e-ISSN : 2600-8793

Available online at
https://jcrinn.com/
https://crinn.conferencehunter.com/

Journal of Computing Research and Innovation

Journal of Computing Research and Innovation 9(1) 2024

A Review Study of Microarray Data Classification with the Application of Dimension Reduction

Sharifah Nadia Syed Hasan[1*], Noor Wahida Jamil[2]

[1,2]*College of Computing, Informatics and Mathematics, UiTM Melaka Campus, Jasin Branch, Melaka, Malaysia*

ARTICLE INFO

Article history:
Received: 31 January 2024
Revised: 27 February 2024
Accepted: 27 February 2024
Online first: 1 March 2024
Published 1 March 2024

Keywords:
Microarray
Classification
Dimension Reduction
Feature Extraction
Feature Selection

DOI:
10.24191/jcrinn.v9i1.424

ABSTRACT

Background. The growth of gene expression or microarray data, mainly in cancer disease, has become a game changer for feature selection techniques in handling complex data. Hence, the advancement of Deoxyribonucleic acid (DNA) microarray technology has made it feasible to measure the expression level of thousands of genes with the ability to diagnose early detection. This extensive study is conducted to review and analyse literature related to applying various dimensionality reduction approaches to predict microarray data. This study is aimed for the Data Science and Medical Sciences disciplines with the goal of extending future research and broader interdisciplinary collaboration efforts.
Methods. The systematic review of this study is based on the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines and reported in accordance with the PRISMA statement. Other than that, a systematic search is conducted using two search engines including, Scopus and Web of Science (WoS), from 2018 to 2022 by inputting the "feature extraction," "feature selection," "classification," and "microarray" as keywords. Based on the inclusion and exclusion criteria, the final articles available for review are 53 articles. Specifically, this study reports on the performance of feature selection approaches and the empirical comparisons of classification techniques used on the microarray dataset.
Results. According to the analysis, part of the included articles is mostly hybrid and novel approaches proposed for gene selection. Many novel and hybrid methods were developed to produce a good performance in terms of accuracy and computational efficiency. Moreover, the hybrid methods are proven effective in reducing dimensions and selecting relevant features. Besides, machine learning techniques are still the top interest among researchers for classification despite the emergence of deep learning approaches.

1. INTRODUCTION

Cancer is the largest cause of mortality globally, accounting for over 10 million fatalities in 2020, or roughly one in every six deaths (World Health Organization, 2022). However, note that many cancers can be cured through early detection and appropriate treatment. Therefore, the advancement of DNA microarray

[1*] Corresponding author. *E-mail address*: nadiahsnkaff@gmail.com

https://doi.org/10.24191/jcrinn.v9i1

technology has made it feasible to measure the expression level of thousands of genes that cause cancer and contribute to diagnosis along with the curative effect related to cancer (Lee et al., 2021).

Microarray data play a significant role in diagnosing various cancer and gene expression analyses. Due to the fact that microarray data allows the expression value of thousands of genes to be revealed at once, microarray data have been utilized extensively to identify the gene-cancer and gene-gene relationships of Leukemia cancer (Bilen et al., 2020). In most microarray data, the high number of genes that the microarrays accommodate is the majority irrelevant to the diseases (Othman et al., 2020). Identifying the key genes from high-dimensional gene expression data is essential to enhance cancer detection and provide specific treatment regimens. An efficient algorithm is crucial for gene selection, mainly when dealing with large microarray data.

Many studies related to microarray gene expression have been conducted to test different classification techniques for cancer data. However, if no preprocessing data measures are taken beforehand, the high dimensionality of microarray data can adversely affect generalization ability. Thus, one way to potentially solve this issue is by selecting relevant features to produce optimal classification performance. Feature selection approaches have been suggested for gene selection to select the most significant and informative genes. In recent years, unsupervised feature selection methods have received much interest in gene selection as they can discover the most discriminating subsets of genes, namely the potential information in biological data (Lu et al., 2021). Iochins Grisci et al. (2019) acknowledge that feature selection is a valuable approach in aiding biomarker identification due to its ability to locate a subset of genes with higher discriminatory power.

Gene expression or microarray data is well-known for its high dimensionality, limited sample size, and irrelevant features that lead to complex analysis. According to Ke et al. (2018), the complexity of high-dimensional data reduced through feature selection has a variety of potential benefits, including (i) limiting the overfitting of classifiers, (ii) improving prediction accuracy, and (iii) minimizing computational time and complexity. On the other hand, El Kafrawy et al. (2021) also agreed with previous relevant biomedical investigations that thorough feature selection testing resulted in high classification accuracy, successful time complexity resolution, and effective informative gene selection. Note that properly implementing feature selection methods would eliminate the concerns mentioned earlier and save significant computational resources (Khan et al., 2019).

Despite the abundance of feature selection literature, the "curse of dimensionality" remains challenging to achieve maximum classification accuracy. Over the last decade, numerous feature selection strategies have been presented to consider a more accurate and effective manner of classifying microarray data. As a result, this review aims to examine previous papers, summarizing significant findings on the applicability and efficacy of feature selection approaches in dealing with microarray data.

2. MATERIALS AND METHODS

The systematic review of this study relied on the PRISMA guidelines and is reported following the PRISMA statement.

2.1 Identification

This paper systematically searched for the most relevant papers on feature selection applied to microarray data. Therefore, a query search is run through different prominent digital databases, Scopus and WoS (Table 1). In both databases, searches were conducted by inputting the following keywords "feature extraction," "feature selection," "classification," and "microarray." The initial search resulted in an extensive list of publications, with 784 papers from Scopus and 204 from WoS. Subsequently, the search query results were refined to consider only the recent five years' publications from 2018 to 2022. The search was also limited to journals, articles, and English.

Table 1. Search string

Database	Search String
Scopus	TITLE-ABS-KEY (**"feature extraction"** AND **"feature selection"** AND **"microarray"** AND **"classification"**) AND (LIMIT-TO (OA , **"all"**)) AND (LIMIT-TO (PUBYEAR , **2022**) OR LIMIT-TO (PUBYEAR , **2021**) OR LIMIT-TO (PUBYEAR , **2020**) OR LIMIT-TO (PUBYEAR , **2019**) OR LIMIT-TO (PUBYEAR , **2018**)) AND (LIMIT-TO (DOCTYPE , **"ar"**)) AND (LIMIT-TO (EXACTKEYWORD , **"Feature Extraction"**) OR LIMIT-TO (EXACTKEYWORD , **"Feature Selection"**) OR LIMIT-TO (EXACTKEYWORD , **"Microarray Data"**) OR LIMIT-TO (EXACTKEYWORD , **"Classification"**)) AND (LIMIT-TO (LANGUAGE , **"English"**)) AND (LIMIT-TO (SRCTYPE , **"j"**))
Web of Sciences	**feature extraction* feature selection* microarray* classification** (Topic) and **2022** or **2021** or **2020** or **2019** or **2018** (Publication Years) **Open Access** and **Articles** (Document Types) and **English**

2.2 Screening

The refinement process of the search query has left several publications for review. However, all duplicates from Scopus and WoS are removed through a screening procedure. The titles of the papers from both databases were screened based on the established selection criteria.

2.3 Eligibility

After removing the duplicate articles, all articles eligible for review were retrieved. For the third phase, 66 articles have been accessed for eligibility. The titles and abstract were rigorously reviewed at this stage to ensure that the inclusion criteria were met regarding their relevance to the present research aims (Table 2).

Hence, the studies were considered eligible if they met the following requirements:

(i) Articles that are centered around "feature extraction," "feature selection," and "microarray data."

(ii) Only considering articles that utilize microarray data for analysis.

(iii) Primary research instead of review papers.

(iv) The entire manuscript is considered rather than just an abstract.

The study is restricted to only microarray data as this study aims to investigate the ability of dimensionality reduction and machine learning in this specific area of gene expression or microarray. As a result, thirteen (13) reports were excluded since they were unrelated to the current study. Finally, fifty-three (53) articles are available for review.

Table 2. Selection criterion for searching

Criteria	Inclusion	Exclusion
Language	English	Non-English
Timeline	2018-2022	< 2018
Literature Type	Journal (only articles)	Journal (book chapter, proceeding)
Document Type	Article	Review, Conference Paper, Book Chapter, Book

3. DATA ABSTRACTION AND ANALYSIS

The following phase presents the outcomes of the steps from the first identification phase to the last phase of the included studies (Fig. 1). The difference between the number of publications in each database is drastic due to the number of similar articles that cause for elimination of duplicate papers. The diversity of the search engine could also explain the reason behind this. Note that the Scopus and WoS databases have different built-in search engines that may result in a different number of publications retrieved. After applying the different criteria of inclusion and exclusion, only the most relevant papers to the field of interest are kept. The data were independently extracted by synthesizing the articles for the review process.

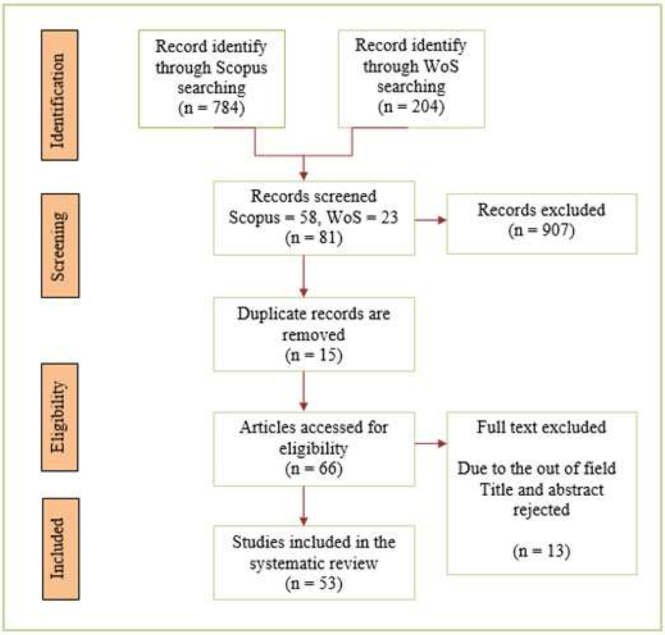
Identification
Screening
Eligbility
Included
Record identify through Scopus searching (n = 784)
Record identify through WoS searching (n = 204)
Records screened Scopus = 58, WoS = 23 (n = 81)
Records excluded (n = 907)
Duplicate records are removed (n = 15)
Articles accessed for eligibility (n = 66)
Full text excluded
Due to the out of field Title and abstract rejected
(n = 13)
Studies included in the systematic review (n = 53)

Fig. 1. Flowchart of the proposed search study

4. RESULTS

A total of 53 articles were analyzed and summarized using several main themes, including feature selection, feature extraction, machine learning, deep learning, and normalization.

4.1 Feature (Gene) Selection for Microarray Data

Microarray data analysis is practical for overcoming gene expression profile problems. The applications of microarray analysis are crucial mainly for cancer classification. Having samples of microarrays in cancer research has always been one of the most concerning issues for researchers in designing the classifiers. This task remains challenging due to the drawback of microarray data consisting

https://doi.org/10.24191/jcrinn.v9i1

of thousands of genes, but only a few are instructive (Yu et al., 2022). The vast amount of genetic data makes it difficult to identify biological markers. Although machine learning is commonly used to detect these markers, the performance depends heavily on the size and data quality available.

On top of that, the microarray data are computationally complex, and the genes are both directly and indirectly correlated, which necessitates efficient feature selection and classification techniques (El Kafrawy et al., 2021). For this matter, two main techniques are employed to select the informative genes: feature extraction and feature selection techniques. Feature selection approaches are an active research area that provides both dimensionality reduction and data interpretability in searching for significant and non-redundant features despite the advancement of machine learning (Roffo et al., 2021). Fig. 2 suggests the methodological procedure for the gene selection of microarray data from the data collection steps to preprocessing and classification.

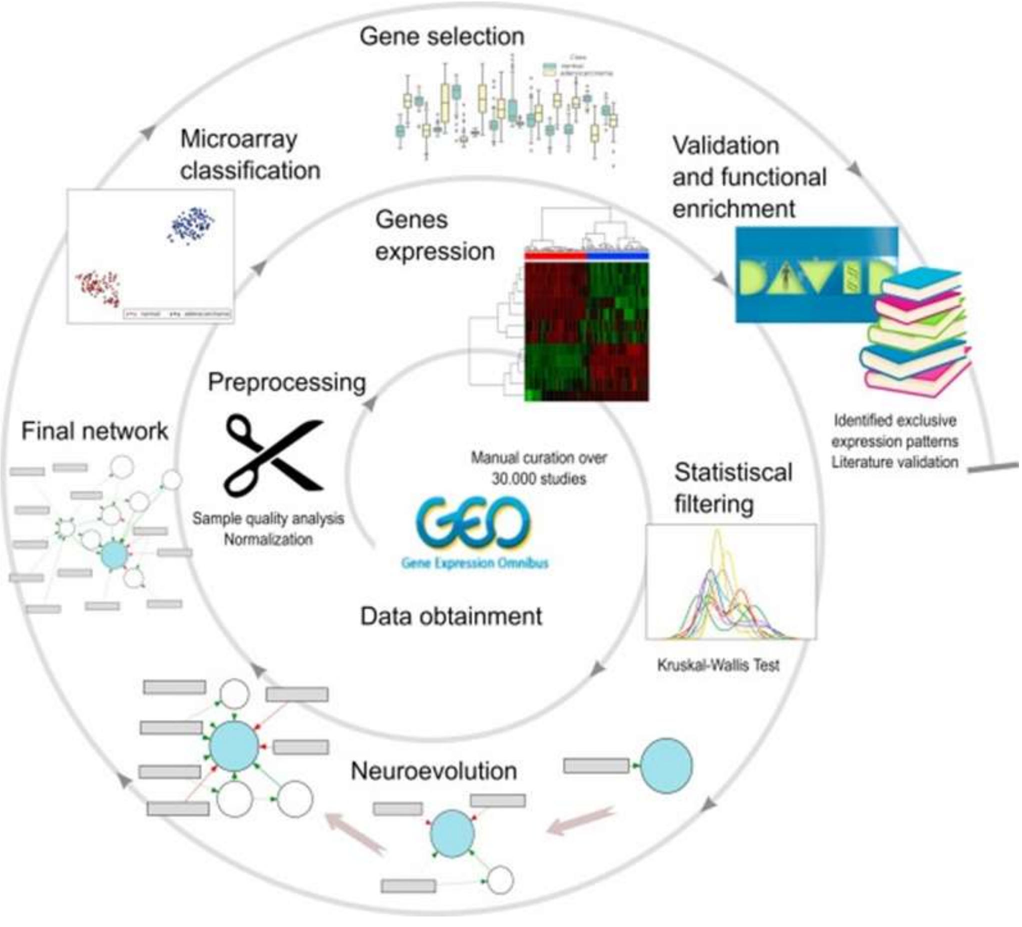
Gene selection
Microarray classification
Validation and functional enrichment
Genes expression
DAVID
Identified exclusive expression patterns Literature validation
Preprocessing
Final network
Sample quality analysis Normalization
Manual curation over 30.000 studies
GEO
Gene Expression Omnibus
Statistiscal filtering
Data obtainment
Kruskal-Wallis Test
Neuroevolution

Fig. 2. Flowchart of the feature (gene) selection methodological process for microarray data.

Source: Iochins Grisci et al. (2019)

4.2 Filter Approach

In order to alleviate the overfitting of classifiers on microarray data, a feature selection mechanism is utilized as a preprocessing procedure prior to classification. Filter feature selection is one of the main approaches widely used in microarray analysis. The filter technique selects the important feature subset by assessing each feature using some independent test before applying classifiers (Parhi et al., 2022). The feature selection process is conducted during the preprocessing step without incorporating any specific learning model (Sharifai & Zainol, 2021). Other than that, filter methods work independently from any classifier and focus on the intrinsic properties of the data (e.g., correlation, variance, locality, Information Gain (InG)) to evaluate the value of each feature subset (i.e., distances between classes or statistical dependencies) (Alshamlan, 2021; Song et al., 2021). However, since the filter approach examines each feature separately, they disregard the individual performance of the feature with respect to the group, despite the fact that features in a group may have a joint effect in a machine learning task (Liu et al., 2018). Consequently, this major drawback of the filter techniques may result in highly correlated features that subsequently affect the classification accuracy (Prabhakar & Lee, 2022).

Specifically, the filter techniques that examine each feature independently are considered univariate. This is because they overlook the feature interaction problem resulting in a feature subset redundancy. Therefore, multivariate filter techniques such as minimum Redundancy-Maximum Relevance (mRMR) are much preferable compared to univariate (Song et al., 2021). The filter methods can be split into two groups which are univariate and multivariate methods. The features in the univariate method are assessed based on their association with other feature-class pairings, where the "relationship" between a feature and the class label is considered. Here, mutual information and chi-square are the common filter metrics. Meanwhile, for multivariate, the characteristics are accessed based on the discrimination performance, where the sets that are better at discriminating are more likely to provide optimal classification accuracy. However, choosing a filter strategy for gene selection is crucial since different techniques may yield different results depending on the dataset (Tripathy et al., 2022).

Despite the limitations of the filter method, this approach is still considered and frequently used as a preprocessing step to remove statistically insignificant genes. Furthermore, this method can be applied to low-dimensional rather than high-dimensional datasets. On top of that, having flexibility and low time complexity allows the filter method to be efficiently integrated with other algorithms (Ke et al., 2018). As a result, this solves the incompatibility issue between data and algorithm and eliminates the high computing cost heuristic algorithms (Bilen et al., 2020).

Numerous studies have employed filter methods for feature selection. Based on the systematic review conducted, it is observed that the filter-based feature selection is mostly integrated with other techniques to perform classification. Tripathy et al. (2022) also mentioned that finding a filter method that can extract superior features from the datasets of relevant applications is highly challenging. Therefore, applying multiple filter approaches in the preprocessing phase instead of a single one is better for classification.

Thus, Tripathy et al. (2022) have implemented a pipeline of filter-based feature reduction combining four feature ranking algorithms, which are Correlation-Based Feature Selection (CBFS), Chi-Square Test (CST), InG, and Relief Feature Selection (RFS). Note that 16 pipelines are created based on the four feature rankings to minimize the feature subset and provide more valuable features. Based on the findings, they conclude that the 5th feature reduction sequence (FR5) out of the 16 feature reduction sequences (FR1 to FR16), i.e., (CBFS → CST → RFS → InG), is the best method for feature reduction combination. They also determined that the filter approach works successfully for high-dimensional data and can increase the classifier efficiency while requiring little computational effort.

In a study by Roffo et al. (2021), they developed a fast graph-based feature filtering strategy that ranks and selects features by evaluating alternative subsets of features as pathways on a graph and works in either an unsupervised or supervised setting. The experiments are performed on 11 different benchmarks, including five DNA microarray datasets, and a comparison of the proposed framework is made with 18 alternative feature selection methods. These findings indicate that Infinite Feature Selection (Inf-FS) practically achieves better in every circumstance.

On the other hand, Alshamlan (2021) conducted an experiment utilizing four different filter methods: mRMR, Joint Mutual Information (JMI), F-score, and Double Input Symmetrical Relevance (DISR). This is to determine the best filter approach to enhance the performance of the previously proposed Firefly and Support Vector Machine (FF-SVM) algorithm. This approach is applied to five different microarray datasets, and the outcome shows that the F-score performed better than other filter methods. Note that applying the F-score filter approach with the FF algorithm has produced an outstanding result in classification performance.

Alternatively, Castillo et al. (2019) examined the transcriptomic data to identify the biomarkers for each of the previously stated forms of leukemia. To acquire the gene expression signature, they integrated two gene quantification methods: microarray and RNA-seq. Thus, this has led the researchers to design a multiclass study using samples from the four main types of leukemia to measure gene expression. Hence, with the mRMR feature selection algorithm, impressive results were obtained by considering the first ten genes in the ranking. Furthermore, the biological analysis of the small subset of ten genes finds a substantial link between nine genes and leukemia disease.

An innovative gene selection approach has been proposed by Rostami et al. (2022), incorporating the concept of Community Detection with Node Centrality (CDNC). The CDNC technique is grouped as a filter gene selection approach due to its selection search strategy that considers relevance and similarity criteria. Here, the community detection method has resulted in a positive outcome in dealing with irrelevant and redundant genes. The proposed method significantly reduces redundancy between selected genes and improves gene selection efficacy simultaneously.

Mazumder & Veilumuthu (2019) introduced an enhanced approach to filter feature selection techniques based on Joe's Normalized Mutual Information (JNMIF) for gene selection. In this study, the researchers compared the JNMIF technique with other well-known mutual information-based feature selections and evaluated them on seven benchmark microarray cancer datasets. Note that the proposed approach performed exceptionally well and greatly improved classification accuracy. They also suggested extending the proposed method by considering feature-feature redundancy.

Apart from that, Cilia et al. (2019) conducted a comparative analysis of ranking-based feature selection, namely, CST, RFS, Gain Ratio, InG, and Symmetrical Uncertainty, along with the state-of-the-art feature selection. This includes sequential forward floating search, fast-correlation-based filter, and mRMR. The comparison research has proven that the feature-selection procedure for studying the DNA microarray dataset is crucial. This approach not only reduces the complexity of the feature space but has dramatically enhanced classification performance using a small feature set.

A new filter feature selection based on criteria fusion was developed by Ke et al. (2018) to improve the classification model prediction performance. The Score-based Criteria Fusion (SCF) feature selection method is tested on five gene expression microarray datasets and three low-dimensional datasets. Subsequently, the experimental results indicate that SCF can uncover more discriminative features than other feature selection techniques when dealing with high-dimensional microarray data. Furthermore, to discover a better combination of features, using SCF as a preprocessing method with other feature selection approaches is advised.

Brankovic et al. (2019) proposed a novel multivariate filter feature selection based on distance Correlation (dCor) to tackle the high dimensionality problem and data distribution complexity. Other than that, the dCor index seems to be a highly resilient criterion regarding overfitting and redundancy. The results tested on seven microarray datasets are highly encouraging. In fact, the classifiers attain excellent accuracy levels while employing just a minimal number of features.

Noh et al. (2023) utilizes filtering techniques and a logistic regression model to extract essential features. A combination of hybrid chi-square and hybrid information gain (hybrid IG) is employed for feature selection, with logistic regression as the classifier. The results demonstrate the effectiveness of hybrid IG, especially in high-dimensional breast and prostate cancer data. The top 50 and 22 features yield the highest classification accuracies of 86.96% and 82.61%, respectively, when integrating hybrid information gain and logistic function (hybrid IG + LR) with a sample size of 75. The study suggests potential applications for multiclass classification of multidimensional medical data in the future, utilizing data from different domains.

4.3 Wrapper Approach

The significant difference between the wrapper and filter approach is that the wrapper method, in contrast, evaluated the quality of the feature subset using a classifier. They are wrapped around a specific learning approach to generate the final classifier based on the classification error estimation of the specified feature subset (Prabhakar & Lee, 2022). Note that the wrapper method is potentially more accurate than the filter method as it requires training the learning algorithm during evaluation (Parhi et al., 2022). The conceptual simplicity of the wrapper approach benefits researchers in a way that it is unnecessary to comprehend the process of selecting variables that affect induction besides allowing them to generate and test (Manita & Korbaa, 2020).

Although wrapper approaches can yield better performance for classification accuracy, they also suffer from computational inefficiency and may have difficulty analyzing gene expression data. For this reason, they are frequently employed with near-optimal search algorithms, producing satisfactory results while minimizing computing costs (Cilia et al., 2019). According to Yu et al. (2021), standard techniques that are employed for wrapper feature selection include stability selection, Recursive Feature Elimination (RFE), Genetic Algorithm (GA), K-Nearest Neighbor (KNN), and Particle Swarm Optimization (PSO). Consequently, the wrapper technique implements evolutionary or bio-inspired algorithms to drive the search process (Qasem & Saeed, 2021).

To address the feature selection problem, Cao et al. (2019) employed Multi-Objective Evolutionary Algorithms (MOEAs) and simultaneously considered classification error, feature number, and feature redundancy. Various methods, such as feature number constraint, distributed parallelism, and sample-wise parallelism, have lowered the time consumption. Compared to many state-of-the-art MOEAs, the suggested algorithms outperform them regarding optimal performance and time efficiency.

On the other hand, Luo et al. (2019) incorporate the Recursive Feature Elimination (RFE) with an Improved SVM (ISVM-RFE (FPD)) for the feature selection process, where FPD stands for F-statistic, Pearson Correlation Coefficient (PCC) and Distance Correlation Coefficient (DCC). In comparison to the existing SVM-based feature selection algorithms, ISVM-RFE (FPD) considers not only the intrinsic properties of the data but also both linear and nonlinear correlations between features. The performance of the proposed method demonstrated that ISVM-RFE (FPD) achieved better results in terms of the recall rate of positive samples (rr_p) and G-mean (G) compared to the existing SVM-based feature selection algorithms.

Baardwijk et al. (2022) utilized the sequential forward feature selection, which is one of the wrapper methods along with the Banff-Human Organ Transplant (B-HOT) for diagnosing kidney transplants. It includes sequential forward feature selection and increases the classifier's performance in predicting genes.

https://doi.org/10.24191/jcrinn.v9i1

The B-HOT with the model attained an average accuracy of 0.921 and area under the ROC curve (AUC) of 0.965 and 0.982 for non-rejection and antibody-mediated rejection (ABMR), respectively.

Note that the ovarian cancer microarray data is diagnosed through the integrated approach by Prabhakar and Lee (2020). The integrated approach involves two consecutive steps, which are the standard gene selection techniques such as Correlation Coefficient, T-Statistics, and Kruskal-Wallis Test, and optimization algorithms including Central Force Optimization (CFO), Lightning Attachment Procedure Optimization (LAPO), Genetic Bee Colony Optimization (GBCO) and Artificial Algae Optimization (AAO). Here, the Kruskal Wallis Test with GBCO and classification using the SVM-Radial Basis Function (RBF) Kernel approach produces the highest results, with a high classification accuracy of 99.48%. Similar outcomes are also achieved when incorporating the Correlation Coefficient with AAO and Logistic Regression, resulting in a 99.48% accuracy.

4.4 Embedded Approach

Embedded approaches are similar to the wrapper feature selection as they are both assigned to learning algorithms. They integrate the learning component of the model with feature selection in such a manner that searching for an ideal subset of features and building the classifiers are done simultaneously (Momenzadeh et al., 2019). This method combines the benefits and properties of both filter and wrapper approaches. Thus, to reduce dimensionality and computational efficiency, it combines the filter approach and later adopts the wrapper method to obtain high classification accuracy (Almugren & Alshamlan, 2019). The embedded approach utilized filter and wrapper methods as preprocessing steps to minimize computational time and computational cost. Although wrappers do better computationally than embedded techniques, the latter allow for classifier-specific judgments that do not fit with any other classifier (El Kafrawy et al., 2021). Unlike the wrapper technique, the embedded feature selection does not require training the new model for each feature subset. Based on Khan et al. (2019), feature selection is known to be part of the model construction throughout the embedding procedure. Hence, the embedded approach is commonly used for feature selection, including Decision Trees (DTs), LASSO, and Ridge Regression (Hamraz et al., 2021).

An example of feature selection employing an embedded approach is proposed by Yang et al. (2018). This method employed the regularization approach and robust Logistic Regression for gene selection. Note that a new sparse Logistic Regression model based on the least absolute deviation and Lq ($0<q<1$) regularization is applied to the gene expression data to reduce dimensionality and noise level. Consequently, this study utilized the GA based on the global search strategy to produce the best results. The study presents remarkable classification accuracy and absolute error rate, especially in a high-noise environment.

4.5 Hybrid Approach

Researchers in the state of art literature have extensively employed hybrid approaches. Hybrid feature selection methods often integrate two or more feature selection algorithms from distinct search strategies consecutively and progressively (Hamim et al., 2021). Other than that, the hybrid approach combines the strength of both methodologies to overcome the disadvantages of the individual approach by reducing the complexity of selecting relevant features.

Some hybrid methods discussed here tend to integrate the filter, wrapper, or embedded feature selection methods and even create a novel feature selection technique for gene selection in cancer classification. For example, the Feature Subset Selection with Optimal Adaptive Neuro-Fuzzy Inference System (FSS-OANFIS) hybrid method has an excellent performance in selecting features compared to other well-known feature selection approaches (Hilal et al., 2022). Subsequently, high classification results are obtained with the hybrid feature selection optimization model, Elastic Net (EN) as a gene selection method along with three optimizer techniques: Social Ski-Driver (SSD), Randomized Search Cross-Validation (RS), and ENCV and SVM as a classifier (Qaraad et al., 2021).

On the other hand, Prabhakar et al. (2021) adopt a methodology combining the ranking feature selection and optimization techniques. A comprehensive analysis of the lung cancer classification has resulted in the highest accuracy of 99.10% using the Relief-F test and Artificial Fish Swarm Optimization (AFSO) with the DT classifier. They suggested that future research focus on different feature selection techniques and optimization approaches with deep learning methods for effectively classifying microarray data.

A robust feature selection framework was created by Şahín and Dírí (2019) based on Long Short-Term Memory (LSTM) recurrent neural networks and trained with the Artificial Immune Recognition System (AIRS). Here, the LSTM is embedded into AIRS and evaluated on six microarray datasets for practical sequential learning problems. The experimental results of the framework indicate an improvement in the classification accuracy, proving the proposed framework's effectiveness for gene expression analysis.

Alternatively, Lu et al. (2021) proposed a unique differential expression and feature selection approach, GEOlimma, that integrates pre-existing microarray data from the Gene Expression Omnibus (GEO) based on the Limma technique for differential expression analysis. The proposed technique modifies the well-known Linear Models for microarray and Ribonucleic acid sequencing (RNA-Seq) data ("Limma") approach. Note that the GEOlimma approach efficiently identifies differential expression genes better than the standard Limma technique.

A novel feature selection approach based on Supervised Locally Linear Embedding and Spearman's Rank Correlation Coefficient (SLLE-SC^2) was developed by Xu et al. (2018) to enhance tumor classification. The outcomes indicate that the SLLE-SC^2 technique performed well in selecting the informative genes with higher specificity than comparable methods. Apart from that, the small number of features increases the algorithm's efficacy while improving microarray data interpretation.

Another research that deals with biomarkers for identifying Esophageal Squamous Cell Carcinoma (ESCC) through the serum microRNAs (miRNAs) is presented by Zheng et al. (2019). This research joined five gene selection algorithms, including the false discovery rate procedure, family-wise error rate procedure, LASSO Logistic Regression, hybrid huberized SVM, and SVM utilizing the squared-error loss with the EN penalty to choose the differentially expressed miRNAs. According to the cross-validation results, the three miRNA-based classifiers can successfully discriminate ESCC patients from healthy controls.

As an alternative to improve cancer detection, Murugesan and Balamurugan (2023) introduces a hybrid approach for selecting crucial genes related to breast cancer detection and proposes an advanced classification model called Hyper-heuristic Adaptive Universum Support Vector Machine (HH-AUSVM). The hybrid gene selection method combines Mutual Information Maximization (MIM) and Improved Moth Flame Optimization (IMFO) in two stages, effectively addressing scalability issues. The HH-AUSVM classifier integrates Adaptive Universum learning and hyper-heuristics-based parameter optimization to handle class sample imbalances. Evaluation on breast cancer gene expression datasets demonstrates significantly improved detection rates with high accuracies (95.67%, 96.52%, 97.97%, and 95.5%) and reduced processing time (4.28, 3.17, 9.45, and 6.31 seconds). This approach presents a promising solution for early identification and treatment of breast cancer patients.

Table 3. Comparison of hybrid approach for the classification of microarray data

No	Studies	Hybrid Method	Classification Method	No of Selected Genes	No of Genes	Classification Accuracy (%)	Dataset
1	Murugesan and Balamurugan (2023)	MIM IMFO	HH-AUSVM	7108	17814	95.67%	Breast Cancer
2	Al-Rajab et al. (2023)	IG+GA MRMR+PSO	SVM NB DT DNN RF KNN	22 35 68	2000 7457 22278	D1 -94.7 (DNN) D2 - 97.2 (SVM, NB, RF, KNN) D3 - 93.9 (NB, RF)	Colon Cancer
3	Noh et al. (2023)	Chi-square IG	LR	168 102	2905 12600	86.96 82.61	Breast Cancer Prostate Cancer
4	Hilal et al. (2022)	FSS	OANFIS	38 102 47 62	7129 12600 4026 2000	89.47	Leukemia, Prostate, DLBCL Stanford, and Colon Cancer
5	Qaraad et al. (2021)	EN	SVM	102 181 104 62 118 49 72	12600 12533 22283 2000 22215 7129 7129		Prostate, Lung, Breast, Colon, Breast, Breast, Leukemia
6	Prabhakar et al. (2021)	Relief-F AFSO	DT	181	12533	99.10	Lung Cancer
7	Xu et al. (2018)	PCA SLLE-SC2	SVM C4.5 NB KNN	72 62 203 102	7129 2000 12600 12600	100 (KNN)	Leukemia, Colon, Lung, Prostate
8	Zheng et al. (2019)	HHSVM SESVM	LR SVM			80 79	Esophageal squamous cell carcinoma (ESCC)
9	Tripathy et al. (2022)	Feature Ranking TOPSIS	DT LR RF KNN	40 62 98 8	7129 2000 1213 7086		Brain Tumor Colon Cancer Breast Cancer Adenoma Cancer
10	Roffo et al. (2021)	Infinite FS	CNN		2000 7129 12533 4026 6033	91.1 95.2 94.7 95.8 99.9	Colon Leukemia Lung Lymphoma Prostate
11	Alshamlan (2021)	FF-SVM	SVM	72 83 96 72 62	7129 2308 7129 7129 2000	92.7 100 97.5 99.5 92.6	Leukemia2 SRBCT Lung Leukemia1 Colon
12	Rostami et al. (2022)	CDNC	SVM	13 17 20 23 28	2000 2328 7129 10509 12600	88.89 82.79 91.16 83.91 91.82	Colon SRBCT Leukemia Prostate Lung
13	Cilia et al. (2019)	Feature Ranking	DT, RF, KNN, NN	50 10 51	17816 2000 7129	91.96 91.94 90.69	Breast Colon Leukemia

				128	12533	98.07	Lymphoma
				308	4023	81.92	Lung
				500	2190	87.91	Ovarian
14	Ke et al. (2018)	SCF	SVM, KNN		2308		SRBCT
					7129		Leukemia
					2000		Colon
					9182		Carcinomas
					10509		Prostate

4.6 Ensemble Approach

In the past several years, researchers have been drawn to ensemble feature selection approaches since they combine disparate feature subsets and may eventually better approximate the ideal subset of features. Zhao et al. (2019) proposed an ensemble model consisting of three functional modules to improve the prediction performance of colorectal cancer. Other than that, the functional model includes the mRMR method to reduce dimensionality and a hybrid sampling algorithm Random Under-Sampling and Boosting (RUSBoost) to solve the class imbalance issue. The third module focuses on the Whale Optimization Algorithm (WOA) to determine the optimal parameters of a Mixed Kernel Function (MKF) based SVM classifier. This approach has achieved a high G-means value of 93.65% with the WOA and MKF-SVM model for cancer detection.

Within the exact scope of the ensemble approach, Hengpraprohm and Jungjit (2020) presented an ensemble feature selection approach with Entropy and Signal Noise Ratio (EnSNR) for selecting relevant features. Note that the EnSNR approach has achieved a promising result and high classification ability for the prediction of breast cancer, requiring only a tiny amount of processing time to produce the EnSNR feature subset.

Alternatively, Giordano et al. (2018) proposed a methodology for characterizing and predicting gene signatures of tobacco smoke exposure from human blood gene expression data. To be precise, the gene selection approach employed in this study consists of three different techniques, including Extra Trees, LASSO using the Least Angle Regression (LARS) algorithm, and the RFE-SVM estimator. The signature linked with smoking effects is robustly depicted by the comparative study of disease association, pathway analysis, and gene ontology terms enrichment of the signatures derived with the three gene selection approaches (Extra-Trees, LASSO-LARS, and RFE-SVM). By combining several gene selection methods, this study can generate gene signatures with high classification performance, as well as novel findings in genetic signatures for biomarkers of tobacco smoking exposure.

4.7 Feature Extraction

The growth of microarray datasets sparked a new line of research in bioinformatics and machine learning. However, it is critical to emphasize the importance of data preprocessing prior to classification tasks. Applying the classification method without considering data noise may influence the biological significance of the selected genes, which could eventually reduce the quality of the outcome (Iochins Grisci et al., 2019). Raw data, such as microarrays, usually contain noise that should be removed to boost classification accuracy. Therefore, the right approach for dimensionality reduction is crucial, especially in high-dimensional gene expression data.

Although most studies adopted the feature selection approach for reducing dimensions, some studies incorporated the feature extraction technique and hybridized the approach for better classification performance. In a study by Zhang et al. (2018), they integrate feature extraction, Principal Component Analysis (PCA), and autoencoder feature selection techniques to identify relevant features from gene

https://doi.org/10.24191/jcrinn.v9i1

expression profiles. The AdaBoost classifier, an ensemble classifier with boosting algorithm, is utilized to accurately identify breast tumors to achieve a more generalized classification outcome. Hence, the analysis suggested that the proposed method has excellent generalization ability and substantially enhances the prediction performance.

4.8 Deep Learning Algorithm

Recently, deep learning methods have achieved outstanding performance over machine learning algorithms in classifying gene expression data. The preprocessing step has been one of the prominent roles in predicting cancer types. Nevertheless, the right approach for classification would help correctly identify various microarrays. Several machine-learning techniques have indicated significant improvements in cancer classification. Nonetheless, certain issues with this technique make the classification process challenging (Alshareef et al., 2022). The deep learning technique is a subset of machine learning that uses a layered structure to create sophisticated modules that can comprehend complex data. With these advances, deep learning algorithms can be adapted to multiple domains in conventional machine learning approaches, including speech recognition and image classification (Zhang et al., 2018).

In a study utilizing deep learning techniques for detecting cancer microarray data, Othman et al. (2020) confirmed that applying a Deep Neural Network (DNN) algorithm can uncover the genes that contribute to cancer diseases in both real-world and clinical applications. Subsequently, the DNN classifier ability is validated through the performance measure rates at which the DNN algorithm is able to analyze gene correlations. From these experiments, it can be inferred that developing a robust classifier model is crucial in a gene selection process to further biological research.

On the other hand, Almarzouki (2021) proposed a Convolution Neural Network (CNN) algorithm to predict cancer types. The CNN classifiers were trained to classify different tumors without labeling them. The experimental findings with the CNN model achieved an accuracy rate of 93.43% using the k-fold cross-validation technique. As a result of this analysis, the possibility of discovering particular medications to treat early diagnosis is made more accessible.

With the motivation of the Artificial Intelligence (AI) technique, the detection rate can be improved using metaheuristic algorithms to overcome the dimensionality curse. Alshareef et al. (2022) built an AI-based Feature Selection with a Deep Learning model for Prostate Cancer Diagnosis (AIFSDL-PCD). The findings demonstrated that the optimal DNN algorithm managed to produce satisfactory outcomes. However, the AIFSDL-PCD technique has resulted in superior performance with higher accuracy of 0.9719. From the following Fig. 3, it is observed that the AIFSDL-PCD technique is a proficient tool for prostate cancer detection and classification.

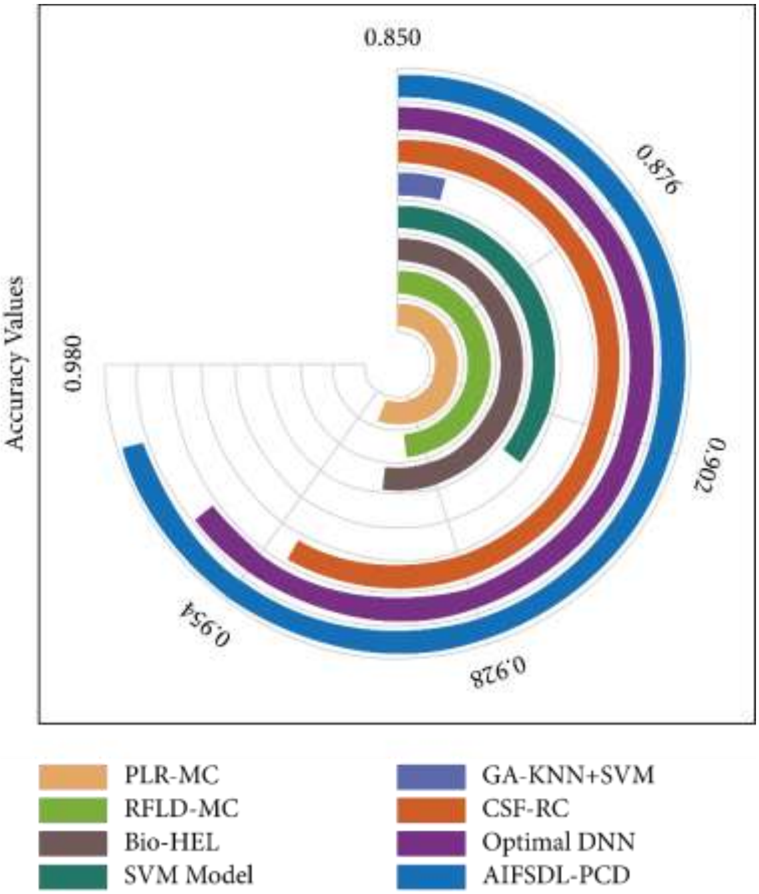

Fig. 3. Comparative analysis of AIFSDL-PCD approach with existing techniques.

Source: Alshareef et al. (2022)

Alternatively, Zhang et al. (2018) developed an unsupervised feature learning framework to detect various traits from gene expression profiles by incorporating a PCA approach with an auto-encoder neural network. The novel approach with deep learning provides superior prediction ability in predicting clinical outcomes of breast cancer. Based on the analysis results, they concluded that the features extracted automatically by the neural network demonstrated an excellent generalization ability which explicitly increased the prediction performance. Consequently, they emphasized that the data are less prone to overfitting due to the complex structure of deep learning algorithms.

4.9 Machine Learning Algorithm

Despite the rise of the deep learning approach, machine learning still gained considerable attention from researchers in many fields, especially gene expression and genomics. Machine learning algorithms have been widely utilized on microarray data with complementary objectives involving data classification and gene selection. Over the past few years, machine learning has been applied to several clinical diagnostics and evaluated with various algorithms. In most cases, SVM has been considered the best approach for microarray data classification since they consistently produce the best result compared to other algorithms (Iochins Grisci et al., 2019).

SVM has an advantage in dealing with high-dimensional datasets, making it an ideal classifier (Sun et al., 2018). More remarkably, the F-score and FF feature selection method and the SVM algorithm have achieved high accuracy, between 94% and 100%, for all five microarray datasets (Almugren & Alshamlan, 2019). SVM has long been praised for its superior classification efficiency and intrinsic feature selection ability. Using the EN and Probabilistic SVM (EPSVM) for the microarray data indicates that SVM delivered better performance than conventional methods for all datasets with accuracy above 75%. Note that the highest accuracy was obtained at 99.58% on the small round blue cell tumor (SRBCT) dataset (Yuan et al., 2020). However, different datasets and preprocessing approaches may yield different classification results.

Apart from that, Babichev and Škvor (2020) employed four binary classifiers, including Logistic Regression (GLM), Support Vector Machine (SVM), Decision Tree (CART), and Random Forest (RF), for the classification of lung cancer disease. The experimental results revealed that the Logistic Regression classifier is inefficient for classifying the high-dimensional vectors of microarray data. Significantly better results are obtained with CART and RF when compared with SVM in classifying the cancer data.

Similarly to Angulo (2018), they utilized machine learning algorithms such as Ripper-k, C4.5, SVM, and Naïve Bayes (NB) to classify microarray cancer datasets. This study achieves the highest processing time and accuracy when utilizing Ripper-k and C4.5 as classifiers with Probabilistic Attribute-Value Integration for Class Distinction (PAVICD) feature selection approach. Nevertheless, SVM obtained the best result utilizing the genetic bee colony wrapper algorithm for gene selection.

Al-Rajab et al. (2023) introduces a Hybrid Machine Learning Feature Selection Model (HMLFSM) to enhance the classification of colon cancer genes. This paper focuses on the importance of early detection of colon cancer and the limitations of traditional invasive methods. The model incorporates a multifilter hybrid approach, involving Information Gain (IG) and Genetic Algorithms (GA) in a two-phase feature selection, along with minimum Redundancy Maximum Relevance (mRMR) coupled with Particle Swarm Optimization (PSO). Testing on three colon cancer genetic datasets demonstrates significant accuracy improvements (95%, ~97%, and ~94% accuracies for datasets 1, 2, and 3, respectively) compared to other models. The results highlight the effectiveness of the proposed approach in improving classification accuracy by selecting relevant genes and eliminating irrelevant ones, emphasizing the importance of selective input feature extraction in enhancing predictive performance for colon cancer gene analysis.

4.10 Normalization

One of the crucial phases of preprocessing gene expression data is normalization. Normalization is a technique for standardizing the range of independent data features. Note that most features are available in continuous values, where each feature is measured on a distinct scale and has a varied range of potential values. Since continuous gene expression values are present in microarray datasets, preprocessing the expression data with an optimum normalization strategy is advised (Ramasamy & Kandhasamy, 2018). Similarly, Gálvez et al. (2020) stated that each microarray series requires a normalization procedure since it was processed from various platforms.

Ramasamy and Kandhasamy (2018) employed two fuzzy normalization techniques, Fuzzy Set (FS) and Intuitionistic FS (IFS), to normalize the microarray datasets. Both fuzzy normalization approaches illustrate a significant improvement in the accuracy rate. This indicates that the normalization technique is essential to enhance the efficiency of gene selection. On the other hand, Shibata et al. (2020) utilized the preprocessing approach using min-max normalizations to eliminate experimental bias and rescale the value range among features. The normalization approach significantly contributes to easing the gene selection process, improving the data quality subsequently. The plots of dataset H indicated that samples from

https://doi.org/10.24191/jcrinn.v9i1

different classes were separated using min-max normalization (Fig. 4). Yuan et al. (2020) also suggested further study of the impact of various normalization techniques on the class-feature-based multiclass classification approach.

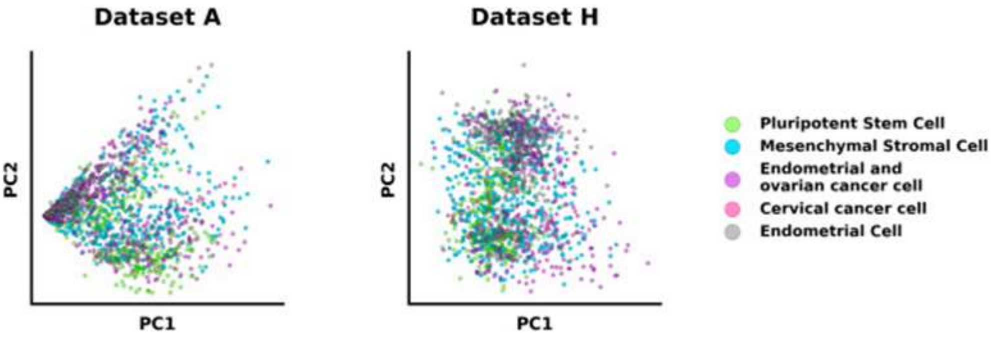

Fig. 4. PCA plots of dataset A (the raw dataset) and dataset H (min-max normalization).

Source: Shibata et al. (2020)

5. DISCUSSION

This study has provided a comprehensive overview of the outcomes obtained using microarray datasets. Based on the studies reviewed above, it is noticed that most of the microarray data used in the study are related to gene expression cancer data. Regarding microarray, gene expression data is known to have such a vast dimension with its complexity in analyzing gene expression features. Since microarray data is known to have errors and biases, the normalization procedure is vital in the earlier stage of data preprocessing. The outcomes of the subsequent analysis are highly dependent on the preprocessing steps. It is also important to highlight that cancer classification is one of the most significant uses of microarray data analysis.

Most of the studies reviewed widely apply machine learning algorithms such as DT, NB, RF, SVM, and KNN. Moreover, it is observed that the SVM algorithm has been used frequently by several researchers. The performance of SVM seems to yield better results for classification. Besides, the deep learning approach has proliferated and contributed significantly to the classification analysis. Other than that, the performance of deep learning seems to be far better than machine learning algorithms. However, both approaches have demonstrated outstanding performance by producing a small number of selected genes when applied to microarray classification problems.

Many authors have introduced hybrid methods as well as ensemble techniques to improve classification performance. This can be supported by Giordano et al. (2018), that the ensemble approach is more robust than a single selector, providing useful insight for researchers and practitioners. Based on this literature review, hybrid methods have been observed for their ability to achieve good accuracy ranging from 80% to 100%, while significantly reducing the number of features. Hence, feature selection is believed to be essential to largely boost classification accuracy.

In conclusion, many novel and hybrid methods have been developed to produce a good performance in terms of accuracy and computational efficiency. The hybrid method has gained immense popularity as it incorporates multiple types of feature selection methods. Consequently, hybrid methods are also proven effective in reducing dimensions and selecting relevant features. The classification technique is as important as the feature selection method in providing optimal classification performance.

In summary, by using hybrid method, it will boost the cancer data adequacy with only relevant features selected and reduced dimensions, hence is expected in improving the cancer diagnosis, treatment or any future research dealing with microarray data.

6. CONCLUSION

This paper reviews relevant studies on dimensionality reduction involving feature extraction and feature selection for the classification of microarray data. This study highlighted a useful preprocessing tool that may benefit future gene expression data analytics studies. A wide range of this review study deepens researchers' understanding of the importance of applying feature selection methodologies while processing microarray data. Based on the literature review, many in-depth studies have been conducted throughout the years to automate microarray data analysis. In this paper, we have reviewed studies on feature selection for microarray data, with a focus on metaheuristic-based hybrid techniques. To improve feature selection, many studies proposed hybrid techniques by combining two or more feature selection approaches, such as filtering and wrapping. Robust feature selection has resulted in reliable and high classification accuracy with a small number of genes selected, which is beneficial in the screening and diagnosis of microarray data.

Therefore, future researchers are suggested to extend the combination of other feature selection methods and classifiers in solving the multi-feature and multi-class classifications problem encountered in practice. Hence, this study is expected to be applicable to Data Science and Medical Sciences for the opportunity for cross-disciplinary collaborations. Besides, the substantial number of related articles reviewed evaluates and validates previous methodologies and tools proposed.

7. CONFLICT OF INTEREST STATEMENT

The authors declare that this research was conducted in the absence of any self-benefits or financial conflicts that could potentially influence the research work and declare the absence of conflicting interests with the funders.

8. AUTHORS’ CONTRIBUTIONS

Sharifah Nadia Syed Hasan conceptualised the central research idea, adopted the theoretical framework from existing research, designed the research, carried out the research, wrote and revised the article. Noor Wahida Jamil supervised research progress, anchored the review, revisions and approved the article submission.

9. REFERENCES

Al-Rajab, M., Lu, J., Xu, Q., Kentour, M., Sawsa, A., Shuweikeh, E., Joy, M., & Arasaradnam, R. (2023). A hybrid machine learning feature selection model—HMLFSM to enhance gene classification applied to multiple colon cancers dataset. *PLOS ONE, 18*(11), e0286791. https://doi.org/10.1371/journal.pone.0286791

Almarzouki, H. Z. (2021). Deep-learning-based cancer profiles classification using gene expression data profile. *Journal of Healthcare Engineering, 2022*. https://doi.org/https://doi.org/10.1155/2022/4715998

Almugren, N., & Alshamlan, H. M. (2019). New bio-marker gene discovery algorithms for cancer gene expression profile. *IEEE Access, 7,* 136907–136913. https://doi.org/10.1109/ACCESS.2019.2942413

Alshamlan, H. M. (2021). An effective filter method towards the performance improvement of FF-SVM algorithm. *IEEE Access*, *9*, 140835–140840. https://doi.org/10.1109/ACCESS.2021.3119233

Alshareef, A. M., Alsini, R., Alsieni, M., Alrowais, F., Marzouk, R., Abunadi, I., & Nemri, N. (2022). Optimal deep learning enabled prostate cancer detection using microarray gene expression. *Journal of Healthcare Engineering, 2022*. https://doi.org/https://doi.org/10.1155/2022/7364704

Angulo, A. P. (2018). Gene selection for microarray cancer data classification by a novel rule-based algorithm. *Information*, *9*(1), 6. https://doi.org/https://doi.org/10.3390/info9010006

Baardwijk, M. van, Cristoferi, I., Ju, J., Varol, H., Minnee, R. C., Reinders, M. E. J., Li, Y., Stubbs, A. P., & Groningen, M. C. C. (2022). A decentralized kidney transplant biopsy classifier for transplant rejection developed using genes of the Banff-Human organ transplant panel. *Frontiers in Immunology*, *13*. https://doi.org/https://doi.org/10.3389/fimmu.2022.841519

Babichev, S., & Škvor, J. (2020). Technique of gene expression profiles extraction based on the complex use of clustering and classification methods. *Diagnostics*, *10*(8). https://doi.org/https://doi.org/10.3390/diagnostics10080584

Bilen, M., H. Isik, A., & Yigit, T. (2020). A new hybrid and ensemble gene selection approach with an enhanced genetic algorithm for classification of microarray gene expression values on leukemia cancer. *International Journal of Computational Intelligence Systems*, *13*(1), 1554–1556. https://doi.org/https://doi.org/10.2991/ijcis.d.200928.001

Brankovic, A., Hosseini, M., & Piroddi, L. (2019). A distributed feature selection algorithm based on distance correlation with an application to microarrays. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(6), 1802–1815. https://doi.org/10.1109/TCBB.2018.2833482

Cao, B., Jianwei Zhao, Yang, P., Yang, P., Liu, X., Qi, J., Simpson, A., Elhoseny, M., Mehmood, I., & Muhammad, K. (2019). Multiobjective feature selection for microarray data via distributed parallel algorithms. *Future Generation Computer Systems*, *100*, 952–981. https://doi.org/https://doi.org/10.1016/j.future.2019.02.030

Castillo, D., Galvez, J. M., Herrera, L. J., Rojas, F., Valenzuela, O., Caba, O., Prados, J., & Rojas, I. (2019). Leukemia multiclass assessment and classification from Microarray and RNA-seq technologies integration at gene expression level. *PLoS One*, *14*(2). https://doi.org/https://doi.org/10.1371/journal.pone.0212127

Cilia, N. D., Stefano, C. De, Fontanella, F., Raimondo, S., & Freca, A. S. di. (2019). An experimental comparison of feature-selection and classification methods for microarray datasets. *Information*, *10*(3), 109. https://doi.org/https://doi.org/10.3390/info10030109

El Kafrawy, P., Fathi, H., Qaraad, M., Kelany, A. K., & Chen, X. (2021). An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data. *IEEE Access*, *9*, 155353–155369. https://doi.org/10.1109/ACCESS.2021.3123090

Gálvez, J. M., Castillo-Secilla, D., Herrera, L. J., Valenzuela, O., Caba, O., Prados, J. C., Ortuño, F. M., &

Rojas, I. (2020). Towards improving skin cancer diagnosis by integrating microarray and RNA-Seq Datasets. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 2119–2130. https://doi.org/10.1109/JBHI.2019.2953978

Giordano, M., Tripathi, K. P., & Guarracino, M. R. (2018). Ensemble of rankers for efficient gene signature extraction in smoke exposure classification. *BMC Bioinformatics*, *19*(48). https://doi.org/https://doi.org/10.1186/s12859-018-2035-3

Hamim, M., Mouden, I. El, Ouzir, M., Moutachaouik, H., & Hain, M. (2021). A novel dimensionality reduction approach to improve microarray data classification. *IIUM Engineering Journal*, *22*(1). https://doi.org/https://doi.org/10.31436/iiumej.v22i1.1447

Hamraz, M., Gul, N., Raza, M., Khan, D. M., Khalil, U., Zubair, S., & Khan, Z. (2021). Robust proportional overlapping analysis for feature selection in binary classification within functional genomic experiments. *PeerJ Computer Science*. https://doi.org/https://doi.org/10.7717/peerj-cs.562

Hengpraprohm, S., & Jungjit, S. (2020). Ensemble feature selection for breast cancer classification using microarray data. *Intelegencia Artificial*, *23*(65), 100–114. https://doi.org/https://doi.org/10.4114/intartif.vol23iss65pp100-114

Hilal, A. M., Malibari, A. A., Obayya, M., Alzahrani, J. S., Alamgeer, M., Mohamed, A., Motwakel, A., Yaseen, I., Hamza, M. A., & Zamani, A. S. (2022). Feature subset selection with optimal adaptive neuro-fuzzy systems for bioinformatics gene expression classification. *Computational Intelligence and Neuroscience*. https://doi.org/https://doi.org/10.1155/2022/1698137

Iochins Grisci, B., Cesar Feltes, B., & Dorn, M. (2019). Neuroevolution as a tool for microarray gene expression pattern identification in cancer research. *Journal of Biomedical Informatics*, *89*, 122–133. https://doi.org/https://doi.org/10.1016/j.jbi.2018.11.013

Ke, W., Wu, C., Wu, Y., & Xiong, N. N. (2018). A New filter feature selection based on criteria fusion for gene microarray data. *IEEE Access*, *6*, 61065–61076. https://doi.org/10.1109/ACCESS.2018.2873634

Khan, Z., Naeem, M., Khalil, U., Khan, D. M., Aldahmani, S., & Hamraz, M. (2019). Feature selection for binary classification within functional genomics experiments via interquartile range and clustering. *IEEE Access*, *7*, 78159–78169. https://doi.org/10.1109/ACCESS.2019.2922432

Lee, J., Choi, I. Y., & Jun, C.-H. (2021). An efficient multivariate feature ranking method for gene selection in high-dimensional microarray data. *Expert Systems with Applications*, *166*, 113971. https://doi.org/10.1016/j.eswa.2020.113971

Liu, X.-Y., Liang, Y., Wang, S., Yang, Z.-Y., & Ye, H.-S. (2018). A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. *IEEE Access*, *6*, 22863–22874. https://doi.org/10.1109/ACCESS.2018.2818682

Lu, L., Townsend, K. A., & Jr., B. J. D. (2021). GEOlimma: Differential expression analysis and feature selection using pre-existing microarray data. *BMC Bioinformatics*, *22*(44). https://doi.org/https://doi.org/10.1186/s12859-020-03932-5

Luo, K., Wang, G., Li, Q., & Tao, J. (2019). An improved SVM-RFE based on F -Statistic and mPDC for gene selection in cancer classification. *IEEE Access*, *7*, 147617–147628.

https://doi.org/10.1109/ACCESS.2019.2946653

Manita, G., & Korbaa, O. (2020). Binary political optimizer for feature selection using gene expression data. *Computational Intelligence and Neuroscience*. https://doi.org/https://doi.org/10.1155/2020/8896570

Mazumder, D. H., & Veilumuthu, R. (2019). An enhanced feature selection filter for classification of microarray cancer data. *ETRI Journal*, *41*(3), Ramachandran Veilumuthu. https://doi.org/https://doi-org.ezaccess.library.uitm.edu.my/10.4218/etrij.2018-0522

Momenzadeh, M., Sehhati, M., & Rabbani, H. (2019). A novel feature selection method for microarray data classification based on hidden Markov model. *Journal of Biomedical Informatics*, *95*, 103213. https://doi.org/https://doi.org/10.1016/j.jbi.2019.103213

Murugesan, V., & Balamurugan, P. (2023). Breast cancer classification by gene expression analysis using hybrid feature selection and hyper-heuristic adaptive universum support vector machine. *International Journal of Electrical and Computer Engineering Systems*, *14*(3). https://doi.org/10.32985/IJECES.14.3.1

Noh, S. S. M., Ibrahim, N., Mansor, M. M., & Yusoff, M. (2023). Hybrid filtering methods for feature selection in high-dimensional cancer data. *International Journal of Electrical and Computer Engineering*, *13*(6). https://doi.org/10.11591/ijece.v13i6.pp6862-6871

Othman, M. S., Raja Kumaran, S., & Mi Yusuf, L. (2020). Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.3029890

Parhi, P., Bisoi, R., & Dash, P. K. (2022). Influential gene selection from high-dimensional genomic data using a bio-inspired algorithm wrapped broad learning system. *IEEE Access*, *10*, 49219–49232. https://doi.org/10.1109/ACCESS.2022.3170038

Prabhakar, S. K., & Lee, S.-W. (2020). An integrated approach for ovarian cancer classification with the application of stochastic optimization. *IEEE Access*, *8*, 127866–127882. https://doi.org/10.1109/ACCESS.2020.3006154

Prabhakar, S. K., & Lee, S.-W. (2022). Transformation based tri-level feature selection approach using wavelets and swarm computing for prostate cancer classification. *IEEE Access*, *8*, 127462–127476. https://doi.org/10.1109/ACCESS.2020.3006197

Prabhakar, S. K., Rajaguru, H., & Won, D.-O. (2021). A holistic performance comparison for lung cancer classification using swarm intelligence techniques. *Journal of Healthcare Engineering*. https://doi.org/https://doi.org/10.1155/2021/6680424

Qaraad, M., Amjad, S., Manhrawy, I. I. M., Fathi, H., Hassan, B. A., & Kafrawy, P. El. (2021). A Hybrid Feature selection optimization model for high dimension data classification. *IEEE Access*, *9*, 42884–42895. https://doi.org/10.1109/ACCESS.2021.3065341

Qasem, S. N., & Saeed, F. (2021). Hybrid feature selection and ensemble learning methods for gene selection and cancer classification. *International Journal of Advanced Computer Science and Applications*, *12*(2). https://doi.org/10.14569/IJACSA.2021.0120225

Ramasamy, P., & Kandhasamy, P. (2018). Effect of intuitionistic fuzzy normalization in microarray gene selection. *Turkish Journal of Electrical Engineering and Computer Sciences*, *6*(3), 1141–1152. https://doi.org/10.3906/elk-1708-105

Roffo, G., Melzi, S., Castellani, U., Vinciarelli, A., & Cristani, M. (2021). Infinite feature selection: A graph-based feature filtering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(12), 4396–4410. https://doi.org/10.1109/TPAMI.2020.3002843

Rostami, M., Forouzandeh, S., Berahmand, K., Soltani, M., Shahsavari, M., & Oussalah, M. (2022). Gene selection for microarray data classification via multi-objective graph theoretic-based method. *Artificial Intelligence in Medicine*, *123*. https://doi.org/https://doi.org/10.1016/j.artmed.2021.102228

Şahín, C. B., & Dírí, B. (2019). Robust feature selection with LSTM recurrent neural networks for artificial immune recognition system. *IEEE Access*, *7*, 24165–24178. https://doi.org/10.1109/ACCESS.2019.2900118

Sharifai, A. G., & Zainol, Z. B. (2021). Multiple filter-based rankers to guide hybrid grasshopper optimization algorithm and simulated annealing for feature selection with high dimensional multi-class imbalanced datasets. *IEEE Access*, *9*, 74127–74142. https://doi.org/10.1109/ACCESS.2021.3081366

Shibata, M., Okamura, K., Yura, K., & Umezawa, A. (2020). High-precision multiclass cell classification by supervised machine learning on lectin microarray data. *Regenerative Therapy*, *15*, 195–201. https://doi.org/https://doi.org/10.1016/j.reth.2020.09.005

Song, S., Chen, X., Tang, Z., & Todo, Y. (2021). A two-stage method based on multiobjective differential evolution for gene selection. *Computational Intelligence and Neuroscience*. https://doi.org/https://doi.org/10.1155/2021/5227377

Sun, L., Zhang, X., Xu, J., Wang, W., & Liu, R. (2018). A gene selection approach based on the fisher linear discriminant and the neighborhood rough set. *Bioengineered*, *9*(1), 144–151. https://doi.org/https://doi-org.ezaccess.library.uitm.edu.my/10.1080/21655979.2017.1403678

Tripathy, J., Dash, R., Pattanayak, B. K., Mishra, S. K., Mishra, T. K., & Puthal, D. (2022). Combination of reduction detection using TOPSIS for gene expression data analysis. *Big Data and Cognitive Computing*, *6*(1), 24. https://doi.org/https://doi.org/10.3390/bdcc6010024

World Health Organization. (2022). *Cancer*. Retrieved from, https://www.who.int/news-room/fact-sheets/detail/cancer

Xu, J., Mu, H., Wang, Y., & Huang, F. (2018). Feature genes selection using supervised locally linear embedding and correlation coefficient for microarray classification. *Computational and Mathematical Methods in Medicine*. https://doi.org/https://doi.org/10.1155/2018/5490513

Yang, Z.-Y., Liang, Y., Zhang, H., Chai, H., Zhang, B., & Peng, C. (2018). Robust sparse logistic regression with the $(0 < q < 1)$ regularization for feature selection using gene expression data. *IEEE Access*, *6*, 68586–68595. https://doi.org/10.1109/ACCESS.2018.2880198

Yu, K., Huang, M., Chen, S., Feng, C., & Li, W. (2022). GSEnet: Feature extraction of gene expression data and its application to Leukemia classification. *Mathematical Biosciences and Engineering*, *19*(5), 4881–4891. https://doi.org/10.3934/mbe.2022228

Yu, K., Xie, W., Wang, L., & Li, W. (2021). ILRC: A hybrid biomarker discovery algorithm based on improved L1 regularization and clustering in microarray data. *BMC Bioinformatics*, *22*(514). https://doi.org/https://doi.org/10.1186/s12859-021-04443-7

Yuan, L., Sun, Y., & Huang, G. (2020). Using class-specific feature selection for cancer detection with gene expression profile data of platelets. *Sensors*, *20*(5). https://doi.org/https://doi.org/10.3390/s20051528

Zhang, D., Zou, L., Zhou, X., & He, F. (2018). Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *IEEE Access*, *6*, 28936–28944. https://doi.org/10.1109/ACCESS.2018.2837654

Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D., & Lyu, C. (2019). Whale optimized mixed kernel function of support vector machine for colorectal cancer diagnosis. *Journal of Biomedical Informatics*, *92*, 103124. https://doi.org/https://doi.org/10.1016/j.jbi.2019.103124

Zheng, D., Ding, Y., Ma, Q., Zhao, L., Guo, X., Shen, Y., He, Y., Wei, W., & Liu, F. (2019). Identification of serum MicroRNAs as novel biomarkers in esophageal squamous cell carcinoma using feature selection algorithms. *Frontiers in Oncology*. https://doi.org/https://doi.org/10.3389/fonc.2018.00674