithms for breast cancer classification. Journal of Advance Computer Research, 10 (2), 31–52. OnlineAvailable http://jacr.iausari.ac.ir/article_663924.html.
[18]. Liang, S., Ma, A., Yang, S., Wang, Y., & Ma, Q. (2018). A review of matched-pairs fea- ture selection methods for gene expression data analysis. Computational and Structural Biotechnology Journal, 16, 88–97. 10.1016/j.csbj.2018.02.005.
[19]. Veerabhadrappa, & Rangarajan, L. (2010). Bi-level dimensionality reduction methods using feature selection and feature extraction. International Journal of Computer Applications, 4 (2), 33–38.
[20]. Augenlicht, L. H., Wahrman, M. Z., Halsey, H., Anderson, L., Taylor, J., & Lip- kin, M. (1987). Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. Cancer Research, 47(22), 6017–6021.
[21]. Augenlicht, L. H., Taylor, J., Anderson, L., & Lipkin, M. (1991). Patterns of gene expression that characterize the colonic mucosa in patients at genetic risk for colonic cancer. Proceedings of the National Academy of Sciences of the United States of America, 88(8), 3286–3289.
[22]. Remeseiro López, B., & Bolon Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in Biology and Medicine, 112.
[23]. Rafii, F., Hassani, B. D. R., & Kbir, M. A. (2017). New approach for microarray data decision making with respect to multiple sources. In Proceedings of the second international conference on big data, cloud and applications (pp. 1–5).
[24]. Chen, C.-R., Shu, W.-Y., Tsai, M.-L., Cheng, W.-C., & Hsu, I. C. (2012). THEME: A web tool.

[25]. Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics, 2015.
[26]. Ventimiglia, G., & Petralia, S. (2013). Recent advances in DNA microarray technology: An overview on production strategies and detection methods. Bionanoscience, 3(4), 428–450.
[27]. Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Inf. Science (Ny)., 282, 111–135.
[28]. Acunzo, M., Romano, G., Wernicke, D., & Croce, C. M. (2015). MicroRNA and cancer – A brief overview. Advances in Biological Regulation, 57, 1–9.
[29]. Hammond, S. M. (2015). An overview of MicroRNAs. Advanced Drug Delivery Reviews.
[30]. de, M. M., Monobe, S., & da Silva, R. C. (2016). Gene expression: An overview of methods and applications for cancer research. Veterinária e Zootecnia, 23(4), 532–546.
[31]. Kong, C. S., Yu, J., Minion, F. C., & Rajan, K. (2011). Identification of biologically significant genes from combinatorial microarray data. ACS Combination Science, 13(5), 562–571.
[32]. Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset selection. Soft Computers, 12(2), 111–120.
[33]. Pihur, V., Datta, S., & Datta, S. (2008). Finding common genes in multiple cancer types through meta–analysis of microarray experiments: A rank aggregation approach. Ge- nomics, 92(6), 400–403.
[34]. Qi, Y., Sun, H., Sun, Q., & Pan, L. (2011). Ranking analysis for identifying differentially expressed genes. Genomics, 97(5), 326–329.
[35]. Kira, K., & Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. AAAI, 2, 129–134.
[36]. Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and in- formation gain criteria. Annals of Mathematics and Artificial Intelligence, 41(1), 77–93.
[37]. Golub, T. R., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science (80-.), 286(5439), 531–537.
[38]. Khodarev, N. N., et al. (2003). Receiver operating characteristic analysis: A general tool for DNA array data filtration and performance estimation. Genomics, 81(2), 202–209. Kim, J., et al. (2005). Identification of potential biomarkers of genotoxicity and carcinogenicity in L5178Y mouse lymphoma cells by cDNA microarray analysis. Environmental and Molecular Mutagenesis, 45(1), 80–89.
[39]. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226–1238.
[40]. He, X., Cai, D., & Niyogi, P. (2006). Laplacian score for feature selection. Advances in neural information processing systems (pp. 507–514).
[41]. Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of Bioinformatics and Computational Biology, 3(02), 185–205.
[42]. Huang, H.-L., & Chang, F.-L. (2007). ESVM: Evolutionary support vector machine for automatic feature selection and classification of microarray data. Biosystems, 90(2), 516–528.
[43]. Zheng, C.-H., Huang, D.-S., Kong, X.-Z., & Zhao, X.-M. (2008). Gene expression data classification using consensus independent component analysis. Genomics Proteomics Bioinformatics, 6(2), 74–82.
[44]. Zhou, N., & Wang, L. (2007). A modified T-test feature selection method and its application on the HapMap genotype data. Genomics Proteomics Bioinformatics, 5(3–4), 242–249.
[45]. Mohammadi, M., Noghabi, H. S., Hodtani, G. A., & Mashhadi, H. R. (2016). Robust and stable gene selection via maximum–minimum correntropy criterion. Genomics, 107(2–3), 83–87.
[46]. Cai, R., Hao, Z., Yang, X., & Wen, W. (2009). An efficient gene selection algorithm based on mutual information. Neurocomputing, 72(4–6), 991–999.
[47]. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine Learning, 46(1–3), 389–422.

[48]. Hall, M. A., & Smith, L. A. (1998). Practical feature subset selection for machine learning. Hambali, M. A., & Gbolagade, M. D. (2016). Ovarian cancer classification using hybrid synthetic minority over-sampling technique and neural network. Journal of Advance Computer Research, 7(4), 109–124.
[49]. Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the twentieth international conference on machine learning (ICML-03) (pp. 856–863).
[50]. Zhao, Z., & Liu, H. (2007). Searching for Interacting Features. In Proceedings of the twentieth international joint conference on artificial intelligence (pp. 1156–1161).
[51]. Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Machine learning proceedings 1992 (pp. 249–256). Elsevier.
[52]. Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In Proceedings of the European conference on machine learning (pp. 171–182).
[53]. Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. Artificial Intelligence, 151(1–2), 155–176.
[54]. Yu, H., Gu, G., Liu, H., Shen, J., & Zhao, J. (2009). A modified ant colony optimization algorithm for tumor marker gene selection. Genomics Proteomics Bioinformatics, 7(4), 200–208.
[55]. Xiong, W., Cai, Z., & Ma, J. (2008). A DSRPCL-SVM approach to informative gene analysis.
[56]. Garro, B. A., Rodríguez, K., & Vázquez, R. A. (2016). Classification of DNA microarrays using artificial neural networks and ABC algorithm. Applied Soft Computing, 38, 548–560. Geman, O., Chiuchisan, I., Covasa, M., Doloc, C., Milici, M.-R., & Milici, L.-D. (2016). Deep learning tools for human microbiome big data. In Proceedings of international workshop soft computing applications (pp. 265–275).
[57]. Hancer, E., Xue, B., Karaboga, D., & Zhang, M. (2015). A binary ABC algorithm based on advanced similarity scheme for feature selection. Applied Soft Computing, 36, 334–348. Hardin, J., & Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis, 44(4), 625–638.
[58]. Zhao, Z., Wang, L., & Liu, H. (2010). Efficient spectral feature selection with minimum redundancy. In Proceedings of the twenty-fourth AAAI conference on artificial intelligence. Zhao, Z., Wang, L., Liu, H., & Ye, J. (2011). On similarity preserving feature selection.
[59]. Kar, S., Das Sharma, K., & Maitra, M. (2015). Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications, 42(1), 612–627.
[60]. Vukusic, I., Grellscheid, S. N., & Wiehe, T. (2007). Applying genetic programming to the prediction of alternative mRNA splice variants. Genomics, 89(4), 471–479.
[61]. Inza, I., Sierra, B., Blanco, R., & Larrañaga, P. (2002). Gene selection by sequential search wrapper approaches in microarray cancer class prediction. Journal of Intelligent & Fuzzy Systems, 12(1), 25–33.
[62]. Sharma, A., Imoto, S., & Miyano, S. (2011). A top-r feature selection algorithm for microarray gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(3), 754–764.
[63]. Wanderley, M. F. B., GardeuX, V., Natowicz, R., & de Pádua Braga, A. (2013). GA-KDE-Bayes: An evolutionary wrapper method based on non-parametric density estimation applied to bioinformatics problems. ESANN.
[64]. Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. Inf. Sci. (Ny)., 181(1), 115–128.
[65]. Anaissi, A., Kennedy, P. J., & Goyal, M. (2011). Feature selection of imbalanced gene expression microarray data. In Proceedings of the 2011 twelfth ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (pp. 73–78).
[66]. Shah, M., Marchand, M., & Corbeil, J. (2011). Feature selection with conjunctions of decision stumps and learning from microarray data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(1), 174–186.

[67]. Canul-Reich, J., Hall, L. O., Goldgof, D. B., Korecki, J. N., & Eschrich, S. (2012). Iterative feature perturbation as a gene selector for microarray data. International Journal of Pattern Recognition and Artificial Intelligence, 26(05), Article 1260003.
[68]. Nie, F., Huang, H., Cai, X., & Ding, C. H. (2010). Efficient and robust feature selection via joint ℓ2, 1-norms minimization. In Advances in neural information processing systems (pp. 1813–1821).
[69]. Zhang, R., Nie, F., & Li, X. (2018). Feature selection under regularized orthogonal least square regression with optimal scaling. Neurocomputing, 273, 547–553.
[70]. Zhang, R., Nie, F., Li, X., & Wei, X. (2019). Feature selection with multi-view data: A survey. Information Fusion, 50, 158–167.
[71]. Wang, G., Song, Q., Xu, B., & Zhou, Y. (2013). Selecting feature subset for high dimensional data via the propositional FOIL rules. Pattern Recognition, 46(1), 199–214.
[72]. Fukunaga, K. (2013). Introduction to statistical pattern recognition. Elsevier.
[73]. Li, H., Jiang, T., & Zhang, K. (2004). Efficient and robust feature extraction by maximum margin criterion. Advances in neural information processing systems (pp. 97–104).
[74]. Luo, Y., Tao, D., Xu, C., Li, D., & Xu, C. (2013). Vector-valued multi-view semi-supervised learning for multi-label image classification. In Proceedings of the twenty-seventh AAAI conference on artificial intelligence.
[75]. Yang, Y., Shen, H. T., Ma, Z., Huang, Z., & Zhou, X. (2011). L2, 1-norm regularized dis- criminative feature selection for unsupervised. In Proceedings of the twenty-second international joint conference on artificial intelligence.
[76]. Zhao, Z., Wang, L., Liu, H., & Ye, J. (2011). On similarity preserving feature selection.
[77]. Wang, S., Tang, J., & Liu, H. (2015). Embedded unsupervised feature selection. In Proceedings of the twenty-ninth AAAI conference on artificial intelligence.
[78]. Cai, D., Zhang, C., & He, X. (2010). Unsupervised feature selection for multi-cluster data. In Proceedings of the sixteenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 333–342).
[79]. Hou, C., Nie, F., Li, X., Yi, D., & Wu, Y. (2013). Joint embedding learning and sparse regression: A framework for unsupervised feature selection. IEEE Transaction on Cybernetics., 44(6), 793–804.
[80]. Krzanowski, W. J. (1987). Selection of variables to preserve multivariate data structure, using principal components. Journal of the Royal Statistical Society, Series C (Applied Statistics), 36(1), 22–33.
[81]. Cai, D., He, X., & Han, J. (2007). Semi-supervised discriminant analysis. In Proceedings of the eleventh international conference on ICCV (pp. 1–7).
[82]. Liu, Y., Nie, F., Wu, J., & Chen, L. (2013). Efficient semi-supervised feature selection with noise insensitive trace ratio criterion. Neurocomputing, 105, 12–18.
[83]. Chang, X., Nie, F., Yang, Y., & Huang, H. (2014). A convex formulation for semi-supervised multi-label feature selection. In Proceedings of the twenty-eighth AAAI conference on artificial intelligence.
[84]. Wang, L., Chu, F., & Xie, W. (2007). Accurate cancer classification using expressions of very few genes. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(1), 40–53.
[85]. Rangasamy, M. (2009). An efficient statistical model based classification algorithm for classifying cancer gene expression data with minimal gene subsets. International Jour- nal of Cyber Society and Education, 2(2), 51–66.
[86]. Martín-Merino, M., & De Las Rivas, J. (2009). Improving k-nn for human cancer classifi- cation using the gene expression profiles. In Proceedings of the international symposium on intelligent data analysis (pp. 107–118).
[87]. Revathy, N., & Amalraj, R. (2011). Accurate cancer classification using expressions of very few genes. International Journal on Computer Applications, 14(4), 19–22.
[88]. Ghorai, S., Mukherjee, A., Sengupta, S., & Dutta, P. K. (2010). Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(3), 659–671.

[89]. El Akadi, A., Amine, A., El Ouardighi, A., & Aboutajdine, D. (2011). A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowledge and Information System, 26(3), 487–500.
[90]. Rajeswari, P., & Reena, G. S. (2011). Human liver cancer classification using microarray gene expression data. International Journal of Computer Applications, 34(6), 25–37.
[91]. Sahu, B., & Mishra, D. (2012). A novel feature selection algorithm using particle swarm optimization for cancer microarray data. Procedia Engineering, 38, 27–31.
[92]. Swathi, S., Babu, G. A., Sendhilkumar, R., & Bhukya, S. N. (2012). Performance of ART1 network in the detection of breast cancer. In Proceedings of international conference on computer design and engineering (ICCDE 2012): 49 (pp. 100–105).
[93]. Dev, J., Dash, S., Dash, S., & Swain, M. (2012). A classification technique for microarray gene expression data using PSO-FLANN. International Journal on Computer Science and Engineering, 4(9), 1534–1535.
[94]. Abeer, M. M., Basma, A. M., El-Sayed, M. E., & Abdel-Badeeh, M. S. (2013). Applying a statistical technique for the discovery of differentially expressed genes in microarray data, pp. 220–227.
[95]. Shreem, S. S., Abdullah, S., & Nazri, M. Z. A. (2014). Hybridizing harmony search with a Markov blanket for gene selection problems. Information Science (Ny)., 258, 108–121.
[96]. Abeer, M. M., & Basma, A. M. (2014). A hybrid reduction approach for enhancing cancer classification of microarray data. International Journal of Advance Research in Artificial Intelligence, 3(10).
[97]. Alshamlan, H. M., Badr, G. H., & Alohali, Y. A. (2015). Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational Biology and Chemistry, 56, 49–60.
[98]. Lee, C.-P., & Leu, Y. (2011). A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing, 11(1), 208–213.
[99]. Zibakhsh, A., & Abadeh, M. S. (2013). Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function. Engineering Applications of Artificial Intelligence, 26(4), 1274–1281.
[100]. Hancer, E., Xue, B., Karaboga, D., & Zhang, M. (2015). A binary ABC algorithm based on advanced similarity scheme for feature selection. Applied Soft Computing, 36, 334–348. Hardin, J., & Rocke, D. M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis, 44(4), 625–638.
[101]. Saeid, M. M., Nossair, Z. B., & Saleh, M. A. (2020). A microarray cancer classification technique based on discrete wavelet transform for data reduction and genetic algorithm for feature selection. In Proceedings of the 2020 forth international conference on trends in electronics and informatics (ICOEI)(48184) (pp. 857–861).
[102]. Hengpraprohm, S., & Jungjit, S. (2020). Ensemble feature selection for breast cancer classification using microarray data. Inteligencia Artificial, 23(65), 100–114.
[103]. Saqib, P., Qamar, U., Khan, R. A., & Aslam, A. (2020). MF-GARF: Hybridizing multiple filters and GA wrapper for feature selection of microarray cancer datasets. In Proceedings of the 2020 twenty-second international conference on advanced communication technology (ICACT) (pp. 517–524).
[104]. Zhang, J., Xiong, Y., & Min, S. (2019). A new hybrid filter/wrapper algorithm for feature selection in classification. Analytica Chimica Acta, 1080, 43–54. https://doi.org/10.1016/j.aca.2019.06.054
[105]. World Cancer Day 2023: Close the care gap. (n.d.). PAHO/WHO | Pan American Health Organization. https://www.paho.org/en/campaigns/world-cancer-day-2023-close-care-gap.
[106]. Banerjee, A., Bandyopadhyay, A., Ghosh, S., & Bandyopadhyay, A. (2023). Revolutionizing Oncology: Cutting-edge Classification Methods for Microarray Data. Journal of Harbin Engineering University, 44(10), 1199-1217.
[107]. Bhaumik, L., Bandyopadhyay, A., & Bandyopadhyay, A. (2023). Exploring the depths: Harnessing the Power of Deep Learning in Marine Ecology. Journal of Survey of Fisheries Sciences, 10(1), 3779-3796.