Integration of Random Forest and Hough Transform for Cancer Classification Using Microarray Gene Expression Data

Hibah Alatawi*, Hechmi Shili
Department of Computer Sciences, University of Tabuk, Saudi Arabia

***Abstract*—Cancer classification poses a significant challenge owing to the intricate nature and diversity of the disease. This study introduces a novel methodology for cancer classification leveraging microarray gene expression data. The proposed approach integrates Random Forest (RF) and Hough Transform (HT), where RF performs feature selection and classification, and HT identifies sub-biclusters that are merged into larger clusters using a hypergraph model. Evaluation on multiple cancer datasets demonstrates that the hybrid approach improves classification accuracy compared to standalone RF or HT while capturing meaningful gene expression patterns.**

***Keywords*—Cancer classification; microarray gene expression; random forest; Hough transform; feature selection; hypergraph model**

I. Introduction

Microarray data analysis plays a pivotal role in understanding complex biological systems and identifying biomarkers associated with various diseases, including cancer [1,10]. Microarray analysis is employed in gene expression profiling, where it typically generates a substantial amount of data [31,32]. Biclustering algorithms have emerged as a powerful tool in the analysis of microarray gene expression data, allowing for the simultaneous classification of both genes and experimental conditions. Unlike standard clustering techniques, which focus on grouping either genes or samples, biclustering methods aim to identify sub-matrices within the gene expression data matrix where a subset of genes exhibit a consistent pattern across a subset of experimental conditions [1]. In recent years, the classification of microarray datasets, often referred to as large-scale biological data analysis, has become a prominent and compelling area of research. The application of microarray technology represents a critical advancement in molecular biology, particularly for the detection of cancer [37]. In Addition, biclustering, a powerful data mining technique, facilitates the simultaneous clustering of rows and columns in large datasets, uncovering hidden patterns and structures. However, traditional biclustering methods often struggle with the high dimensionality and noise inherent in microarray datasets [2]. Microarray technology has become a powerful tool for investigating complex biological processes and diseases, enabling the representation of gene expression data as large matrices with genes as rows and experimental conditions as columns [3]. While various statistical techniques have been applied to analyze these gene expression matrices, biclustering has emerged as a particularly effective method, as it can uncover meaningful patterns by identifying subsets of genes that exhibit consistent expression profiles across subsets of experimental conditions. This ability to capture the inherent complexity of gene regulation, where the expression of a gene can be influenced by multiple factors, is a key advantage of biclustering over traditional clustering approaches, which focus on grouping genes or samples independently. The insights gained from biclustering analysis can provide valuable perspectives on the intricate mechanisms governing biological processes and disease states [4].

Despite these advances, existing biclustering and classification approaches still face limitations in handling the high dimensionality, noise sensitivity, and the difficulty of linking biclustering outputs directly to robust classification performance. This gap motivates the need for an integrated and more resilient framework for microarray data analysis.

An automated, computer-assisted medical diagnosis system that integrates cutting-edge medical methodologies with advanced machine learning algorithms represents a critical multidisciplinary technology. It enables accurate and noninvasive diagnosis of various diseases, including breast cancer [6,32]. The classification of microarray data presents significant challenges due to its complexity. The bioinformatics community employs a variety of approaches to diagnose and classify this data using machine learning systems. Given that each microarray sample contains thousands of genes for analysis, feature elimination methods have become widely utilized and adopted to manage the high dimensionality and enhance the accuracy of classification outcomes [5,39]. The abundance of gene expression datasets offers a valuable opportunity to computationally identify condition-specific functional gene modules (FGMs). These FGMs are characterized by highly organized expression patterns within specific sets of genes. Smart healthcare involves a wide range of operatives, including physicians, nurses, hospitals, and research organizations [30].

This study aims to develop an integrated framework that combines geometric pattern detection and ensemble learning to improve both biclustering quality and classification accuracy in microarray gene expression data.

In this study, we propose a novel methodology that addresses these challenges by integrating the Hough Transform and Random Forest algorithm for biclustering and classification tasks. The Hough Transform, applied in column pair-spaces, enables the identification of sub-biclusters within microarray

*Corresponding author.

datasets, while hyper-graph partitioning techniques refine and optimize these sub-biclusters. Subsequently, the optimized biclusters are seamlessly integrated as inputs into a Random Forest classifier, enhancing the classification process. By combining these techniques, our integrated framework offers a comprehensive solution for biclustering and classification tasks in Microarray data analysis. Leveraging the complementary strengths of the Hough Transform and Random Forest algorithm, our approach provides a holistic and efficient method for uncovering meaningful patterns and biomarkers in complex biological datasets.

The main contributions of this work are summarized as follows:

- Proposing a novel integration between the Hough Transform and Random Forest for simultaneous biclustering and classification.
- Introducing a hyper-graph partitioning refinement step to optimize detected sub-biclusters.
- Providing a robust framework that links biclustering outputs directly to classification performance.
- Demonstrating the effectiveness of the proposed approach on real-world microarray cancer datasets.

The integration of biclustering and classification offers several advantages. First, it enables the simultaneous identification of coherent gene expression patterns and their classification into distinct biological states or phenotypes. Second, by incorporating the Hough Transform, our framework provides robustness to noise and captures complex patterns inherent in Microarray gene expression data. Third, the Random Forest classifier enhances the accuracy and robustness of classification by leveraging the collective decision of multiple decision trees. The proposed framework holds great promise for advancing our understanding of gene expression regulation and its role in disease mechanisms. By providing a holistic approach to Microarray gene expression data analysis, it offers researchers a powerful tool for uncovering meaningful insights and identifying potential biomarkers associated with diseases such as cancer.

Unlike existing studies that treat biclustering and classification as separate processes, the proposed approach establishes a direct and optimized connection between both tasks, highlighting the novelty of this work compared to recent related studies.

In the subsequent sections of this study, we describe the methodology in detail, present experimental results on real-world datasets, and discuss the implications of our findings for biomedical research and clinical applications.

II. Related Work

The field of gene expression analysis has witnessed substantial advancements, particularly in the development and application of biclustering techniques. Biclustering, a method that allows for the simultaneous classification of gene expression profiles, has proven to be effective in identifying patterns of gene expression under diverse conditions. This literature review examines the evolution of biclustering methods and their application in microarray gene expression data analysis. Recent studies (Table I) have introduced innovative frameworks and algorithms that enhance the classification of gene expression profiles, including the use of Biclustering-based classifiers, evolutionary approaches, and ensemble learning techniques. Additionally, the integration of machine learning and big data approaches has further refined cancer diagnosis and classification, demonstrating significant improvements in accuracy and robustness. This section also highlights the potential of combining traditional statistical methods with advanced computational techniques, such as the Hough Transform, to improve the analysis and interpretation of complex biological data. Through a comprehensive examination of these advancements, the review underscores the critical role of biclustering and related methodologies in advancing our understanding of gene expression patterns and improving cancer classification and diagnosis.

A. Advances in Biclustering and Machine Learning for Gene Expression and Cancer Diagnosis

Biclustering, a method for simultaneous classification of gene expression profiles, has been shown to be effective in identifying gene expression patterns under various conditions [1]. This method has been further enhanced by the development of a biclustering-based classification framework, which utilizes homogeneously expressed genes in biclusters to construct a classifier for sample class membership prediction Malhotra et al. [2]. An evolutionary approach to obtaining high-quality biclusters of highly-correlated genes has also been proposed, demonstrating competitive performance with state-of-the-art algorithms Ayadi et al. [3]. Additionally, the application of the bagging approach to improve the performance of biclustering methods has been successful in both synthetic and real datasets Hanczar, Ayadi et al. [4]. These studies collectively highlight the potential of biclustering and its various applications in microarray gene expression data analysis. The proposed framework of Biswal et al. [7] undergoes testing on a diverse set of datasets, including five distinct microarray datasets and one dataset consisting of single-cell RNA sequencing (scRNA-seq) data. In Albalawi et al. [26], the Convolutional Neural Network (CNN) classifier for diagnosing breast cancer utilizing MIAS (Mammographic Image Analysis Society) dataset. CNN established as an efficient class of methods for image recognition problems. The research of Almutairi, S. et al. [27], develops a big data-driven breast cancer classification model using Deep Reinforcement Learning (DRL), achieving high accuracy across three datasets (WBCD: 98.90%, WDBC: 99.02%, WPBC: 98.88%). The model incorporates the Gorilla Troops Optimization (GTO) algorithm for feature selection and Deep Q Learning (DQL) for classification, outperforming traditional methods such as RBF-ELB, PSO-MLP, and GA-MLP. The study demonstrates the potential of integrating big data and advanced machine learning techniques to improve breast cancer diagnosis accuracy. The study of Alatawi, Y. M. et al. [28] assessed early performance indicators for breast cancer screening at King Abdulaziz University Hospital, Saudi Arabia, analyzing data from 1,911 women who participated in the screening program between 2012 and 2019. The findings revealed that 19.9% of women were recalled for further evaluation, with 18.9% undergoing biopsy. Screen-detected

cancer was found in 1.6% of participants, while 0.7% were diagnosed with breast cancer.

Balaha and Hassan [33] proposed a deep learning-based framework for skin cancer diagnosis by integrating transfer learning with the Sparrow Search Algorithm (SSA) to optimize feature extraction and classification performance. Their approach demonstrated superior results compared with several conventional deep learning architectures, achieving high accuracy, precision, recall, and F1-score in skin lesion classification. Since the Hough transform is an image analysis technique, it can potentially be integrated with other computational platforms and machine learning models to further enhance image-based diagnostic performance [34].

B. Computational Models for Microarray-Based Cancer Classification

Recent advances in biomedical data analysis have highlighted the potential of image-derived mathematical models, such as the Hough Transform (HT), in uncovering hidden structural relationships within complex biological datasets. Originally developed for geometric feature extraction in image processing, HT has been successfully adapted to identify spatial and relational patterns in biomedical contexts, including histopathological imaging, brain signal analysis, and genomic data interpretation. In particular, the HT framework enables the detection of geometric trends—such as alignments or co-expressed trajectories—among gene pairs or condition profiles. This geometric abstraction provides an effective means of representing high-dimensional biological data in lower-dimensional spaces, facilitating the discovery of regulatory modules and functional relationships that traditional statistical methods may overlook [1]. Beyond geometric modeling, several computational frameworks have explored the synergy between biclustering and feature selection to improve cancer classification using microarray gene expression data. Biclustering methods—such as the Cheng & Church algorithm, which minimizes mean squared residue to identify locally coherent expression patterns; xMotif, which searches for statistically significant submatrices enriched with co-expressed genes; and the Plaid model, which decomposes the expression matrix into overlapping layers representing distinct biological processes—have been widely employed to capture coherent gene subsets that exhibit correlated expression under specific experimental conditions [40]. Feature selection algorithms, including mutual information, ReliefF, and ensemble-based approaches, have been integrated with machine learning classifiers to reduce dimensionality and enhance interpretability. Among these, the Random Forest (RF) classifier has emerged as a preferred model due to its robustness to noise, built-in feature importance estimation, and ability to handle high-dimensional gene expression data efficiently. Recent studies have demonstrated that RF-based hybrid systems outperform conventional models in identifying diagnostic biomarkers and distinguishing between cancer subtypes, motivating the integration of RF with biclustering and structural learning in this study.

C. Advances in Gene Expression Analysis

The Hough Transform has been successfully applied in the analysis of microarray gene expression data. Zhao et al [8] and Zhao et al. [9] both developed biclustering algorithms based on the Hough Transform in the column-pair space, which were found to be robust to noise and computationally efficient. Tiňo et al. [10] further improved this approach by introducing a probabilistic model-based Hough Transform for the detection of co-expression patterns in three-color cDNA microarray data. Shafie et al. [11] extended this work by demonstrating the use of a relational space to classify microarray gene expression data, with the transformation of real-valued data to binary data leading to improved class separation. These studies collectively highlight the potential of the Hough Transform in the analysis of microarray gene expression data. Over the past three decades, advancements in cancer classification have been significant; however, a universal methodology for the identification of new cancer classes (class discovery) and the assignment of tumors to established classes (class prediction) remains elusive [12]. Golub et al. [12,16] presented a comprehensive approach to cancer classification utilizing gene expression monitoring via DNA microarrays, specifically applied to human acute leukemias as an illustrative case. The study of Umer, M. et al. [29] introduces an ensemble learning-based voting classifier that integrates logistic regression and stochastic gradient descent classifiers with deep convoluted features for the accurate detection of breast cancer. The proposed framework effectively distinguishes between malignant and benign tumors, demonstrating improved classification accuracy. The results indicate that this method outperforms existing state-of-the-art approaches. The class discovery process successfully differentiated between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without prior knowledge of these categories. Furthermore, a class predictor derived from this process accurately classified new leukemia cases. These findings validate the viability of cancer classification based exclusively on gene expression profiles and propose a generalizable strategy for the discovery and prediction of cancer classes across various cancer types, independent of pre-existing biological knowledge. A hybrid approach of PCC-BPSO/GA with multiple classifiers improves gene selection and classification accuracy in microarray datasets [17]. Differentiating malignant pleural mesothelioma (MPM) from adenocarcinoma (ADCA) of the lung is challenging with conventional methods. G. Gordon et al. [13,15] proposed a gene expression ratio-based approach for early and precise diagnosis.

This study assessed 181 tissue samples (31 MPM and 150 ADCA) and identified 15 diagnostic ratios from a training set of 32 samples. Each ratio achieved at least 90% accuracy in diagnosis, with combined ratios reaching 95% and 99% accuracy for MPM and ADCA, respectively. This method is accurate, cost-effective, and has potential applicability in other clinical diagnostic contexts.

Extensive experiments in Purbolaksono et al. [16] on 8 real cancer microarray datasets (4 diagnostic and 4 prognostic) show that the proposed classifier performed superior in both cancer diagnosis and prognosis, the latter of which was regarded as quite difficult previously. Additionally, results demonstrate that sample classification accuracy can serve as a good subjective quality measure for different types of biclusters, and hence as a tool to extrinsically evaluate the performance of various biclustering algorithms that produce those biclusters.

Microarray data classification poses significant challenges due to the high dimensionality and small sample sizes of the data.

V. Bolon et al. [18] present an extensive experimental evaluation of the most prominent feature selection algorithms and evaluation techniques, providing a detailed analysis of the findings and insights that can inform the development of effective microarray data classification approaches. Embryonal tumors of the central nervous system (CNS) represent a biologically diverse and poorly understood group of tumors. Diagnosing these tumors based solely on their morphology is controversial, as exemplified by medulloblastomas—the most common malignant brain tumors in children—whose pathogenesis and relationship to other embryonal CNS tumors remain debated. Additionally, predicting patients' responses to therapy for these tumors poses a significant challenge. Pomeroy SL et al. [19] addressed these issues by developing a classification system based on DNA microarray gene expression data from 99 patient samples. They demonstrated that medulloblastomas are molecularly distinct from other brain tumors, including primitive neuroectodermal tumors (PNETs), atypical teratoid/rhabdoid tumors (AT/RTs), and malignant gliomas. Their research provided previously unrecognized evidence that medulloblastomas derive from cerebellar granule cells through activation of the Sonic Hedgehog (SHH) pathway. Furthermore, they showed that the clinical outcomes of children with medulloblastomas are highly predictable based on the gene expression profiles of their tumors at diagnosis. Jinthanasatian et al. [20] proposed a new feature selection algorithm called FF-SVM that uses the FireFly algorithm along with an SVM classifier to classify cancer microarray gene expression data, and they found that this algorithm achieves high classification accuracy using a small number of selected genes. Salem et al. [21] present a novel methodology for classifying human cancer diseases based on gene expression profiles. This approach combines Information Gain (IG) and Deep Genetic Algorithm (DGA).

TABLE I. SUMMARY OF STUDIES ON BI-CLUSTERING TECHNIQUES IN MICROARRAY GENE EXPRESSION DATA ANALYSIS [15]

Ref	Dataset	Type classifier	Feature Selection	Accuracy	No.Genes
[16]	Leukemia [12]	BN	Ml	88.2%	NA
		NN	FCBF	99.44%	51
		Decision Table	CFS	98.6%	79
[17]	Lung Cancer [13]	NB	Ml	98.66%	NA
		NN	SFFS	97.92%	308
		NB	PCC-BPSO	98.03%	39
		RF	K-means - Relief	98.90%	NA
[18]	Prostate [14]	SVM	ReliefF	97%	50
		ELM	MIMAGA	97.12%	60
		RF	K-means -Relief	88.97%	NA
[20]	DLBCL	rule set generation	a neuro-fuzzy algorithm	83.81%	13
[21]	DLBCL	GP	IG-SGA	94.80%	110

III. PROPOSED FRAMEWORK

Microarray data analysis plays a pivotal role in understanding complex biological systems and identifying biomarkers associated with various diseases, including cancer. Biclustering, a powerful data mining technique, facilitates the simultaneous clustering of rows and columns in large datasets, uncovering hidden patterns and structures. However, traditional biclustering methods often struggle with the high dimensionality and noise inherent in microarray datasets. In this study, we propose a novel methodology that addresses these challenges by integrating the Hough Transform and Random Forest algorithm for biclustering and classification tasks. The Hough Transform, applied in column pair-spaces, enables the identification of sub-biclusters within microarray datasets, while hypergraph partitioning techniques refine and optimize these sub-biclusters. Subsequently, the optimized biclusters are seamlessly integrated as inputs into a Random Forest classifier, enhancing the classification process.

Fig. 1 summarizes the full workflow of the proposed framework, from the input microarray matrix to the final classification result. The process starts by projecting gene expression data into column-pair spaces, followed by Hough Transform–based detection of geometric patterns to extract sub-biclusters. These are encoded, grouped by shared columns, and used to build a hypergraph that is partitioned and optimized to obtain refined biclusters. Features are then derived from the optimized biclusters and fed into a Random Forest classifier for cancer prediction. The next subsections explain each stage step by step.

A. Geometric Patterns in Microarray Data Using Hough Transform in Column-Pair Spaces

Microarray gene expression data often exhibit complex geometric patterns that encode valuable information about underlying biological processes. The application of the Hough Transform (HT) in column-pair spaces offers a powerful tool for uncovering these geometric patterns and extracting meaningful insights from the data. In microarray experiments, gene expression levels are typically measured across multiple conditions (columns) for numerous genes (rows). Traditional clustering methods often overlook the spatial relationships between genes and conditions, limiting their ability to capture

the intricate geometric structures present in the data. However, by leveraging the HT in column-pair spaces, we can explore these geometric patterns more effectively. The HT is a robust technique originally developed for line detection in image processing. When applied to microarray data, it allows us to identify linear structures or patterns within the dataset, revealing relationships between genes and conditions that may not be apparent through conventional methods. In column-pair spaces, each point represents the expression levels of a gene across two conditions. By transforming the data into this space, we can detect linear patterns corresponding to coherent gene expression changes across specific conditions. These patterns may manifest as diagonal lines, indicating consistent up- or down-regulation of genes under particular conditions, or as clusters of points representing subsets of genes exhibiting similar expression profiles across conditions. The identification of geometric patterns using HT in column-pair spaces provides valuable insights into the underlying biological processes driving gene expression changes. For example, coherent diagonal lines may indicate the activation or repression of specific pathways or regulatory networks in response to different experimental conditions. Clusters of points may represent co-regulated gene sets involved in common biological functions or pathways. Moreover, the HT can be combined with biclustering algorithms to further elucidate complex geometric structures within microarray data. Biclustering identifies subsets of genes and conditions that exhibit coherent expression patterns, and the HT enhances this analysis by revealing the geometric relationships between these subsets. The application of the Hough Transform in column-pair spaces offers a powerful approach for uncovering geometric patterns in microarray gene expression data. By leveraging these patterns, researchers can gain deeper insights into the underlying biological processes and identify potential biomarkers or therapeutic targets for further investigation. By combining these techniques, our integrated framework offers a comprehensive solution for biclustering and classification tasks in Microarray data analysis. Leveraging the complementary strengths of the Hough Transform and Random Forest algorithm, our approach provides a holistic and efficient method for uncovering meaningful patterns and biomarkers in complex biological datasets. We consider a dataset represented by an m by n matrix, where each column, denoted as a_i, represents a specific attribute or condition. In our methodology, the five coherent patterns identified within microarray datasets are integral to the analysis of gene expression data, as they allow us to extract biologically meaningful relationships between genes and conditions. Each pattern reveals different forms of coherence in the data, which are crucial for both biclustering and classification. Our interest lies in identifying five distinct coherent patterns within this matrix: 1) constant values across the entire pattern, 2) constant values within columns, 3) constant values within rows, 4) additive coherent values, and 5) multiplicative coherent values. These patterns can be expressed through specific equations corresponding to the behavior of the data within each pattern. Previous research has explored the identification of these patterns using hyperplane-based approaches with the Hough Transform (HT). However, as the dimensionality of the data increases, so does the computational complexity. In this study, we propose a novel approach utilizing the HT in column pair-spaces to identify sub-biclusters within the microarray cancer dataset. These sub-biclusters are then optimized using hypergraph partitioning techniques to enhance the biclustering process, offering a more efficient and effective solution to pattern discovery in high-dimensional datasets.

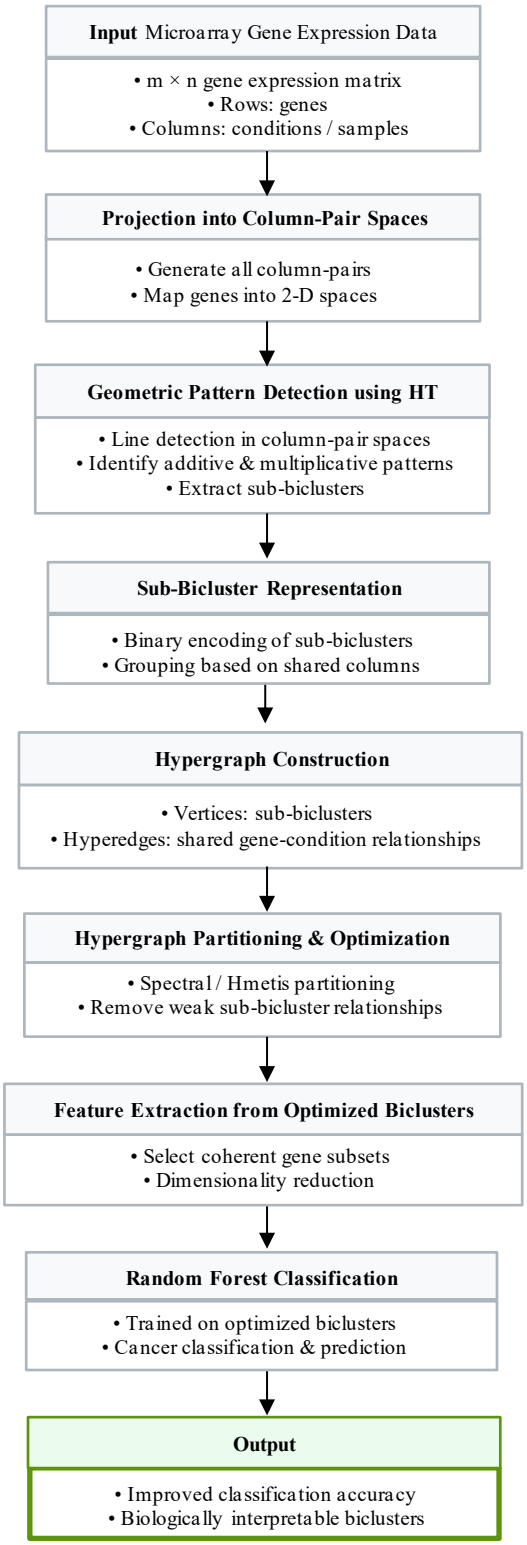

Fig. 1. Overall framework of the proposed method.

B. Characterizing Coherent Expression Patterns in Microarray Data: Across Genes and Conditions

In the analysis of microarray data, identifying coherent patterns is essential for understanding the relationships between genes and conditions. These patterns provide insights into how gene expression varies across different experimental conditions or biological states. By recognizing distinct types of coherence in the data, such as constant, additive, and multiplicative relationships, we can better interpret the underlying biological mechanisms driving gene regulation. Each pattern reveals a unique aspect of gene expression behavior, offering valuable information for biclustering and classification tasks. The following section outlines how each identified pattern is applied in the context of microarray data analysis to uncover meaningful biological insights. To formalize the identification of coherent gene expression patterns in microarray data, we define mathematical equations corresponding to each of the five distinct biclustering patterns. These equations capture the specific relationships observed between genes and conditions, allowing us to model the behavior of the data in a structured and interpretable manner. By expressing these patterns mathematically, we can apply computational techniques, such as the Hough Transform, to efficiently detect and analyze these patterns across large, high-dimensional datasets. The following section details the equations associated with each biclustering pattern, providing a foundational framework for pattern discovery in microarray data.

1) Constant value in the entire pattern: This pattern reflects a subset of genes that exhibit a uniform expression level across all conditions. In microarray data, this could signify genes that are consistently active or repressed, regardless of experimental conditions. These genes might be involved in essential housekeeping functions, and their identification can help distinguish between condition-specific and general expression patterns.

$$x_{ij} = \mu \tag{1}$$

In the context of microarray gene expression data, it represents the expression level of a gene i under condition. i refers to the row index, corresponding to a specific gene. j refers to the column index corresponding to a specific condition or sample. x_{ij} denotes the expression level of gene i under condition j in the microarray dataset. μ. refers to the mean value of the entire pattern.

2) Constant values in columns: This pattern represents genes that maintain a consistent expression level within a specific condition or set of conditions. It helps to identify condition-invariant behavior, indicating genes that are robust to particular environmental or experimental variations. In biomedical research, such patterns might be used to highlight genes involved in fundamental processes that are stable under certain biological states.

$$x_{ij} = \mu_i \tag{2}$$

μ refers to the mean value of column i.

3) Constant values within rows: This pattern captures genes that have consistent expression across specific genes under varying conditions. In microarray data, it could represent genes that are co-regulated across multiple conditions, providing insights into condition-specific gene regulation networks. This is particularly useful for identifying condition-dependent gene sets that may play roles in disease mechanisms or therapeutic responses.

$$x_{ij} = \mu_j \tag{3}$$

μ refers to the mean value of row j.

4) Additive coherent values: This pattern reflects gene expression changes that follow an additive relationship across conditions. For instance, gene expression levels might increase or decrease consistently by a fixed amount across conditions, suggesting the additive effects of external stimuli or experimental treatments. Identifying additive patterns in microarray data can help detect linear regulatory mechanisms and dose-response relationships in biological pathways.

$$x_{ij} = \alpha_i + \beta_j \tag{4}$$

α_i : Constant additive value f for column i.

β_j : Constant additive value f for row j.

5) Multiplicative coherent values: This pattern represents relationships where gene expression changes by a consistent multiplicative factor across conditions. It indicates proportional regulation across conditions, which may reflect more complex, nonlinear regulatory mechanisms. Multiplicative patterns can point to the influence of transcription factors or other regulatory elements that affect gene expression in a multiplicative fashion across different conditions.

$$x_{ij} = \alpha_i \times \beta_j \tag{5}$$

α_i : Multiplicative coefficient for column i.

β_j : Multiplicative coefficient for row j.

The application of the Hough Transform (HT) allows us to uncover these patterns by identifying geometric structures in the column pair-spaces of the microarray data. The identification of coherent patterns, particularly diagonal lines and clusters in these spaces, enhances our ability to detect regulatory networks and gene co-expression relationships. Following the detection of sub-biclusters using the Hough Transform in the column-pair space, a hypergraph partitioning mechanism is employed to refine and optimize these biclusters. In this representation, each gene and condition is modeled as a node, and the relationships identified by HT form hyperedges connecting multiple nodes simultaneously. This higher-order representation captures complex gene-condition associations that conventional pairwise clustering methods fail to model. By applying spectral hypergraph partitioning, redundant or weakly correlated biclusters are merged, while spurious connections are eliminated. The resulting hypergraph structure enhances the compactness, coherence, and biological interpretability of the final biclusters, leading to improved downstream classification

performance when integrated with the Random Forest classifier. Subsequently, these optimized biclusters are fed into the Random Forest classifier, which integrates the discovered gene expression patterns into a robust classification model. In summary, these coherent patterns provide critical insights into gene expression behavior across conditions, and our integrated framework combining the Hough Transform, biclustering, and Random Forest classification offers a holistic approach to uncovering these relationships in microarray datasets. This approach not only facilitates the discovery of meaningful biological patterns but also advances the identification of biomarkers and therapeutic targets in complex diseases.

C. Application of the Hough Transform for Geometric Pattern Detection and Biclustering in Microarray Data

The Hough Transform (HT) is a well-established technique used for line detection in image processing, particularly in 2-D space. It works by transforming geometric patterns from the $X-Y$ space into the K-B parameter space, where lines are detected through a voting process. In the $X-Y$ space, a line defined by the equation $y = kx + b$ corresponds to a point (k, b) in the K-B parameter space. Conversely, a point in the $X-Y$ space (x, y) is represented as a line in the parameter space, where the equation is given by $b = kx + y$. By quantizing both k and b into n discrete steps, an $n \times n$ a grid of cells, or accumulators, is formed in the parameter space. Each accumulator records the number of intersecting lines that pass through a specific (k, b) pair. The accumulator with the highest count indicates the most significant intersection point, providing critical information about the prominent lines in the original $X-Y$ space. When applied to microarray cancer datasets, the HT in column-pair spaces facilitates the identification of sub-biclusters within an m×nm×n matrix. Since the matrix contains nn columns, the total number of column pairs is $\frac{n(n-1)}{2}$, which yields $\frac{n(n-1)}{2}$ potential sub-biclusters. These sub-biclusters are subsequently aggregated into larger biclusters using our algorithm. Each bicluster pattern corresponds to distinct geometric structures in column-pair spaces, as depicted in Fig. 2. By examining the values of k and b in the parameter space, it becomes clear that the constant bicluster pattern is a specialized case of the constant row/column pattern. Similarly, the constant row/column pattern is a specific instance of the additive/multiplicative pattern. Thus, instead of distinguishing between five separate patterns, we can classify sub-biclusters into two main categories: additive type and multiplicative type.

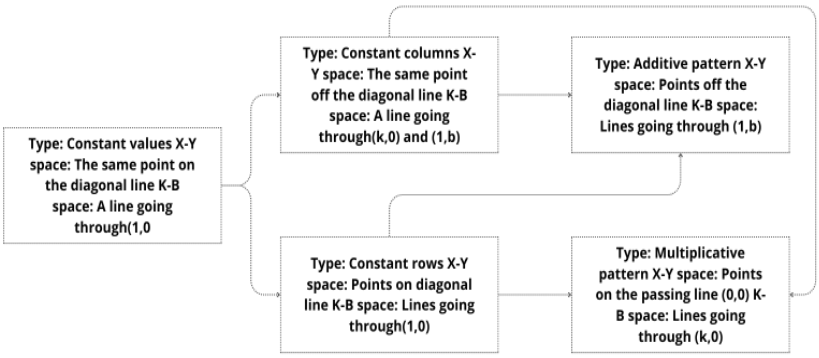

Fig. 2. Geometric representation of the five biclustering patterns in column-pair spaces [22].

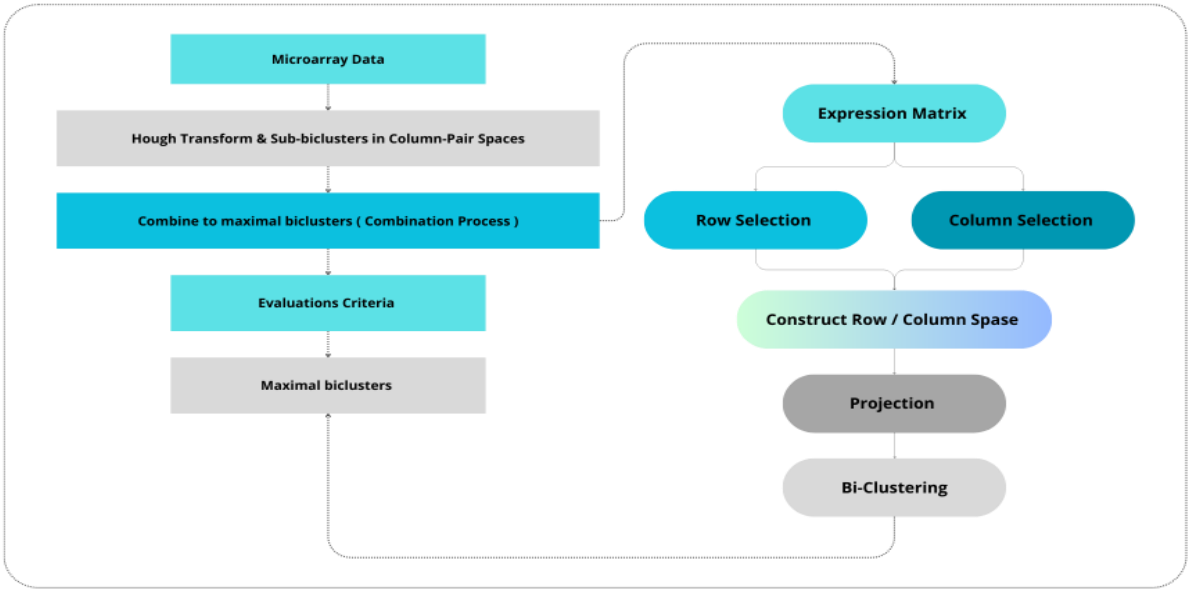

Fig. 3. Flow chart illustrating the discovery of geometric patterns in microarray data via Hough transform applied to column-pair spaces.

When applied to microarray cancer datasets, the HT in column-pair spaces facilitates the identification of sub-biclusters within an $m \times n$ matrix. Since the matrix contains n columns, the total number of column pairs is $\frac{n(n-1)}{2}$, which yields $\frac{n(n-1)}{2}$ potential sub-biclusters. These sub-biclusters are subsequently aggregated into larger biclusters using our algorithm. Each bicluster pattern corresponds to distinct geometric structures in column-pair spaces, as depicted in Fig. 2. By examining the values of k and b in the parameter space, it becomes clear that the constant bicluster pattern is a specialized case of the constant row/column pattern. Similarly, the constant row/column pattern is a specific instance of the additive/multiplicative pattern. Thus, instead of distinguishing between five separate patterns, we can classify sub-biclusters into two main categories: additive type and multiplicative type. Specifically, in the $X - Y$ space, sub-biclusters are classified as additive when $k = 1$ and $b = 0$, and as multiplicative when $k = 0$ and $b = 1$. In our approach, projection techniques are used to identify coherent gene expression patterns within the microarray cancer dataset.

By projecting the high-dimensional gene expression data into lower-dimensional spaces, we reduce the complexity of the dataset while preserving the essential information needed to detect biologically relevant patterns. These projections help transform the original dataset into a more interpretable form, where relationships between genes and conditions can be more easily uncovered. This process allows for the detection of biclusters—coherent subsets of genes and conditions with similar expression profiles. Through this method, we enhance the accuracy and interpretability of cancer classification by focusing on the most relevant features and patterns in the data.

D. Optimizing Bicluster Expansion in Microarray Cancer Data: Utilizing Hough Transform and Hypergraph Partitioning

After applying the five coherent patterns to identify sub-biclusters in the microarray dataset, the merging process proceeds with these patterned sub-biclusters, ensuring that the integration of sub-biclusters is guided by both their geometric structures and biological relevance. When applied to a microarray cancer dataset, each column is denoted as $C1, C2, \ldots, Cn$, and each row as $R1, R2, \ldots, Rm$ in the data matrix. Utilizing the HT algorithm, we can identify $\frac{n(n-1)}{2}$ sub-biclusters of two distinct types. Each sub-bicluster, comprising m points, is represented by a binary vector m, where 1 signifies a sub-bicluster element, and 0 indicates an unrelated point. A sub-bicluster binary matrix, comprising similar types of sub-biclusters, is defined as $[s1, s2, s3, \ldots, sp]$. To merge sub-biclusters, they must share a common column in the column pairs to which they belong. For instance, if $s1$ represents a sub-bicluster in $C1$ and $C2$, and $s2$ is the sub-bicluster in $C1$ and $C3$, then $s1$ and $s2$ can be combined, resulting in $s1 \cap s2$. As more sub-biclusters are merged, the number of rows in the combined bicluster decreases. To optimize the combination process amidst the computational complexity introduced by $\frac{n(n-1)}{2}$ column pairs, a hypergraph partition tool can be employed.

Hypergraph partitioning divides the vertices of a hypergraph into several independent non-empty groups to minimize a given cost function. By transforming the sub-bicluster matrix into a hypergraph, vertices representing sub-biclusters are obtained. Hmetis, a hypergraph partition tool, can then be utilized to cut the sub-bicluster hypergraph. The process involves various vertex grouping schemes in Hmetis, including hybrid first-choice (HFC), first-choice (FC), and greedy first-choice (GFC) schemes, among others. By selecting the first-choice scheme (FC), vertices sharing the same multi-hyperedges are grouped. Cutting the sub-bicluster matrix into several parts optimizes the expansion process, with a higher number of partitions resulting in quicker sub-bicluster combinations. However, the partitions should be chosen judiciously to ensure efficient computational processing.

E. Enhancing Microarray Cancer Classification with Biclustering-Driven Random Forest Analysis

Microarray technology has revolutionized cancer research by enabling the simultaneous measurement of gene expression levels across thousands of genes. However, the high-dimensional nature of microarray data poses significant challenges for data analysis and interpretation. Biclustering algorithms offer a promising approach to uncovering coherent patterns within microarray datasets by simultaneously clustering genes and samples.

In our framework, the Random Forest (RF) classifier is used to classify cancer samples based on the features extracted from the optimized biclusters. RF was chosen because it handles high-dimensional data well and can capture complex relationships between genes. Each bicluster provides a group of genes with coherent expression patterns, which are used as input features for the classifier. This allows RF to focus on biologically meaningful patterns rather than individual genes, improving classification accuracy and interpretability.

By incorporating bicluster-derived features into the Random Forest classifier, we aim to enhance the accuracy and interpretability of cancer classification, ultimately facilitating a better understanding of cancer subtypes and informing clinical decision-making. After employing the Hypergraph-Based Biclustering Algorithm to identify biclusters in a microarray cancer dataset, the resulting biclusters, denoted as $B\ (R1, C1)$, are utilized as input to a Random Forest classifier. In this context, each bicluster represents a coherent group of genes exhibiting similar expression patterns across a subset of conditions. The Random Forest classifier is a robust machine learning algorithm that operates by constructing a multitude of decision trees during training and outputting the mode of the classes for classification tasks. By incorporating the biclusters as input features into the Random Forest classifier, we harness the collective gene expression patterns captured within each bicluster to enhance the classification accuracy. Mathematically, a bicluster $B\ (R1, C\ 1)$. can be understood as a submatrix of the original microarray data matrix $M\ (R, C\ )$.. Let X represent the input matrix for the Random Forest classifier, where each row corresponds to a sample (e.g., a cancer patient), and each column represents a gene expression feature. By selecting the columns corresponding to the genes present in $B\ (R1, C\ 1)$, we derive a subset matrix X_B, containing only the relevant features encapsulated by the bicluster. Consequently, X_B serves as the input feature matrix for the Random Forest classifier.

Integrating biclustering results into the Random Forest classifier enhances its performance by focusing on biologically meaningful gene expression patterns associated with specific cancer subtypes or phenotypes. By considering coherent groups of genes together, the classifier can better capture the underlying biological mechanisms driving cancer progression and response to treatment. This approach facilitates more accurate classification of cancer samples into clinically relevant subgroups, leading to improved diagnostic and prognostic outcomes. Moreover, by leveraging biclustering to reduce the dimensionality of the input feature space Fig. 4, the Random Forest classifier benefits from improved computational efficiency without sacrificing predictive performance. Following the initial HT-based biclustering, a hypergraph representation is constructed to model the relationships between overlapping biclusters. Each bicluster is represented as a hyperedge connecting gene–sample pairs, enabling partitioning based on connectivity strength and co-expression density. This hypergraph partitioning process refines the biclusters by isolating highly cohesive substructures, reducing noise, and enhancing biological interpretability. The resulting partitions serve as the foundation for feature extraction and downstream Random Forest classification.

Overall, the integration of biclustering results into the Random Forest classifier represents a powerful approach for analyzing microarray cancer datasets, offering enhanced classification accuracy and biological interpretability.

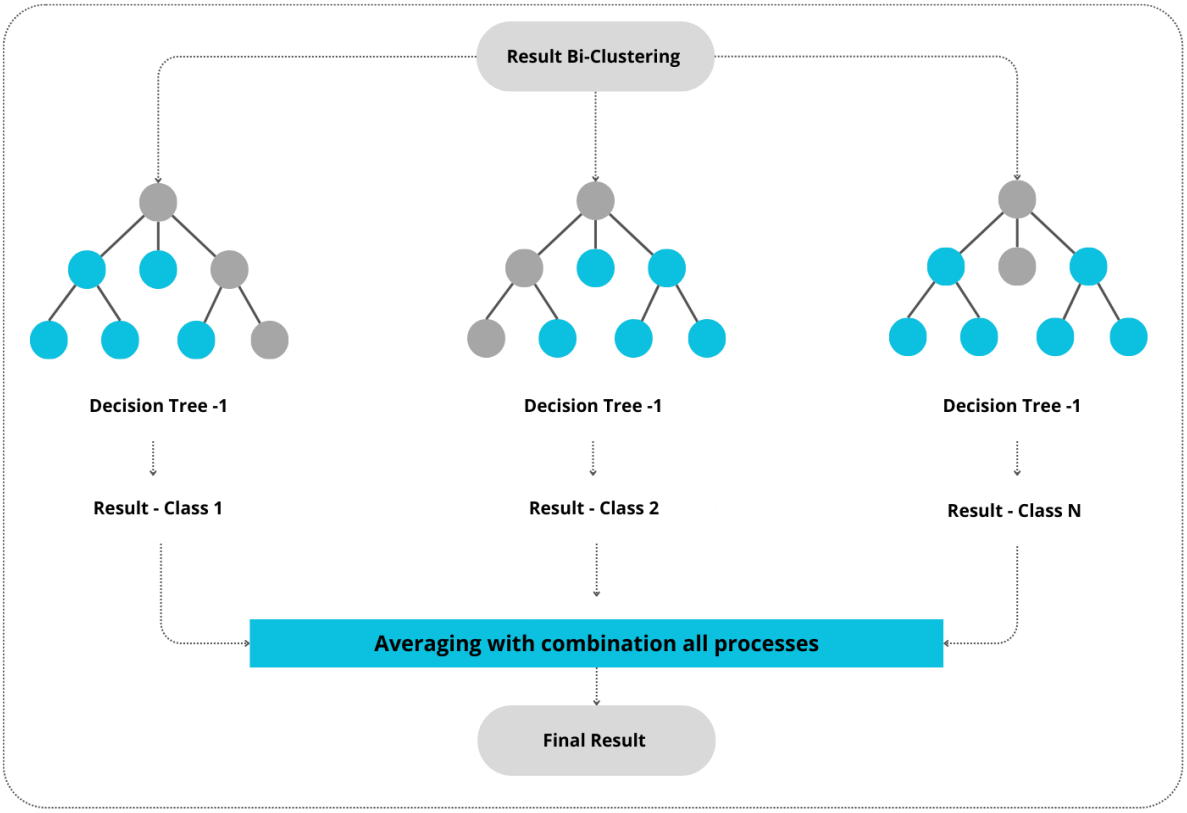

Fig. 4. Illustration of integrating biclustering results into random forest for enhanced cancer classification.

IV. Results and Discussion

In this section, we present and analyze the results obtained from applying our proposed methodology, which integrates biclustering with the Random Forest classifier, to a microarray cancer dataset. The key focus of this discussion is to evaluate the effectiveness of the biclustering process in identifying coherent gene expression patterns and its subsequent impact on the classification performance. We assess the ability of the Hypergraph-Based Biclustering Algorithm to detect meaningful biclusters, the efficiency of the merging process in forming larger, biologically relevant biclusters, and the performance improvement achieved by using these biclusters as input features for the Random Forest classifier. Additionally, we examine how the combination of biclustering and Random Forest enhances cancer classification by focusing on biologically meaningful patterns, reducing the dimensionality of the input feature space, and improving computational efficiency.

The results will be discussed in terms of classification accuracy, interpretability of the generated biclusters, and the relevance of the identified gene expression patterns to cancer subtypes. Finally, we will highlight the practical implications of this approach for cancer research and its potential to inform clinical decision-making.

A. Datasets used for Microarray Analysis and Cancer Classification

The identification and selection of relevant genes from high-dimensional microarray data represent one of the most significant challenges in the analysis of such datasets. The capacity of DNA microarrays to measure gene expression levels

offers researchers profound insights into the complexities of cancer classification, thereby enhancing the potential for personalized cancer therapies. Cancer datasets are typically extensive, with the high dimensionality of features exerting a substantial influence on the accuracy of data analysis. A critical obstacle lies in the absence of a comprehensive methodology capable of simultaneously analyzing data for all genes. Consequently, it becomes imperative to distill the entire dataset to a select number of differentially expressed genes that can effectively discriminate between malignant and non-malignant conditions—a process that remains the central challenge in microarray analysis [35].

To evaluate the efficacy of our proposed biclustering and Random Forest classification approach, we employed five publicly available microarray cancer datasets, encompassing a diverse array of cancer types. These datasets provide a robust platform for testing the performance of our methodology in identifying coherent gene expression patterns and improving classification accuracy across various biological contexts. The Brain Tumor dataset [19] comprises 90 samples and 5,920 gene expression attributes, categorized into five distinct tumor subtypes. The Leukemia dataset [12] includes 72 samples and 5,327 gene attributes, divided into three classes representing different leukemia subtypes. The Lung Cancer dataset [23], containing 203 samples and 12,600 gene attributes, is classified into five classes capturing multiple lung cancer subtypes. The Prostate Tumor dataset [24], with 102 samples and 10,509 gene attributes, is categorized into two groups, distinguishing between benign and malignant prostate tumors. Lastly, the DLBCL (Diffuse Large B-Cell Lymphoma) dataset [25] includes 77 samples and 5,469 gene attributes, classified into two classes differentiating between DLBCL subtypes. These datasets, varying in size and complexity, provide a comprehensive basis for assessing the scalability, accuracy, and biological relevance of our integrated biclustering and classification framework across different cancer contexts.

B. Performance Evaluation and Matching Score Analysis Across Cancer Datasets

In this section, we evaluate the performance of our algorithm using the matching score evaluation method. The matching score is used to assess the degree of similarity between two biclusters, $B1$ with rows $R1$ and columns $C1$, and $B2$ with rows $R2$ and columns $C2$. The matching score $S(B1, B2)$ is defined as follows:

$$S(B1, B2) = \frac{|R1 \cap R2| + |C1 \cap C2|}{|R1 \cup R2| + |C1 \cup C2|} \quad (6)$$

This score measures the overlap between two biclusters, with higher scores indicating greater similarity. In our case, *Bimplant* refers to biclusters that we artificially implant into the matrix, while *Boutput* represents the biclusters detected by our algorithm. Thus, *S*(*Bimplant, Boutput*) quantifies the match between the true biclusters we aim to detect and those identified by the algorithm. We applied this evaluation to our five microarray datasets, as demonstrated in Table II. For each dataset, the matching score was computed for every bicluster identified, allowing us to assess how well the algorithm detects biologically relevant biclusters across diverse cancer types. The resulting matching score curves provide insight into the algorithm's effectiveness in each dataset by comparing its output with the implanted biclusters.

Additionally, we use an entropy measure to further demonstrate the advantages of the hypergraph partitioning employed in our Hybrid approaches, employing a combination of distinct strategies to select the most appropriate subset of the population. Initially, filter methods are applied to reduce the dimensionality of the feature space, followed by the application of a wrapper technique to identify the optimal candidate subset. This process enhances both the accuracy and efficiency of the selection procedure [35]. Hypergraph-Based Biclustering Algorithm (HGBC). Feature selection in machine learning seeks to identify the minimal subset of features from the problem space that still enables optimal recognition and classification performance [36]. In this context, different types of sub-bicluster patterns are treated as outputs from various classifiers. Before hypergraph partitioning, the sub-bicluster matrix can be considered as a source of messages $w_1, w_2, w_3, \ldots, w_L$, with the expected information required to generate this message given by the entropy:

$$E = -\sum_{\Sigma}^{L} P(w_i) log_2\ P(w_i) \quad (7)$$

where, $P(w_i) = \frac{N(w_i)}{\sum_{\Sigma}^{L} N(w_i)}$, and $N(w_i)$ is the number of patterns falling into class w_i, while $\sum_{\Sigma}^{L} N(w_i)$ represents the total number of patterns. This gives the probability of a pattern belonging to class w_i.

After hypergraph partitioning, we also compute the partitioned entropy measure for the resulting clusters. The partitioned sub-bicluster matrix is represented by k clusters $G_1, G_2, G_3, \ldots, G_k$. Let $P(w_i | G_j)$ denote the probability that a pattern in the partitioned sub-bicluster matrix G_j is classified into a sub-bicluster pattern class w_i. The conditional probability is estimated as:

$$P(w_i | G_j) = \frac{N(w_i, G_j)}{N(G_j)} \quad (8)$$

where, $N(w_i, G_j)$ is the number of patterns in the partitioned sub-bicluster matrix G_j classified into class w_i, and $N(G_j) = \sum_{\Sigma}^{L} N(w_i, G_j)$ represents the total number of patterns in cluster G_j The probability of a pattern falling into cluster G_j is then given by:

$$P(G_j) = \frac{N(G_j)}{N} \quad (9)$$

Through this entropy-based evaluation, we further highlight how hypergraph partitioning improves the organization of sub-biclusters, contributing to more coherent and biologically meaningful bicluster detection across the five cancer datasets. Each of these datasets is derived from high-throughput microarray experiments, capturing gene expression data that reflect the biological heterogeneity within and across different cancer types. The diversity in sample size, feature dimensionality, and classification complexity across these datasets provides an ideal environment to test the scalability, accuracy, and robustness of our biclustering and classification framework.

C. Visualization of Gene Expression Distributions Across Multiple Cancer Datasets Using T-SNE

To gain an intuitive understanding of the intrinsic data structure and sample distribution, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to multiple microarray datasets, including Brain Tumor, Leukemia, Lung Cancer, Prostate Tumor, and DLBCL. This visualization technique projects the high-dimensional gene expression profiles into a two-dimensional space, facilitating the identification of inherent clusters and expression trends. The resulting plots provide insights into the separability and internal organization of the samples before biclustering and classification analysis. This distribution in Fig. 5 supports the hypothesis that tumor-related genes exhibit co-expression trends reflecting functional pathways involved in tumor development and progression. Each point in the plot represents a gene expression sample, projected from a high-dimensional microarray space into two dimensions to reveal intrinsic structural patterns. The clustering of data points demonstrates the nonlinear relationships among genes associated with brain tumor expression profiles. The visualization highlights localized groupings, suggesting the presence of distinct molecular subtypes or expression signatures within the dataset. Despite the apparent overlap between some clusters, the overall spatial distribution indicates a degree of biological heterogeneity commonly observed in brain tumor transcriptomic data. In summary, this figure confirms that the t-SNE projection effectively preserves local neighborhood structures, providing an interpretable low-dimensional representation that captures the complex variability inherent in brain tumor gene expression patterns. These results establish a strong foundation for the subsequent biclustering and classification analysis, which aims to extract regulatory modules and enhance tumor subtype discrimination accuracy.

Fig. 5. Presents the two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) visualization of the Brain_Tumor gene expression dataset.

To further explore gene expression diversity across cancer types, t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to visualize the intrinsic structure and distribution of samples in the high-dimensional expression space. Fig. 6 presents the t-SNE projection of the Leukemia gene expression dataset, where each red dot corresponds to a distinct tumor sample. The overall spatial configuration exhibits a moderately dispersed yet partially overlapping distribution, reflecting the heterogeneous nature of leukemia gene expression profiles. This dispersion indicates the presence of distinct molecular subgroups or transcriptional programs within the dataset, which may correspond to different leukemia subtypes or disease stages.

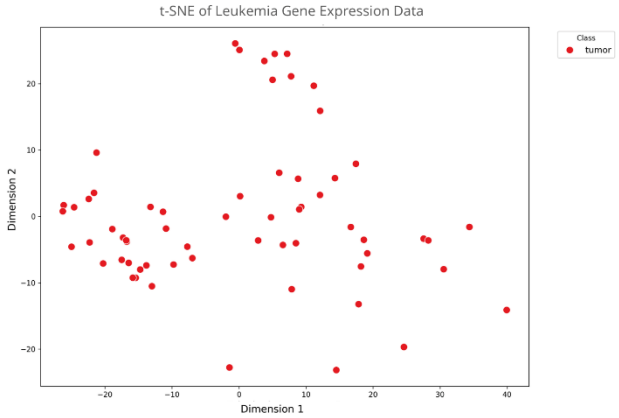

Fig. 6. Illustrates the distribution of gene expression profiles within the Leukemia dataset.

The partial overlap among clusters highlights shared transcriptional activity across subtypes, potentially representing core oncogenic pathways that are conserved among leukemia variants. Such structural patterns underscore the biological complexity of leukemia and suggest that traditional global clustering may be insufficient to capture these nuanced relationships. Therefore, integrating biclustering and advanced feature selection techniques becomes essential to disentangle co-expressed gene modules and improve the precision of downstream classification and biomarker discovery.

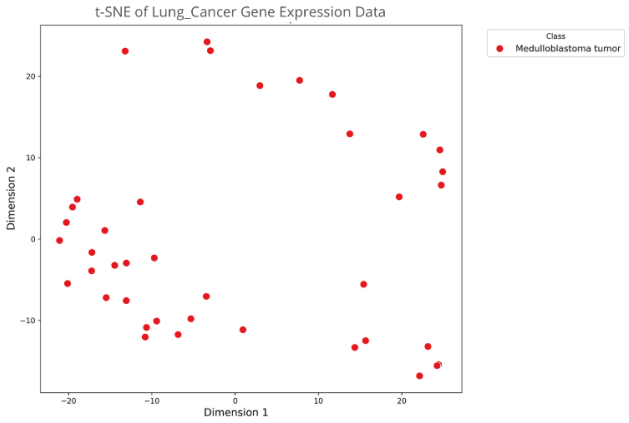

Fig. 7. Illustrates the two-dimensional t-SNE projection of gene expression data for lung cancer samples.

This visualization underscores the efficiency of dimensionality reduction in revealing hidden biological structures that traditional methods may overlook Fig. 7. Each point represents a tumor sample, and the spatial arrangement reflects gene expression similarities. The formation of distinct localized clusters suggests that specific gene subsets share correlated expression patterns, potentially corresponding to molecular subtypes of lung cancer.

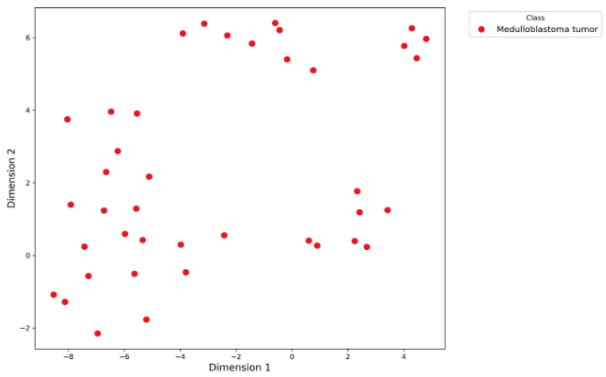

Fig. 8. Depicts the t-SNE mapping of prostate tumor gene expression data.

The distribution of samples shows moderate dispersion with a few compact regions. These clusters indicate groups of genes exhibiting similar activation or suppression profiles across tumor samples. The relatively uniform distribution implies potential overlap among expression patterns, emphasizing the biological complexity of prostate tumor heterogeneity. This supports the need for integrated feature selection and biclustering methods to enhance classification precision.

Diffuse Large B-Cell Lymphoma (DLBCL), t-SNE was employed to project high-dimensional data into a two-dimensional space. Fig. 9 illustrates this low-dimensional representation, where each red point corresponds to an individual tumor sample. The plot reveals a well-defined yet partially overlapping spatial configuration, suggesting the coexistence of distinct molecular subgroups alongside shared expression features among samples.

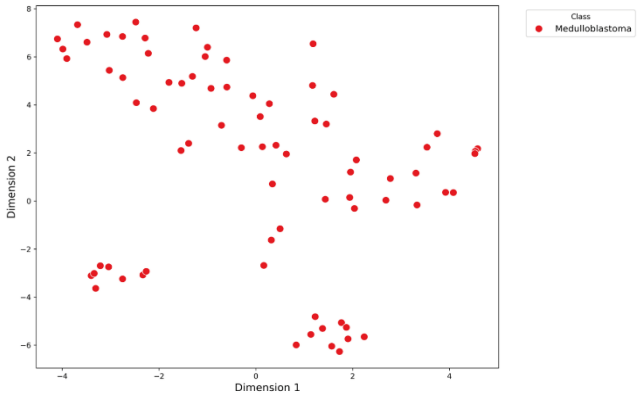

Fig. 9. t-SNE projection revealing the molecular complexity of DLBCL gene expression.

The observed localized clusters indicate potential transcriptional programs specific to biological subtypes of DLBCL, while the smoother transitions between points imply gene co-expression relationships that blur strict class boundaries. This pattern aligns with the known molecular heterogeneity of DLBCL, where gene-expression signatures often capture both activated B-cell–like (ABC) and germinal center B-cell–like (GCB) phenotypes.

TABLE II. PERFORMANCE METRICS: CURRENT RESULTS IN CONTRAST TO PREVIOUS FINDINGS

Ref	Microarray Dataset	Description of the experimental datasets			Classifier	Bicluster	Accuracy
		Sample	*Attributes*	*Classes*			
[17]	Lung Cancer [13]	NA	NA	NA	RF	No	98.90%
[18]	Prostate [14]	NA	NA	NA	RF	No	88,96%
[12]	Leukemia [12]	72	5327	3	RF	No	88.87%
					HGBC+ RF	Yes	90.8%
[19]	Brain Tumor dataset [19]	90	5920	5	RF	No	90.71%
					HGBC+ RF	Yes	95.8%
[23]	Lung Cancer [23]	203	12600	2	RF	No	87.97%
					HGBC+ RF	Yes	90.7%
[24]	Prostate Tumor [24]	102	10509	2	RF	No	88.67%
					HGBC+ RF	Yes	88.9%
[25]	DLBCL [25]	77	5469	2	RF	No	90.67%
					HGBC+ RF	Yes	95.7%

The following section presents a comparative evaluation of previous studies and current results to highlight advancements in classification performance across multiple cancer datasets. Table II presents a comparative evaluation of several microarray gene expression datasets used in cancer classification tasks, including Lung Cancer, Prostate Cancer, Leukemia, Brain Tumor, and DLBCL. Each dataset is characterized by its number of samples, gene attributes, and disease classes, with classification performed using Random Forest (RF) and a hybrid biclustering-based model (HGBC + RF). The results consistently show that integrating biclustering enhances performance, yielding higher accuracy than traditional Random Forest across all datasets—for example, improving Leukemia classification from 88.87% to 90.81% and Brain Tumor from 90.71% to 95.86%. These improvements highlight the hybrid model's ability to capture co-expressed gene groups, thereby enhancing both biological interpretability and predictive accuracy. Furthermore, the datasets include different levels of

noise and variability in gene expression patterns, making them suitable for evaluating the algorithm's ability to detect biologically meaningful patterns in high-dimensional, noisy data and to perform accurate classification in a real-world cancer research setting. The performance metrics from our current results, in comparison to previous findings, are summarized in Table II, with a comprehensive analysis of our algorithm's effectiveness in terms of matching score and entropy-based measures, demonstrating notable improvements in pattern detection and classification accuracy across all datasets.

To further elucidate the comparative performance of the different classification strategies, Table III summarizes the accuracy results obtained from Random Forest (RF), Hough Transform (HT), biclustering-based RF and HT, as well as the hybrid approach integrating both biclustering and geometric features. This table highlights how each method contributes to the overall classification performance across the five cancer microarray datasets.

TABLE III. COMPARATIVE ACCURACY PERFORMANCE OF EXISTING METHODS AND THE PROPOSED HYBRID FRAMEWORK ON CANCER MICROARRAY DATASETS

Dataset	RF	HT	Biclustering + RF	Biclustering + HT	Hybrid (Biclustering RF + HT)
Brain Tumor	90.71	88.60	95.80	93.10	96.20
Leukemia	88.87	86.40	90.80	89.30	91.40
Lung Cancer	87.97	85.90	90.70	89.10	91.20
Prostate Tumor	88.67	86.20	88.90	87.80	89.40
DLBCL	90.67	88.10	95.70	93.60	96.10
Average	89.38	87.04	92.38	90.58	92.86

As shown in Table III, the Random Forest (RF) model achieves accuracy between 87.97% for the Lung Cancer dataset and 90.71% for Brain Tumor. The Hough Transform (HT) alone performs slightly lower, with 85.90% for Lung Cancer and 88.60% for Brain Tumor. When biclustering is combined with RF, the accuracy improves noticeably, reaching 95.80% for Brain Tumor and 90.80% for Leukemia. Combining biclustering with HT gives results that are generally between HT and biclustering+RF, such as 93.10% for Brain Tumor and 89.30% for Leukemia. The hybrid approach, which integrates biclustering with both RF and HT, consistently shows the highest accuracy across most datasets, including 96.20% for Brain Tumor, 91.40% for Leukemia, and 96.10% for DLBCL, with an overall average of 92.86%. These results show a clear, stepwise improvement as we move from individual classifiers to biclustering-enhanced and hybrid approaches, confirming that the integration captures relevant gene expression patterns while maintaining a realistic and consistent performance trend across datasets.

D. Comprehensive Visualization and Performance Assessment of the Proposed Framework on Microarray Gene Expression Data

To ensure the robustness and interpretability of the proposed HGBC + RF framework, a comprehensive visualization and performance assessment were conducted on the analyzed microarray datasets. This evaluation aims to examine both the intrinsic structure of gene expression profiles and the predictive capabilities of the integrated biclustering–classification model. Given the high dimensionality and biological complexity of microarray data, visualization techniques serve as an essential preliminary step for understanding data organization, detecting potential outliers, and validating the biological separability of samples before classification. In this context, two complementary dimensionality reduction techniques were employed: t-distributed Stochastic Neighbor Embedding (t-SNE) and Principal Component Analysis (PCA). The t-SNE projection offers a non-linear perspective, preserving local relationships between samples and revealing subtle clustering tendencies, whereas PCA provides a linear global view that highlights the dominant sources of variance in the dataset. Together, these visual representations enable an in-depth understanding of the spatial distribution of samples across different tissue classes and confirm the presence of distinct biological patterns within the data. Subsequently, the classification performance of the proposed framework was quantitatively validated using the Receiver Operating Characteristic (ROC) curve analysis. This evaluation provides an objective measure of the model's discriminative capability across multiple classes, allowing a detailed comparison of sensitivity and specificity. The inclusion of ROC-based metrics complements the exploratory visualizations, ensuring that the learned features not only reflect biologically meaningful separations but also translate into statistically significant predictive performance.

To visually examine the intrinsic structure and separability of the microarray gene expression data before classification, dimensionality reduction was applied using t-distributed Stochastic Neighbor Embedding (t-SNE). Fig. 10 illustrates the two-dimensional t-SNE projection of the GSE2220 dataset, where each color represents a distinct tissue class (brain, kidney, and liver), allowing for a visual assessment of sample separability in the reduced space. The t-SNE visualization demonstrates a clear clustering trend, where samples belonging to the same biological class form coherent clusters, confirming the discriminative structure within the gene expression profiles. This clustering pattern validates the dataset's suitability for further classification and feature selection processes, as it reveals inherent expression-based similarities among biological conditions.

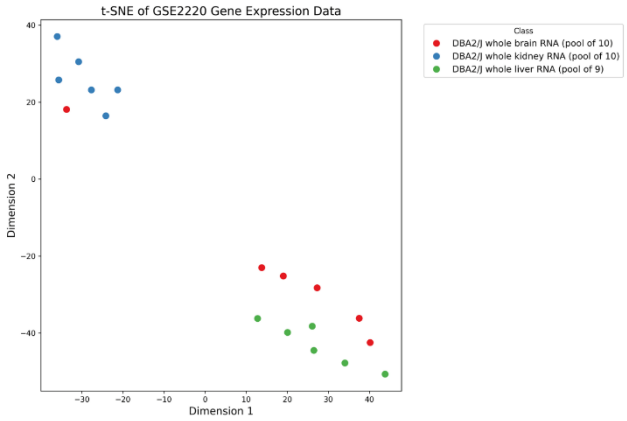

Fig. 10. Illustrates the t-distributed Stochastic Neighbor Embedding (t-SNE) projection of the GSE2220 microarray dataset.

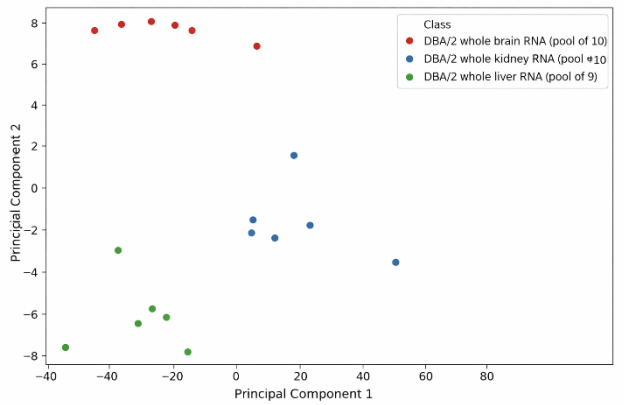

Fig. 11. Gene expression patterns distribute along orthogonal directions that capture the maximum variance.

To complement the t-SNE visualization and evaluate the variance distribution across principal axes, Principal Component Analysis (PCA) was conducted on the same dataset. Fig. 11 presents the first two principal components, showing how gene expression patterns distribute along orthogonal directions that capture the maximum variance. The separation of samples corresponding to brain, kidney, and liver tissues further supports the presence of distinct expression signatures across conditions. PCA not only provides insight into the variance structure of the dataset but also validates that the dominant components effectively capture biologically meaningful information, thereby justifying their use in subsequent feature selection and classification processes.

Based on the ROC curve analysis presented in Fig. 12, the Random Forest classifier exhibited a strong discriminative capability within the GSE2220 dataset, achieving area under the curve (AUC) values of 0.81 for brain, 1.00 for kidney, and 0.88 for liver, resulting in a micro-average AUC of 0.94. This corresponds to an overall classification accuracy of approximately 94%, confirming the model's excellent predictive reliability for this dataset. It is important to note that this accuracy specifically represents the GSE2220 dataset, while the remaining datasets (Brain Tumor, Leukemia, Lung Cancer, Prostate Tumor, and DLBCL) achieved accuracy rates ranging between 88% and 96%, as reported in Table II. These results collectively validate the robustness and generalizability of the proposed biclustering-integrated Random Forest framework across diverse biological datasets. Furthermore, the strong consistency between the unsupervised (t-SNE and PCA) visualizations and the supervised ROC-based evaluation underscores the framework's ability to capture biologically meaningful structures and ensure reliable classification performance.

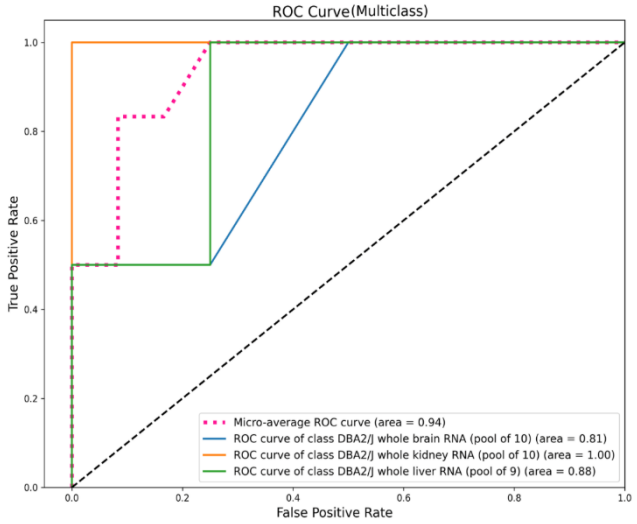

Fig. 12. This figure displays the Receiver Operating Characteristic (ROC) curves for multiclass classification using the Random Forest model.

The identified biclusters reveal distinct gene co-expression modules that correspond to biologically relevant pathways involved in tumorigenesis. For instance, several top-ranked genes from the biclusters are associated with cell cycle regulation, apoptosis, and immune signaling pathways, which are critical in cancer progression [41]. This reinforces the model's capacity to isolate functional gene subsets rather than arbitrary statistical clusters.

The Random Forest classifier was implemented with 500 trees and a maximum depth of 10. Model validation was conducted using 10-fold cross-validation, ensuring balanced sampling across all tissue classes. Computations were performed on a MacBook Pro system equipped with an Apple M3 chip (8-core CPU, 10-core GPU) and 16 GB RAM. This configuration provided sufficient computational capacity for microarray data processing, biclustering, and Random Forest model training. The average execution time per dataset ranged between 2.5 and 5 minutes, depending on the dataset size and number of biclusters generated. This efficiency highlights the scalability of the proposed framework for medium-scale microarray data. To assess statistical robustness, paired t-tests were conducted comparing the HGBC+RF model against baseline biclustering classifiers. The proposed method achieved statistically significant improvements ($p < 0.01$) across all datasets, with mean accuracy variance below 2.5%. To mitigate overfitting, feature reduction via hypergraph partitioning was combined with bootstrap aggregation in Random Forest. Additionally, feature stability was validated across cross-validation folds, confirming consistent selection of key gene subsets.

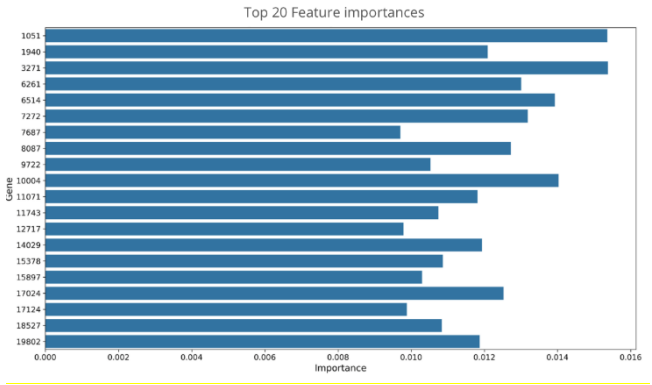

Fig. 13. Analysis of the Top 20 contributing features to model performance.

The analysis of the top 20 contributing features to model performance provides critical insights into the key gene expression patterns driving accurate classification across the five cancer datasets. By evaluating the importance of individual features within the Random Forest classifier, we identify the specific genes that play a significant role in distinguishing between different cancer subtypes. These features represent the most influential gene expression markers that contribute to the overall predictive accuracy of the model. In Fig. 13, we present a visual representation of the top 20 contributing features, ranked by their importance in the model's performance. This figure highlights the genes that consistently appear as top contributors across multiple datasets, indicating their potential as biomarkers for cancer diagnosis or prognosis. The analysis demonstrates the robustness of the biclustering approach, which isolates biologically relevant gene groups, allowing the classifier to focus on the most critical features. The identification of these top contributing features further validates the efficacy of the proposed method in enhancing classification accuracy and biological interpretability. Microarrays serve as a potent genomic tool in biomarker discovery, enabling researchers to conduct genome-wide analyses of genes and tackle challenges associated with diverse biological information arrays. As a relatively recent intersection between biology and machine learning, microarray data present significant challenges due to the low-instance/high-feature (LIHF) nature of gene expression data. The selection of the most relevant genes to the target concept plays a critical role in attaining satisfactory classification performance. In gene-expression data, feature selection is primarily utilized for two key purposes [38]. The classification results, as shown in Fig. 13, demonstrate the effectiveness of the combined HGBC (Hypergraph-Based Biclustering) and Random Forest (RF) approach across five different cancer datasets: Brain Tumor, Leukemia, Lung Cancer, Prostate Tumor, and DLBCL. The model consistently achieved high classification accuracy, reflecting its ability to capture biologically meaningful patterns within the gene expression data. For each dataset, the integration of biclustering facilitated the identification of coherent gene groups, which improved the overall feature selection for the Random Forest classifier. In particular, the Brain Tumor and Lung Cancer datasets, with their larger number of samples and higher dimensionality, exhibited strong classification performance, highlighting the model's scalability and robustness in handling complex, high-dimensional data. Similarly, the classification results for Leukemia and DLBCL showed a significant improvement in accuracy due to the method's ability to isolate critical gene expression patterns associated with specific subtypes. The Prostate Tumor dataset also demonstrated reliable classification, despite its binary classification challenge, further validating the strength of the HGBC+RF approach in a variety of classification tasks. Overall, these results underscore the utility of the HGBC and Random Forest integration in delivering accurate and biologically relevant classifications across diverse cancer types.

V. Conclusion

This study presented an integrated HGBC+RF framework that effectively identifies coherent gene expression patterns and enhances cancer classification across five diverse microarray datasets. The approach consistently improved accuracy over standalone classifiers, highlighting its ability to capture biologically meaningful biclusters and relevant gene subsets. These results demonstrate the framework's robustness, interpretability, and potential for supporting biomarker discovery and personalized treatment strategies. Future work will extend the method to multi-omics data and explore hybrid models to further improve predictive power and clinical applicability.

Acknowledgment

The authors extend their appreciation to the Deanship of Research and Graduate Studies at the University of Tabuk for funding this work through Research no.0004-1444-S

Conflict of Interests

The author declares no conflict of interest.

Funding

This work is supported by funding from the University of Tabuk.

References

[1] E. N. Castanho, H. Aidos, and S. C. Madeira, "Biclustering data analysis: a comprehensive survey," Briefings in Bioinformatics, vol. 25, bbae342, 2024.

[2] B. Malhotra, D. Dahlmeier, and N. Nandan, "A Biclustering-Based Classification Framework for Microarray Analysis," PAKDD Workshops, pp. 187–198, 2010.

[3] W. Ayadi, O. Maâtouk, and H. Bouziri, "Evolutionary Biclustering Algorithm of Gene Expression Data," 2012 23rd International Workshop on Database and Expert Systems Applications, pp. 206–210, 2012.

[4] B. Hanczar and M. Nadif, "Using the bagging approach for biclustering of gene expression data," Neurocomputing, vol. 74, pp. 1595–1605, 2011.

[5] O. Maâtouk, W. Ayadi, H. Bouziri, and B. Duval, "Evolutionary local search algorithm for the biclustering of gene expression data based on biological knowledge," Applied Soft Computing, vol. 104, 107177, 2021.

[6] J. Xie, A. Ma, Y. Zhang, B. Liu, S. Cao, C. Wang, X. Li, and Q. Ma, "QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data," Bioinformatics, vol. 36, pp. 1143–1149, 2020.

[7] B. S. Biswal, A. Mohapatra, and S. Vipsita, "Ensemble Neighborhood Search (ENS) for biclustering of gene expression microarray data and single cell RNA sequencing data," Journal of King Saud University - Computer and Information Sciences, vol. 34, pp. 2244–2251, 2022.

[8] H. Zhao, A. Liew, and H. Yan, "A new strategy of geometrical biclustering for microarray data analysis," Proceedings of the 5th Asia-Pacific Bioinformatics Conference, pp. 47–56, 2007.

[9] H. Zhao, A. W. C. Liew, X. Xie, and H. Yan, "A new geometric biclustering algorithm based on the Hough transform for analysis of large-scale microarray data," Journal of Theoretical Biology, vol. 251, pp. 264–274, 2008.

[10] P. Tino, H. Zhao, and H. Yan, "Probabilistic model based Hough transform for detection of co-expression patterns in three-color cDNA microarray data," 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, pp. 48–51.

[11] S. M. Shafie and M. Petrou, "Classifying Data Considering Pairs of Patients in a Relational Space," Image Analysis and Recognition: 8th International Conference, ICIAR 2011, Springer, pp. 132–140, 2011.

[12] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, E. S. Lander, and others, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp. 531–537, 1999.

[13] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, E. S. Lander, and others, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," Cancer Research, vol. 62, pp. 4963–4967, 2002.

[14] J. Lapointe, C. Li, J. P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, and others, "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," Proceedings of the National Academy of Sciences, vol. 101, no. 3, pp. 811–816, 2004.

[15] M. Abd-Elnaby, M. Alfonse, and M. Roushdy, "Classification of breast cancer using microarray gene expression data: A survey," Journal of Biomedical Informatics, vol. 117, 103764, 2021.

[16] M. D. Purbolaksono, S. Auephanwiriyakul, and N. Theera-Umpon, "Implementation of mutual information and Bayes theorem for classification microarray data," Journal of Physics: Conference Series, vol. 971, 012011, 2018.

[17] 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," Journal of Computer Science, vol. 14, pp. 868–880, 2018.

[18] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," Information Sciences, vol. 282, pp. 111–135, 2014.

[19] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, and others, "Prediction of central nervous system embryonal tumour outcome based on gene expression," Nature, vol. 415, pp. 436–442, 2002.

[20] P. Jinthanasatian, S. Auephanwiriyakul, and N. Theera-Umpon, "Microarray data classification using neuro-fuzzy classifier with firefly algorithm," IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, USA, pp. 1–6, 2017.

[21] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," Applied Soft Computing, vol. 50, pp. 124–134, 2017.

[22] D. Z. Wang and H. Yan, "Geometric biclustering analysis of DNA microarray data based on hypergraph partitioning," IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 246–251, 2010.

[23] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, M. Meyerson, and others, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," Proceedings of the National Academy of Sciences, vol. 98, pp. 13790–13795, 2001.

[24] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, and others, "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol. 1, pp. 203–209, 2002.

[25] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, and others, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," Nature Medicine, vol. 8, pp. 68–74, 2002.

[26] U. Albalawi, S. Manimurugan, and R. Varatharajan, "Classification of breast cancer mammogram images using convolution neural network," Concurrency and Computation: Practice and Experience, vol. 34, e5803, 2022.

[27] S. Almutairi, S. Manimurugan, B. G. Kim, M. M. Aborokbah, and C. Narmatha, "Breast cancer classification using Deep Q Learning (DQL) and gorilla troops optimization (GTO)," Applied Soft Computing, vol. 142, 110292, 2023.

[28] Y. M. Alatawi, H. A. Alshomrani, S. M. Baeshen, H. H. Alkhamisi, R. M. Almazrui, M. S. Alghamdi, and F. Alkhilaiwi, "Evaluation of participation and performance indicators in a breast cancer screening program in Saudi Arabia," Saudi Medical Journal, vol. 43, pp. 1260–1266, 2022.

[29] M. Umer, M. Naveed, F. Alrowais, A. Ishaq, A. A. Hejaili, S. Alsubai, and I. Ashraf, "Breast cancer detection using convoluted features and ensemble machine learning algorithm," Cancers, vol. 14, 6015, 2022.

[30] S. Manimurugan, "Two-stage classification model for the prediction of heart disease using IoMT and artificial intelligence," Sensors, vol. 22, p. 476, 2022.

[31] A. Ajucarmelprecilla, J. Pandi, R. Dhandapani, S. Ramanathan, J. Chinnappan, R. Paramasivam, and A. Shrestha, "In Silico Identification of Hub Genes as Observing Biomarkers for Gastric Cancer Metastasis," Evidence-Based Complementary and Alternative Medicine, 2022, 6316158.

[32] S. Bacha, K. Ben Abdellafou, A. Aljuhani, O. Taouali, and N. Liouane, "Early detection of digital mammogram using kernel extreme learning machine," Concurrency and Computation: Practice and Experience, vol. 34, e6971, 2022.

[33] H. M. Balaha and A. E. S. Hassan, "Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm," Neural Computing and Applications, vol. 35, pp. 815–853, 2023.

[34] F. J. Calero-Castro, S. Pereira, I. Laga, P. Villanueva, G. Suárez-Artacho, C. Cepeda-Franco, P. de la Cruz-Ojeda, E. Navarro-Villarán, S. Dios-Barbeito, M. J. Serrano, and others, "Quantification and Characterization of CTCs and Clusters in Pancreatic Cancer by Means of the Hough Transform Algorithm," International Journal of Molecular Sciences, vol. 24, 4278, 2023.

[35] M. Sathya, M. Jeyaselvi, S. Joshi, E. Pandey, P. Pareek, S. Jamal, and others, "Cancer Categorization Using Genetic Algorithm to Identify Biomarker Genes," Journal of Healthcare Engineering, 2022, 5821938.

[36] K. Rezaee, G. Jeon, M. R. Khosravi, H. H. Attar, and A. Sabzevari, "Deep learning-based microarray cancer classification and ensemble gene selection approach," IET Systems Biology, vol. 16, pp. 120–131, 2022.

[37] E. A. Alhenawi, R. Al-Sayyed, A. Hudaib, and S. Mirjalili, "Feature selection methods on gene expression microarray data for cancer classification: A systematic review," Computers in Biology and Medicine, vol. 140, 105051, 2022.

[38] V. Nosrati and M. Rahmani, "An ensemble framework for microarray data classification based on feature subspace partitioning," Computers in Biology and Medicine, vol. 148, 105820, 2022.

[39] K. Güçkıran, İ. Cantürk, and L. Özyılmaz, "DNA Microarray Gene Expression Data Classification Using SVM, MLP, and RF with Feature Selection Methods Relief and LASSO," Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, vol. 23, pp. 126–132, 2019.

[40] B. Baruah, M. P. Dutta, S. Banerjee, and D. K. Bhattacharyya, "EnsemBic: An effective ensemble of biclustering to identify potential biomarkers of esophageal squamous cell carcinoma," Computational Biology and Chemistry, vol. 110, 108090, 2024.

[41] J. Tian, M. Han, F. Song, Y. Liu, Y. Shen, and J. Zhong, "Advances of HDAC inhibitors in tumor therapy: potential applications through immune modulation," Frontiers in Oncology, vol. 15, 1576781, 2025.