frontiers in Genetics

Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues and Future Research Directions

Nivedhitha Mahendran[1], Durai Raj[1], Kathiravan Srinivasan[1], Chuan-Yu Chang[2*]

[1]VIT University, India, [2]National Yunlin University of Science and Technology, Taiwan

Submitted to Journal:
Frontiers in Genetics

Specialty Section:
Computational Genomics

ISSN:
1664-8021

Article type:
Review Article

Received on:
08 Sep 2020

Accepted on:
29 Oct 2020

Provisional PDF published on:
29 Oct 2020

Frontiers website link:
www.frontiersin.org

Citation:
Mahendran N, Raj D, Srinivasan K and Chang C(2020) Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues and Future Research Directions. *Front. Genet.* 11:1468. doi:10.3389/fgene.2020.603808

Machine Learning Based Computational Gene Selection Models: A Survey, Performance Evaluation, Open Issues and Future Research Directions

Nivedhitha Mahendran[1], Durai Raj Vincent P M[1]*, Kathiravan Srinivasan[1], Chuan-Yu Chang[2]*

[1]School of Information Technology and Engineering, Vellore Institute of Technology (VIT), Vellore 632014, India

[2]Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin 64002, Taiwan

*** Correspondence:**
Durai Raj Vincent P M; Chuan-Yu Chang
pmvincent@vit.ac.in; chuanyu@yuntech.edu.tw

Keywords: Gene Selection, Machine Learning, Microarray Gene Expression, Supervised Gene Selection, Unsupervised Gene Selection.

Abstract

Gene Expression is the process of determining the physical characteristics of living beings by generating the necessary proteins. Gene expression takes place in two steps, translation and transcription. It is the flow of information from DNA to RNA with enzymes' help, and the end product is proteins and other biochemical molecules. Many technologies can capture Gene Expression from the DNA or RNA. One such technique is Microarray DNA. Other than being expensive, the main issue with Microarray DNA is that it generates high-dimensional data with minimal sample size. The issue in handling such a heavyweight dataset is that the learning model will be over-fitted. This problem should be addressed by reducing the dimension of the data source to a considerable amount. In recent years, Machine Learning has gained popularity in the field of genomic studies. In the literature, many Machine Learning-based Gene Selection approaches have been discussed, which were proposed to improve dimensionality reduction precision. This paper does an extensive review of the various works done on Machine Learning-based gene selection in recent years, along with its performance analysis. The study categorizes various feature selection algorithms under Supervised, Unsupervised, and Semi-supervised learning. The works done in recent years to reduce the features for diagnosing tumors are discussed in detail. Furthermore, the performance of several discussed methods in the literature is analyzed. This study also lists out and briefly discusses the open issues in handling the high-dimension and less sample size data.

1 Introduction

Deoxy-ribonucleic Acid (DNA) is a hereditary material containing the genetic information, usually found in the cell's nucleus. The information inside the DNA is made up of a code consisting of four bases, namely, Adenine, Guanine, Cytosine, and Thymine. Adenine pairs with Thymine and Cytosine with Guanine to form base pairs. The base pairs, along with their respective sugar and phosphate molecules, form a Nucleotide. The Nucleotide forms a double helical structure, which looks like a ladder. Gene is the fundamental unit of heredity and is built-up of DNA. Genes are responsible for determining characteristics such as height, color, and many others. Some of the genes manufacture

proteins, and some do not. According to the Human Genome Project, there are approximately around 25,000 genes in humans.

There are two copies of genes in every human; one passed on from the parent; almost all the genes are the same, except a few, less than 1% called the Alleles. They determine the unique physical features of a person. Genes manufacture proteins, and proteins, in turn, say what the cell should do (cell functions). The flow starts with DNA, RNA, and then the proteins. The flow of information determines the type of proteins being produced. The process in which the information contained in DNA is transformed into instructions to form proteins and other biochemical molecules is called gene expression. Gene expression assists the cells to react appropriately to the changing environment. The gene expression involves two critical steps in manufacturing the proteins, Transcription and Translation [1].

- Transcription: The DNA present in the gene will be copied to form an RNA known as the messenger RNA (mRNA). RNA is similar to DNA; however, it has a single-strand, and instead of Thymine, it has Uracil (U).
- Translation: The messages carried from the transcription by the mRNA will be read by the transfer RNA (tRNA) in the Translation phase. The mRNA can read three letters at a time, which constitutes one Amino acid (Amino acids are the building blocks of proteins)

Proteins play a significant role in cell functioning. Gene expression controls everything, such as when to produce protein, when not to, volume, i.e., increasing or decreasing the amount, etc. It is a kind of on/off switch. When this process does not happen as it is supposed to be, genetic disorders, tumors occur. A detailed study of the gene expression will help find the essential biomarkers that cause genetic disorders and tumors.

There are many techniques available to capture the gene expressions such as Northern blot, RNA protection assay, Reverse Transcription – Polymerase Chain Reaction (RT - PCR), Serial Analysis of Gene Expression (SAGE), Subtractive Hybridization, DNA Microarrays, Second Generation Sequencing (NGS) and many others. Among these, the most widely used these days is DNA Microarray [1, 2]. The DNA microarray technology manages to capture gene expressions of thousands of genes simultaneously. However, the Microarray result is enormous, with a high dimension, which makes the analysis challenging. Thus, it is necessary to perform gene selection to handle the high dimensional problem by removing the redundant and irrelevant genes. There are many computation techniques used in the field of bioinformatics been carried out over the years, such as Pattern Recognition, Data Mining, and many others to manage the high dimensional issue, yet ineffective [1].

Hence, in recent years, Machine Learning, which is a part of Artificial Intelligence, has gained the researchers' attention in genomics and gene expression. Machine Learning is the part of Data Science; its primary purpose is to enable a model to train and learn to make decisions on its own in the future. Machine Learning is commonly categorized as Supervised, Unsupervised, and Semi-supervised or Semi-unsupervised learning. The Supervised involves the labeled data; unsupervised learning involves unlabeled data, and the Semi-supervised or Semi-unsupervised involves handling both labeled and unlabeled data. Machine Learning flows through Pre-processing and Classification or Clustering. In gene expression microarray data, machine learning-based feature selection approaches like gene selection approaches will help to select the required genes from the lot.

Feature selection helps in preserving the informative attributes. Feature selection is primarily applied to the high-dimensional data; in simple terms, feature selection is a dimensionality reduction technique [3]. Feature selection assists significantly in the fields, which have too many features and relatively scarce samples, for instance, RNA sequencing and DNA Microarray [4].

The primary intent that feature selection got famous in the recent past is to extract the informative subset of features from the original feature space [4]. Feature selection techniques aids in overcoming the scare of model overfitting, handling the dimension, better interpretation of the feature space, maximizes prediction accuracy, and maximizes the model training time [5, 6]. The outcome of Feature selection is the optimal number of features that are relevant to the given class label, which contributes to the process of prediction.

One more technique for dimensionality reduction is Feature Extraction. Feature Selection is part of Feature Extraction [7]. It is the process of transforming the original feature space into a prominent space, which can be a linear or non-linear combination of the original feature space [8]. The major drawback of using Feature Extraction is that it alters the original feature space; eventually, the data interpretability is lost. Also, the transformation is usually expensive [9].

Gene expression is the flow of genetic information from Deoxy-ribose Nucleic Acid (DNA) to Ribose Nucleic Acid (RNA) to protein or other biomolecule syntheses. Gene expression data is a biological representation of various transcriptions and other chemicals found inside a cell at a given time. As data is recorded directly from DNA, through various experiments, a pertinent computational technique will reveal deep insights about the disease or disorder in the cell, eventually the organism in which the cell belongs [10].

On the one hand, the gene expression data is highly dimensional; also, on the other, the sample size is incompetent. The high dimensionality in the data is due to the vast number of values generated for every gene in a genome in the order of thousands. Advanced technologies, for instance, Microarray, assists in analyzing thousands of proteins in a gene in a particular sample. However, the issue with Microarray is that it is expensive [11].

However, the data with vast feature space will have redundant features with unnecessary information that will lead to overfitting, significantly affecting the model's performance. The primary purpose of implementing the Feature selection or gene selection on gene expression data is to choose the most regulating genes and eliminate the redundant genes that do not contribute to the target class [12].

The gene expression data are usually unlabeled, labeled, or semi-labeled, which leads to the necessity of the concepts of Unsupervised, Supervised, and Semi-supervised feature selection. Unlabeled data has no prior information about the functionalities, whereas it validates the gene selection based on data distribution, variance, and separability. Labeled data consists of meaningful class labels and information about the functionalities. Then gene selection will be performed based on the relevance and importance score of the labeled features. Semi-supervised or Semi-unsupervised combines a small amount of unlabeled data with labeled data and vice versa, which acts as additional information [13]. This paper discusses the importance of feature selection or gene selection to have an improved result. This paper's remaining sections discuss the background and development of feature selection, the steps involved in feature selection, a detailed discussion on various works on gene selection in the literature, the open issues, and future research directions concerning the gene expression data and conclusion.

The feature selection methods can be categorized into Supervised, Unsupervised, and Semi-supervised learning models. The survey works in the literature concentrate on either one of the models; for example, [14] focuses only on the supervised gene selection methods. Some works also concentrate on one particular feature selection strategy; for example, [42] focuses on filter-based techniques. Table 1 shows the comparison of existing reviews with the current survey. Our study categorizes the feature selection strategy into supervised, unsupervised, and semi-supervised methods and discusses the existing approaches in those categories. Also, we have done a detailed discussion of their performances.

2 Gene Selection – Background and Development

Gene Selection is the technique applied to the gene expression dataset, such as DNA Microarray, to reduce the number of genes, which are redundant and less expressive or less informative. Gene Selection has its base in the Machine Learning-based Feature Selection technique, which significantly suits the applications that involve thousands of features [15]. Gene Selection techniques are applied mainly for two reasons: finding the informative and expressive genes and removing the original space's redundant genes. Theoretically, an increase in the number of genes will bring down the model's performance and compromise the generalization by overfitting. The present works on Gene Selection concentrate mainly on finding the relevant genes, and there is limited research in removing the noise and redundant genes [16].

For significant results, it is critical to concentrate on relevancy, redundancy, and complementarity. A gene is considered as relevant when it has necessary information (individually or combined with other genes) about the given class, for example, tumorous or not. According to [17], the feature subset can be classified into strongly relevant, weakly relevant, and irrelevant in technical terms. The weakly irrelevant can again be classified into weakly relevant and redundant features and weakly relevant and non-redundant features. Most of the informative features can be found under strongly relevant and weakly relevant, and non-redundant features [18]. The same approach is followed in the Gene Selection from the gene expression data. Figure 1 shows the representation of the Gene Selection approach.

Many works in literature [19, 20, 21] aim to remove redundancy and relevancy from the data with the Mutual Information algorithm's help in Gene Expression. Many variations in Mutual Information are implemented to tackle these two issues. Along with these two issues, there is one more issue, which many of the existing works fail to address, complementarity. Complementarity is the degree of feature interaction between a gene subset and an individual gene in a given class.

To solve the issues mentioned above, commonly, two approaches are followed in the literature, one is analyzing individual genes, and the other is finding an optimal subset. In analyzing individual genes, the genes are ranked based on their importance scores; genes with a similar score (redundant) and genes with the least score (irrelevant) below a given threshold will be removed. In finding an optimal subset, a search for a minimal subset of genes will be done, satisfying specific criteria and eliminating redundant and irrelevant genes.

In applications such as Text and Genomic Microarray analysis, the central issue is the “Curse of Dimensionality,” where finding the optimal subset of genes is considered an NP-hard problem. Effective learning will be achieved only when the model is trained with relevant and non-redundant

genes. However, with an increase in the genes' dimension, the possible number of optimal gene subsets will also increase exponentially.

In machine learning, feature space is defined as the space associated with a feature vector distributed all over the sample in an n-dimensional space. Moreover, to reduce the dimensionality of such feature space, feature extraction, or feature selection techniques can be used. Feature Selection is a part of the Feature Extraction technique. However, in feature selection, a subset from the original feature space will be formed, whereas, in feature extraction, a new set of feature space will be created that seems to capture the necessary information from the original feature space [22]. The most commonly used feature extraction techniques are Principle Component Analysis (PCA), Independent Component Analysis (ICA), Expectation-Maximization (EM), and Linear Discriminant Analysis (LDA). Some examples of Feature Selection techniques are RELIEF, Conditional Mutual Information Maximization (CMIM), Correlation Coefficient, Information Gain, and Lasso [23].

The major drawback of using Feature extraction is that the data's interpretability will be lost in the transformation. Also, the transformation itself will be expensive sometimes [23]. Therefore, in this paper, we will discuss various Feature Selection techniques used in Gene Selection, which is less expensive and preserves the data's interpretability.

The Gene Selection based on machine learning can be classified into three types, Supervised, Unsupervised, and Semi-Supervised. Supervised Gene Selection utilizes the genes that are labeled already [24]. The input and output labels are known in advance in this method. However, the data continues to grow and overwhelm the process, leading to data mislabeling, making it unreliable. The main issue in deploying Supervised Gene Selection is overfitting, which can be caused by selecting irrelevant or sometimes eliminating the most relevant gene [4].

Unsupervised Gene Selection, unlike Supervised, will not have any labels to guide the selection process [24]. The data used in Unsupervised Gene Selection is unlabelled. That makes it unbiased and serves as an effective way to find the necessary insights into the classification process [26]. The main issue in Unsupervised Gene Selection is that it does not consider the interaction among the Genes (correlation), making the resultant gene subset insignificant in the discrimination task [27].

Semi-supervised or Semi-unsupervised Gene Selection is like an add-on to the Supervised and Unsupervised Gene Selection. A Gene Selection is considered semi-supervised when most of the data is labeled, and a Gene Selection is said to be Semi-unsupervised when most of the data is unlabelled [28]. The labeled data in the Semi-supervised or unsupervised is used to increase the distance between the data points that belongs to different classes, whereas the unlabelled data will help identify the geometrical structure of the feature space [29]. Figure 2 illustrates the overview of the process involved in Gene Selection.

2.1 Steps Involved in Feature Selection

2.1.1 Search Direction

The first stage involved in Feature Selection is to choose a search direction, which serves as a starting point to the process. There are three commonly used search directions,

- Forward Search: In Forward Search, the Search will be started with an empty set, and features are added one by one [30].

- Backward Search: Search will be started with the whole set of genes, and the genes will be eliminated one by one with each iteration.
- Bi-directional: Search involves the advantages of Forward Search and Backward Search. The Search starts from both directions by either adding or removing a gene with each iteration [31]. Other than these, Random Search is also used as a search direction [32].

2.1.2 Search Strategy

A good search strategy should attain fast convergence and provide an optimal solution with efficient computational cost and good global search ability [5]. There are three most widely used searching strategies,

- Sequential: follows a particular order in finding the best feature subset, for instance, Sequential Forward Search, where the search will be carried out from the start to the end [33]. This strategy is prone to feature interaction and has the risk of attaining local minima [32]. Examples: Floating Forward or Backward, Linear Forward Search, Beam Search, Greedy Forward Selection, and Backward Elimination
- Exponential: It is a full-scale search; it guarantees an optimal solution but proves to be expensive. This approach finds all possible feature subsets to choose an optimal subset, which is computationally upscale, especially in high-dimensional datasets such as the Gene Expression Microarray dataset. Some of the examples for Exponential Search are, Exhaustive Search and Branch-and-bound.
- Heuristic Search: It is performed based on a cost measure or a heuristic function, which iteratively improves the solution. Heuristic Search does not always ensure an optimal solution, but it offers an acceptable solution with reasonable time, cost, and memory space [34]. Some examples of Heuristic Search are Best-First Search, Depth-First Search, A* Search, Breadth-First Search, and Lowest-Cost-First Search [35].

2.1.3 Evaluation Criteria

There are currently four types of evaluation methods used widely; they are Filter, Wrapper, Embedded, and Hybrid. Hybrid and Embedded methods are the recent developments in Gene Selection.

a) Filter Feature Selection Approach

Filter helps in identifying the specific abilities of features depending on the inherent properties of the data. The best among the features are identified with relevance score and threshold criteria [36]. The features with a low relevance score will be eliminated.

The significant advantages of filter techniques are that they are not dependent on the classifiers, fast and straightforward in terms of computation, and scaled to the immensely dimensioned dataset [4]. The common disadvantage is that they consider the data's univariate features, which means the features are processed individually [37]. As a result, there are high chances of ignoring the feature dependencies, which leads to the classifiers' poor performance compared to other feature selection approaches. Many multivariate filter techniques are introduced to avoid this to some extent [38, 39, 40, 41].

The examples for filter techniques are Pearson Correlation, Fisher Score, Model-based Ranking, and Mutual Information [42] were done in a detailed survey on the filter techniques applied to Gene Expression Microarray data. Figure 3 is the representation of the process involved in the filter approach in gene selection.

b) Wrapper Feature Selection Approach

Unlike the filter approaches, the wrapper approaches wrap the feature subset selection process around the black box's induction algorithm. Once the search procedure for a feature subspace is defined, various feature subsets will be generated, and the classification algorithm is used to evaluate the selected feature subsets [43]. With this approach, it is possible to select features tailored for the induction algorithm [44]. The classification algorithm's evaluation measures will be optimized while eliminating the features, hence offering better accuracy than the filter approach [45, 46].

The significant advantage of using a wrapper approach, as both feature subset generation and the induction algorithm are wrapped together; the model will have the ability to track the feature dependencies [47]. The common drawback is that it becomes computationally intensive for datasets with high dimensions [46]. Examples of Wrapper techniques are Hill Climbing, Forward Selection, and Backward Elimination. Figure 4 is the representation of the process involved in the wrapper approach.

c) Embedded Feature Selection Approach

In a way, embedded approaches resemble the wrapper approaches, as both depend on the learning algorithm [48]. However, the embedded methods are less computationally intensive than the wrapper methods. The link between the learning algorithm and the feature selection is more robust in embedded methods than the wrapper methods [49]. In the embedded methods, the feature selection is made as a part of the classification algorithm; in other terms, the algorithm will have its built-in approaches to select the essential features [50].

In the literature, it is mentioned that embedded methods combine the benefits of filter and wrapper methods to improve accuracy. The significant difference between other gene selection approaches and embedded approaches is how the genes are selected and the interaction with the learning algorithm [51, 52]. Some examples of embedded approaches are ID3, RF, CART, LASSO, L1 Regression, and C4.5. Figure 5 is the representation of the process involved in the embedded approach.

d) Hybrid Feature Selection Approach

Hybrid methods, as the name suggests, is a combination of two different techniques. Here, it can be two different feature selection approaches or different methods with similar criterion or two different strategies. In most cases, the filter and wrapper approaches are combined to form a hybrid approach [53, 54]. It strives to utilize the benefits of two methods by combining their compatible strengths. Hybrid methods offer better accuracy and computational complexity than the filter and wrapper methods. Also, it is less susceptible to overfitting [55]. Figure 6 is the representation of the process involved in the hybrid approach.

2.1.4 Stopping Criteria

The stopping criteria are a kind of threshold used to inform the classifier when to stop selecting the features [16]. Appropriate stopping criteria will refrain a model from overfitting, thus offer better results, which are computationally cost-effective [4]. Some of the commonly used stopping criteria are as follows,

1. When the search reaches a specific bound, the bound can be several iterations or many features.
2. The results do not improve with a deletion (or addition) of another feature.
3. An optimal subset is found. A subset is said to optimal when the classifier's error rate is less than the preferred threshold.

2.1.5 Evaluating the Results

There are many performance evaluation metrics available in the literature to evaluate and validate the classifier results. In the classification case, i.e., predicting using the categorical attribute, the commonly used error estimation methods are Confusion Matrix, Cross-Validation, and Receiver Optimizer Characteristics (ROC). In the case of regression, i.e., predicting using the continuous attribute, the commonly used error estimation methods are Mean Absolute Error (MAE), Mean Squared Error (MSE), and Coefficient of Determination (R2).

a) ***Confusion Matrix:*** In the case of Multi-class problems, a confusion matrix is the best option to evaluate the classification model [56]. For instance, there are four possible results in a binary classification problem with which the model can be evaluated, True Positive, classified correctly, False Positive, erroneous classification, False Negative, erroneously rejected, and True Negative rejected correctly [57]. Confusion Matrix offers measures such as Accuracy, Precision, Sensitivity, Specificity, and FMeasure to validate the results of a classifier.

b) ***Cross-Validation (CV):*** It is the process of partitioning the available data into k-sets. Here, k can be any integer depending on the number of folds one needs for the classification or regression task (for instance, k = 10, k = 20, etc.) [57, 58]. CV is most commonly used on the Regression and Classification approaches [51]. The main advantage of using CV is that it offers unbiased error estimation, although sometimes it is variable [59].

c) ***Receiver Optimization Characteristics (ROC):*** ROC graphs and curves are commonly used for visualizing the performance of the classifiers and select the one showing better performance [60]. As the researches these days are increasingly concentrated on the classification errors and unbalanced class distribution, ROC has gained a lot of attention [61]. It is the depiction of the trade-offs between the Sensitivity or benefits (TPR) and the Specificity or costs (FPR) [62].

d) ***Root Mean Square Error (RMSE):*** RMSE is a metric commonly used to measure the residuals' standard deviation or prediction scores. In other words, the deviation in predictions from the regression line. It is given by [63],

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x_i})^2}{n}}$$

Where, x_i – Actual or Observed Values

$\bar{x_i}$– Predicted Values

n – Total number of sample

e) Mean Absolute Error: It is the standard measure of the residuals' average magnitude (prediction errors), neglecting their directions. It is given by [63],

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x_i}|$$

Where, x_i – Actual or Observed Values

$\bar{x_i}$– Predicted Values

n – Total number of sample

f) Determination Coefficient (R^2): It is the measure to estimate how much one variable impacts other variables. It is the change in the percentage of one variable concerning the other. It is given by [63],

$$R^2 = \left[\frac{n[\sum(xy) - (\sum x \sum y)]}{\sqrt{n\sum x^2 - (\sum x)^2[n\sum y^2 - (\sum y)^2]}}\right]^2$$

Where, x – first set of values data

y – the second set of values in the data

R – Coefficient of determination

n – Total number of sample

3 Machine Learning based Gene Selection Approaches

3.1 Supervised Gene Selection

Supervised Gene Selection involves the data with labeled attributes. Most of the studies done in recent years have concentrated mainly on enhancing and improving the existing supervised gene selection methods.

For instance, Devi Arockia Vanitha et al., 2014 enhanced the Mutual Information (MI) filter method for selecting the informative gene. Also, Joe's Normalised Mutual Information, an improved version of the standard existing MI approach, was implemented by Maldonado & López, 2018. Filter approaches are independent of the classifiers used. Hence, many works are focused on developing filter technologies. For instance, a novel filter approach is mainly based on the Hilbert-Schmidt

Independence Criterion (SHS) and motivate by Singular Value Decomposition (SVD). Table 2 shows some of the filter-based gene selection techniques used in the literature to select informative genes.

The wrapper approach is computationally intensive than other feature selection approaches. Works on the wrapper feature selection approach are less because of the issue mentioned above. So, most of the research on the wrapper is focused on improving the computational cost. For instance, Wang et al., 2017 implemented a wrapper-based gene selection with Markov Blanket, which reduces the computation time. Many approaches try to enhance the most widely used Support Vector Machine – Recursive Feature Elimination (SVM-RFE), such as Shukla et al., 2018, implemented Support Vector Machine – Bayesian t-test – Recursive Feature Elimination (SVM-BT-RFE), where Bayesian t-test is combined with SVM-RFE to improve the results. Table 3 shows the works done in recent years on Wrapper-based Supervised Gene Selection.

Hybrid Feature Selection is usually the combination of other approaches, mostly filter and wrapper approaches are made into hybrids. For instance, Liao et al., 2014, implemented a filter-wrapper based hybrid approach utilizing the Laplacian score and Sequential Forward and Backward Selection. Also, various works are going on in combining the nature-inspired algorithm. For example, Alshamlan et al., 2015, implemented a Genetic Bee Colony, combining the Genetic Algorithm and Artificial Bee Colony for gene selection. A hybrid of the Salp Swarm Algorithm (SSA) and multi-objective spotted hyena optimizer are implemented in A Sharma et al., 2019. The SSA focuses on diversity, and MOSHO concentrates on convergence. Table 4 consists of the recent works done on Hybrid-based Supervised Gene Selection approaches.

Ensemble Feature Selection is a combination of the outputs from different expert feature selection approaches. Ghosh et al., 2019, combines the outputs of ReliefF, Chi-square, and Symmetrical Uncertainty (SU) with Union and Intersection of top ‘n’ features. Seijo-Pardo et al., 2016, used a ranking aggregation method to various aggregate ranks from Chi-square, InfoGain, mRmR, and ReliefF. Table 5shows the different Ensemble-based Supervised Gene Selection approaches used in recent years.

Embedded methods merge the benefits of filter and wrapper methods, where the learning algorithm has a built-in feature selection approach. Ghosh, Begum, et al., 2019, implemented a Recursive Memetic Algorithm (RMA) with a wrapper-based approach embedded in it. Also, Guo et al., 2017, used L1 Regularization, along with a feature extraction method for selecting the informative genes. Table 6 shows the various Embedded-based Supervised Gene Selection approaches developed in recent years.

3.2 Unsupervised Gene Selection

Unsupervised Gene Selection involves data without any labels. Compared to Supervised Gene Selection, works on Unsupervised are less.

There are many novel works done on filter-based unsupervised gene selection, such as Solorio-Fernández et al., 2017, proposed a filter method for both non-numerical and numerical data. It is a combination of kernel approach and spectrum-based feature evaluation. Also, Liu et al., 2018, developed a Deep Sparse Filtering model considering the deep structures, enhancing the results. Many studies on nature-inspired gene selection and the [88] implemented the MGSACO to minimize redundancy, thereby increasing the dataset's relevancy. One another issue with high-dimensional data is dependency maximization. The work in [85] implemented the Hilbert-Schmidt Independence

Criterion to eliminate the most dependent genes to handle dependency maximization. Table 7 is the collection of works done in recent years on Filter-based Unsupervised Gene Selection approaches.

Filter-based gene selection approaches are not dependent on the learning model; on the contrary, wrapper methods are entirely dependent on the learning model. The dependency makes it complicated and has a high computational cost. Hence, the study on wrapper methods is less concentrated. Same with the unsupervised wrapper gene selection, which is less focused. Xu et al., 2017, has implemented SVM-RFE, a wrapper-based gene selection, on unlabeled data to distinguish high-risk and low-risk cancer patients. Table 8 is an example of a wrapper-based Unsupervised Gene Selection approach.

Hybrid Unsupervised gene selection is also focused on in the literature as much as the filter approach. Li & Wang, 2017, developed a two-stage gene selection approach; it applies the matrix factorization and minimum loss principle. A coarse-fine hybrid gene selection on unlabelled data shows better results than a few other approaches compared to the study. Filter-wrapper hybrid approaches are equally focused on supervised as well as unsupervised gene selection. For instance, Solorio-Fernández et al., 2016, implemented a Laplacian Score Ranking, a filter approach, and Normalised Calinski-Harabasz (LS-WNCH), a wrapper approach as hybrid unsupervised gene selection. It includes the properties of spectral feature selection. Table 9shows the hybrid-based Unsupervised Gene Selection approaches.

Ensemble and embedded approaches are studied less than the filter and hybrid methods. Elghazel & Aussem, 2013, implemented a Random Cluster Ensemble with k-means as the clustering model. The ECE was constructed with different bootstrap samples at every ensemble partitions. They have also calculated out-of-bag feature importance at every ensemble. Jundong Li et al., 2017, developed a Reconstruction-based unsupervised feature selection model, an embedded approach. The model has a filter-based approach embedded in the k-means clustering. Table 10is the example for Ensemble-based, and Embedded-based Unsupervised Gene Selection approaches.

3.3 Semi-Supervised Gene Selection

Semi-supervised gene selection is yet to be explored research area. There are not many works done as much as supervised or unsupervised gene selection. Semi-Supervised or Semi-Unsupervised consists of both labeled and unlabelled data.

Z. Li et al., 2018, combined the benefits of the spectral graph and Mutual Information to develop a Semi-Supervised Maximum Discriminative Local Margin (SemiMM). It takes care of variance, local structure, and MI all at the same time. SVM is used widely in supervised and unsupervised gene selection approaches; in semi-supervised, Ang et al., 2016, implemented a semi-supervised SVM-RFE (S3VM) for selecting the informative genes, and it proves to be successful. Chakraborty & Maulik, 2014, developed a hybrid model; Kernalised Fuzzy Rough Set (KFRS) and S3VM are combined to select the relevant features. The results show that the proposed algorithm is capable of choosing useful biomarkers from the dataset. A semi-supervised embedded approach, Joint Semi-Supervised Feature Selection (JSFS), was developed with a Bayesian approach. The model automatically chooses the informative features and also trains the classifier.

Rajeswari & Gunasekaran, 2015, developed an ensemble-based semi-supervised gene selection to improve the quality of the cluster model. Modified Double Selection based Semi-Supervised Cluster

Ensemble (MDSVM-SSCE) assists in selecting the most relevant genes. Table 11 shows the Semi-Supervised Gene Selection approaches developed in recent years.

4 Performance Analysis and Discussion on the Reviewed Literature

In the literature, the top three datasets used widely are Prostate, Leukaemia, and Colon. Table 12, 13, 14 shows the respective proposed models' performance on the datasets mentioned above, along with the number of genes selected.

All three gene selection methods discussed in this paper has its own merits and demerits. From the literature, it is clear that the Supervised Gene Selection is researched the most in recent years, and the Semi-supervised the least. Even though the Semi-Supervised potential is not tapped upon yet, it seems to be the better one among the three. It takes the advantages of Supervised and Unsupervised Gene Selection approaches. It has both labeled and unlabelled data; thus, it combines both the approaches' benefits, eventually achieving better results. It considers the overlapping genes and handles it with the Unsupervised Gene Selection approach (unlabelled data) and learn and train the learning model with great accuracy and precision with the help of Supervised Gene Selection approaches (labeled data). Figures 7a, 7b, 8, and 9 show that the Supervised Gene Selection performs way better than the other two. Still, it might be because there are considerably significantly fewer works in Unsupervised and Semi-Supervised Gene Selection. The abbreviations for the acronyms used in the plot can be found in table 15. There are several opportunities still untapped in these two areas. We can also notice that many works are concentrated more on Filter approaches as they are simple and computationally effective. However, hybrid approaches are upcoming and promising.

As for the evaluation criteria, in recent years, filter-based approaches are more focused much. Filter methods function independently of the learning model; thus, it is less computationally intensive. As it is less complicated, many researchers target the filter-based approaches in selecting informative genes. Wrapper-based approaches are the least concentrated upon; it is dependent and designed to support the learning model. Wrapper approaches are usually time-consuming and generate high computational overhead. Though other methods are concentrated equally, the hybrid approach proves to be better among the others. Hybrid is a combination of two or more approaches. The most commonly used hybrid method is the Filter-Wrapper combination. In the Hybrid approach, the limitations of the individual approaches are compensated; in other words, it inherits the benefits of two methods. Further, this will minimize computational cost. Hybrid approaches seem to provide better accuracy and reduce over-fitting risks. Apparently, hybrid methods are most suited for high-dimensional datasets such as the gene expression microarray from the literature.

Apart from the discussed literature, many other works focused on nature-inspired and meta-heuristic algorithms in diagnosing cancer. A bio-inspired algorithm is proposed by [107] using the BAT algorithm with more refined and effective multi-objectives. Also, they have proposed a novel local search strategy. Another such BAT inspired algorithm with two-staged gene selection is proposed in [108], wherein the first stage is a filter (Minimum Redundancy and Maximum Relevance) and the second stage is the wrapper consisting of BAT and SVM. Other than that, considerable works are done in Particle Swarm Optimization (PSO) by improving and enhancing the existing algorithm. In [109], the authors implemented a two-phased hybrid gene selection method, combining the improved PSO (iPSO) and Correlation-based Feature Selection (CFS). The proposed method controls the early convergence problem. A recursive PSO is implemented in [110]; it tries to refine the feature space into more fine-grained. They have also combined existing filter-based feature selection methods with

the recursive PSO. KNN and PSO are implemented in [111] to handle the uncertainty involved in choosing the k-value in KNN. In [112], the authors proposed a Binary PSO (BPSO) to improve the interpretability of the gene selected and improve the prediction accuracy of the model. In [113], a nature-inspired algorithm Harmony Search Algorithm (HAS) is embedded with Markov Blanket, which focuses on symmetrical uncertainty.[114] implemented an Ant Colony Optimization based gene selection (ACO) along with Cellular Learning Automata (CLA) as a wrapper method. In another approach [115], a hybrid combining filter and wrapper approaches is implemented using Information Gain (IG) and improved Swarm Optimization to find the optimal gene subset. Information Gain (IG) is also implemented along with SVM in [116] to remove the redundant genes. There are works done in gene selection using the Genetic algorithms with different variations from the existing one. One such work combines the Genetic algorithm and Fuzzy in [117], integrating the two approaches to finding out the optimal gene subset. Genetic Algorithm is also combined with learning automata (GALA) in [118], which improves the time complexity in selecting the gene subset. Statistically, significant models are also implemented, such as the entropy-based measure and rough sets [119] and [120, 121], testing the statistical significance with p-value and fold change. Decision tree and random forest variances are also worked on, such as the four-state-of art Random forest [122], decision tree along with PSO [122], and a guided regularised Random Forest [124]. Various works are focus on improving the interpretability of the features and reducing the feature space with improvements in the existing models [125, 126, 127, 128, 129, 130, 131].

Machine Learning techniques are widely used in modern-day research in the field of bioinformatics. The Machine Learning algorithms are available under different criteria, such as the logic-based algorithms (E.g., Decision Trees, Random Forest), perceptron-based algorithms (Neural Network, Multi-layered Perceptron), and Statistical Learning (Naïve Bayes) [132]. The classification or prediction models used commonly in the literature discussed in this paper mostly include SVM, KNN, Random Forest, Decision Tree, Naïve Bayes, and Logistic Regression. SVM consists of support vectors that assist in classifying a disease or disorder. The classification depends on the formation of a hyperplane that divides binary classes. The SVM locates the hyperplane with the help of the kernel function. A most important advantage of using SVM is to tackle the outliers [133]. KNN works on the assumption that the instances within a dataset will be close to one another. Although KNN is easy to understand and implement the algorithm, it lacks the fundamental principle in choosing the value of k. Also, it is sensitive to the distance or similarity function used. Decision Tree is made up of nodes and branches, used mainly because of their effectiveness and speed in calculations. Decision Trees are highly prone to overfitting and underfitting of the data [134]. Random Forests are the ensemble of Decision Tree. Naïve Bayes is the statistical classification model. Based on the Bayes Theorem, it works on the assumption that all the features in the dataset are independent and equal.

In general, for continuous and multi-dimensional features, neural networks and SVM show better performance. Whereas, in the case of the categorical or discrete features, the logic-based algorithms, such as the rule learners and decision trees, perform better. SVM and others will need a large sample size to produce high accuracy, but Naïve Bayes works on a small dataset. The training time varies for each algorithm; for example, Naïve Bayes trains quickly because of their single pass of the entries. Also, it does not need much storage space during training and testing. On the contrary, during training, KNN based models require huge storage space and more than that during the testing phase.

In terms of interpretability, the logic-based models are interpreted easily, whereas SVM and neural networks are difficult to interpret. They also have the highest number of parameters, which need

optimization and tuning. One algorithm cannot outperform the other. One way to determine the type of algorithm to use is to validate the models and estimate their accuracy and choose the one with better accuracy. Recently, combining the algorithms are proposed to enhance individual algorithm performances. However, the gene expression data has the issue of High Dimension and Low Sample Size (HDLSS), for which machine learning models are less suited. Hence, the Deep Learning and Deep Belief Networks are being researched in recent days and a multi-omics dataset.

In the performance evaluation metrics, the commonly used ones are the Classification Accuracy, Least One Out Cross Validation (LOOCV), k-Fold Cross-Validation, and ROC. Among these, several works use the Classification Accuracy. However, many performance metrics need concentration, such as sensitivity, sensibility, and similarity measures.

5 Open Issues in Gene Expression Data

The gene expression is a biological process; DNA instructions are transformed into a functional product called the proteins. The cells in a living organism do not need proteins all the time. Certain complex molecular mechanisms must turn the genes on and off. If that does not happen, diseases and disorders will follow.

DNA Microarray is a technology used widely in biomedical research to analyze gene expression to discover the disease or disorder, classify, and predict. The DNA microarray data is also used to predict the responses of a drug or therapies given. There are different types of DNA microarray, such as cDNA (complementary Deoxy-Ribose Nucleic Acid), SNP (Single Nucleotide Polymorphism, and CNV (Copy Number Validation) microarrays [135]. cDNA is a DNA without introns and formed from a single-stranded RNA. SNP is the variations that can be found only at a single point in a DNA sequence. CNV is a condition where parts of a genome will be repeated, and the repetition will vary from one individual to another. There are many advanced technologies available to analyze gene expression. Most widely used are cDNA bi-color glass slide and Affymetrix GeneChip.

Many challenges and limitations need to be addressed to extract the required knowledge from the gene expression with great precision. The significant difficulties are as follows [102, 136, 137]:

a) ***Curse of Dimensionality:*** The major issue that is researched upon in machine learning is the overfitting of a learning model. The work in [138] discusses the curse of dimensionality in detail. Microarray is generally high-dimensional data, ranging from hundreds to thousands and more features. Microarray data prove to be hectic in managing. To handle such huge volumes of data, advanced storage systems are required [139, 140].
b) ***The gap between the Researchers and Biologists:*** There is a huge gap among the researchers, biologists and medical practitioners, which led to many unexplored areas in the genomic studies. The opportunity of finding the best techniques and approaches are very less because of the aforementioned gap.
c) ***Redundant and Mislabelled Data:*** Data imbalance and mislabelled data is the most prevailing issue in the Microarray data because of the irregular scanning. The Microarray dataset usually has class imbalance issue, i.e. one class will dominate the entire dataset. When the learning model is trained on a mislabelled and imbalanced data, it will greatly affect the generalization ability of the learning model. Same as the abovementioned issues, redundant and irrelevant data are also the main concern in determining the efficiency of the feature set [141, 142].

d) ***Difficulty in Retrieving the Biological Information:*** There are many clinical challenges in retrieving the biological information. The main aim of genomic studies is to discover the significant changes in the gene expression, clinically or biologically. The difficulty is that not everyone will possess high-ended equipment to capture significant changes. Also, in some of the biological processes, the changes in the expression are very subtle and difficult to be identified with analytical methods. Due to the different range of approaches regarding the experimental design, data access, study and batch of reagents used, the data may be erroneous and biased.

Some of the future directions with which the research in this area can be proceeded are as follows,

a) Enhanced Models for Better Diagnosis of Rare Genetic Disorders

There are various genetic disorders classified under Monogenic and Polygenic disorders. Monogenic disorders are caused because of modifications in a single gene and inherited genetically. It is rare. Unlike Monogenic, Polygenic are commonly occurring and caused because of modifications in several genes. The genetic illnesses of such types are overwhelming in the recent years. Machine Learning classification and prediction models will diagnose the disorders with great accuracy.

b) Cancer Prognosis and Prediction

Cancer is a heterogeneous disease, which is considered to have various subtypes. It is critical to diagnose early to further assist the patients clinically. The importance of grouping high and low risk patients had led to various researches in bioinformatics and machine learning applications. The ability of machine learning models such as Support Vector Machine (SVM), Artificial Neural Networks (ANN) and Bayesian Networks (BN) in the development of classification and predictive models for accurate decisions have to be explored.

c) Collaborative Platforms in Gene Expressions

The individual models in Machine Learning will yield better results when applied on gene expression data. However, hybrid methods prove to be successful at many instances. Along with hybrid methods, more research should be done in combining different gene expression data and clinical reports. It is difficult and exhaustive, yet it will offer greater results.

d) Analysing Drug Response in Gene Expression Data

Predicting a drug response to any genetic disorder or disease is an important step. Many recent efforts in analysing the sensitivity and response to cancer or other diseases are commendable. Still, the main problem in developing a model for drug response is the high dimension and less sample size. The feature selection techniques in Machine Learning assist in reducing the dimensions and improve the accuracy in predicting the drug response.

6 Conclusion

Gene expression Microarray is a high-dimensional database with less sample size. It needs powerful techniques to handle it and preserve the informative genes by minimizing the redundancy and dependency. This paper discusses the works done in the recent years in the gene expression microarray dataset. The papers are selected from the past six years, the focus is mainly on the

supervised, unsupervised and semi-supervised based feature selection in the gene expression data. Further, under those three learning methods, we have chosen papers that concentrate on filter, wrapper, hybrid, embedded and ensemble based gene selection. This study lists out the significant difficulties faced in handling such huge dimensional datasets. To overcome the dimension issues, the gene selection must be made carefully. Although there are a lot of works done in the literature on the gene expression microarray data, there are many open opportunities that need attention. The researches have mainly focused on supervised gene selection with a filter as evaluation methods. The potentials of unsupervised and semi-supervised techniques are yet to be tapped. The semi-supervised technique works with the benefits of supervised and unsupervised techniques combined. Hence, the chances of improved accuracy is high in semi-supervised. The only aim of almost all the works is to achieve higher accuracy the focus on sensitivity, specificity, stability and similarity is scarce. As equally important as the dimensionality issue is the misclassification or mislabelled data. There is a promising future for overcoming these two issues. Another important direction for improvement in gene selection is to develop more ensemble and hybrid evaluation methods. As discussed in the literature, works on hybrid and ensemble are considerably less when compared to filter and wrapper approaches. Hybrid and ensemble methods are capable of providing more accurate results. Apparently, it needs further developments. Research must be done in joint analysis, to combine the clinical reports and the gene expression data. It will help in analysing various aspects and will offer a different perspective. It would serve as a major breakthrough, yet hectic and exhaustive.

7 References

1. Raut, S. A., Sathe, S. R., & Raut, A. (2010, April). Bioinformatics: Trends in gene expression analysis. In 2010 International Conference on Bioinformatics and Biomedical Technology (pp. 97-100). IEEE.
2. Wang, H., & van der Laan, M. J. (2011). Dimension reduction with gene expression data using targeted variable importance measurement. BMC bioinformatics, 12(1), 312.
3. Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In Machine Learning Proceedings 1992 (pp. 249-256). Morgan Kaufmann.
4. Ang, J. C., Mirzal, A., Haron, H., &Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM transactions on computational biology and bioinformatics, 13(5), 971-989.
5. Halperin, E., Kimmel, G., & Shamir, R. (2005). Tag SNP selection in genotype data for maximizing SNP prediction accuracy. Bioinformatics, 21(suppl_1), i195-i203.
6. Sun, L., Zhang, X., Qian, Y., Xu, J., & Zhang, S. (2019). Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification. Information Sciences, 502, 18-41.
7. Cárdenas-Ovando, R. A., Fernández-Figueroa, E. A., Rueda-Zárate, H. A., Noguez, J., & Rangel-Escareño, C. (2019). A feature selection strategy for gene expression time series experiments with hidden Markov models. PloS one, 14(10).
8. [8] Anter, A. M., & Ali, M. (2020). Feature selection strategy based on hybrid crow search optimization algorithm integrated with chaos theory and fuzzy c-means algorithm for medical diagnosis problems. Soft Computing, 24(3), 1565-1584.

9. Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., ...& Haley, C. S. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. Scientific reports, 5, 10312.
10. Koul, N., &Manvi, S. S. (2020). Machine-Learning Algorithms for Feature Selection from Gene Expression Data. In Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications (pp. 151-161). Springer, Singapore.
11. Wahid, A., Khan, D. M., Iqbal, N., Khan, S. A., Ali, A., Khan, M., & Khan, Z. (2020). Feature selection and classification for gene expression data using novel correlation based overlapping score method via Chou's 5-steps rule. Chemometrics and Intelligent Laboratory Systems, 199, 103958.
12. Pearson, W., Tran, C. T., Zhang, M., &Xue, B. (2019, June). Multi-Round Random Subspace Feature Selection for Incomplete Gene Expression Data. In 2019 IEEE Congress on Evolutionary Computation (CEC) (pp. 2544-2551). IEEE.
13. Yang, Y., Yin, P., Luo, Z., Gu, W., Chen, R., & Wu, Q. (2019). Informative Feature Clustering and Selection for Gene Expression Data. IEEE Access, 7, 169174-169184.
14. Kumar, C. A., Sooraj, M. P., & Ramakrishnan, S. (2017). A comparative performance evaluation of supervised feature selection algorithms on microarray datasets. *Procedia computer science*, *115*, 209-217.
15. Dashtban, M., &Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics, 109(2), 91-107.
16. Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F., &Mewes, H. W. (2005). Gene selection from microarray data for cancer classification—a machine learning approach. Computational biology and chemistry, 29(1), 37-46.
17. Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. Journal of machine learning research, 5(Oct), 1205-1224.
18. Vergara, J. R., &Estévez, P. A. (2014). A review of feature selection methods based on mutual information. Neural computing and applications, 24(1), 175-186.
19. Sun, L., & Xu, J. (2014). Feature selection using mutual information based uncertainty measures for tumor classification. Bio-medical materials and engineering, 24(1), 763-770.
20. Hu, Q., Pan, W., An, S., Ma, P., & Wei, J. (2010). An efficient gene selection technique for cancer recognition based on neighborhood mutual information. International Journal of Machine Learning and Cybernetics, 1(1-4), 63-74.
21. Hoque, N., Bhattacharyya, D. K., &Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. Expert Systems with Applications, 41(14), 6371-6385.
22. Jović, A., Brkić, K., &Bogunović, N. (2015, May). A review of feature selection methods with applications. In 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1200-1205). Ieee.
23. Khalid, S., Khalil, T., &Nasreen, S. (2014, August). A survey of feature selection and feature extraction techniques in machine learning. In 2014 Science and Information Conference (pp. 372-378). IEEE.

24. Filippone, M., Masulli, F., &Rovetta, S. (2006, July). Supervised classification and gene selection using simulated annealing. In The 2006 IEEE International Joint Conference on Neural Network Proceedings (pp. 3566-3571). IEEE.
25. Filippone, M., Masulli, F., &Rovetta, S. (2005, September). Unsupervised gene selection and clustering using simulated annealing. In International Workshop on Fuzzy Logic and Applications (pp. 229-235). Springer, Berlin, Heidelberg.
26. Ye, X., & Sakurai, T. (2017, May). Unsupervised Feature Learning for Gene Selection in Microarray Data Analysis. In Proceedings of the 1st International Conference on Medical and Health Informatics 2017 (pp. 101-106).
27. Acharya, S., Saha, S., & Nikhil, N. (2017). Unsupervised gene selection using biological knowledge: application in sample clustering. BMC bioinformatics, 18(1), 513.
28. Ang, J. C., Mirzal, A., Haron, H., &Hamed, H. N. A. (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM transactions on computational biology and bioinformatics, 13(5), 971-989.
29. Sheikhpour, R., Sarram, M. A., Gharaghani, S., &Chahooki, M. A. Z. (2017). A survey on semi-supervised feature selection methods. Pattern Recognition, 64, 141-158.
30. Mohapatra, P., Chakravarty, S., & Dash, P. K. (2016). Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. Swarm and Evolutionary Computation, 28, 144-160.
31. Abinash, M. J., &Vasudevan, V. (2018). A Study on Wrapper-Based Feature Selection Algorithm for Leukemia Dataset. In Intelligent Engineering Informatics (pp. 311-321). Springer, Singapore.
32. Wang, L., Wang, Y., & Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods, 111, 21-31.
33. Chen, Y., & Yao, S. (2017). Sequential search with refinement: Model and application with click-stream data. *Management Science*, *63*(12), 4345-4365.
34. Ruiz, R., Riquelme, J. C., & Aguilar-Ruiz, J. S. (2005, June). Heuristic search over a ranking for feature selection. In *International Work-Conference on Artificial Neural Networks* (pp. 742-749). Springer, Berlin, Heidelberg.
35. Russell, S. J., &Norvig, P. (2016). Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited
36. Hancer, E., Xue, B., & Zhang, M. (2018). Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems*, *140*, 103-119.
37. Saeys, Y., Inza, I., &Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. bioinformatics, 23(19), 2507-2517.
38. Djellali, H., Guessoum, S., Ghoualmi-Zine, N., &Layachi, S. (2017, October). Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection. In 2017 5th International Conference on Electrical Engineering-Boumerdes (ICEE-B) (pp. 1-6). IEEE.

39. Zhou, Y., Wang, P., Wang, X., Zhu, J., & Song, P. X. K. (2017). Sparse multivariate factor analysis regression models and its applications to integrative genomics analysis. Genetic epidemiology, 41(1), 70-80.
40. Brumpton, B. M., & Ferreira, M. A. (2016). Multivariate eQTL mapping uncovers functional variation on the X-chromosome associated with complex disease traits. Human genetics, 135(7), 827-839.
41. Rouhi, A., &Nezamabadi-pour, H. (2018, March). Filter-based feature selection for microarray data using improved binary gravitational search algorithm. In *2018 3rd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)* (pp. 1-6). IEEE.
42. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., ...&Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4), 1106-1119.
43. Blanco, R., Larrañaga, P., Inza, I., & Sierra, B. (2004). Gene selection for cancer classification using wrapper approaches. International Journal of Pattern Recognition and Artificial Intelligence, 18(08), 1373-1390.
44. Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing*, *69*, 541-553.
45. Inza, I., Larrañaga, P., Blanco, R., &Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial intelligence in medicine, 31(2), 91-103.
46. Mohamed, E., El Houby, E. M., Wassif, K. T., & Salah, A. I. (2016). Survey on different methods for classifying gene expression using microarray approach. International Journal of Computer Applications, 975, 8887.
47. Rodrigues, D., Pereira, L. A., Nakamura, R. Y., Costa, K. A., Yang, X. S., Souza, A. N., & Papa, J. P. (2014). A wrapper approach for feature selection based on bat algorithm and optimum-path forest. Expert Systems with Applications, 41(5), 2250-2258.
48. Hernandez, J. C. H., Duval, B., &Hao, J. K. (2007, April). A genetic embedded approach for gene selection and classification of microarray data. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (pp. 90-101). Springer, Berlin, Heidelberg.
49. Huerta, E. B., Hernández, J. C. H., Caporal, R. M., Cruz, J. F. R., &Montiel, L. A. H. (2010). An efficient embedded gene selection method for microarray gene expression data. Research in Computing Sience.
50. Hira, Z. M., &Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in bioinformatics, 2015.
51. Chandrashekar, G., &Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.
52. Vanjimalar, S., Ramyachitra, D., &Manikandan, P. (2018, December). A Review on Feature Selection Techniques for Gene Expression Data. In 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1-4). IEEE.

53. Liu, H., Zhou, M., & Liu, Q. (2019). An embedded feature selection method for imbalanced data classification. IEEE/CAA Journal of AutomaticaSinica, 6(3), 703-715.
54. Apolloni, J., Leguizamón, G., & Alba, E. (2016). Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. *Applied Soft Computing*, *38*, 922-932.
55. Almugren, N., &Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. IEEE Access, 7, 78533-78548.
56. Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., ... &Asadi, H. (2019). Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. American Journal of Roentgenology, 212(1), 38-43.
57. Braga-Neto, U., Hashimoto, R., Dougherty, E. R., Nguyen, D. V., & Carroll, R. J. (2004). Is cross-validation better than resubstitution for ranking genes?. Bioinformatics, 20(2), 253-258.
58. Schaffer, C. (1993). Selecting a classification method by cross-validation. Machine Learning, 13(1), 135-143.
59. Bergmeir, C., &Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192-213.
60. Landgrebe, T. C., &Duin, R. P. (2008). Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. IEEE transactions on pattern analysis and machine intelligence, 30(5), 810-822.
61. Flach, P. A. (2016). ROC analysis. In Encyclopedia of Machine Learning and Data Mining (pp. 1-8). Springer.
62. Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.
63. Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. (2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. Computers and electronics in agriculture, 155, 257-282.
64. Ca, D. A. V., & Mc, V. (2015). Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Computer Science, 47, 13-21.
65. Shukla, Alok Kumar, and DiwakarTripathi. "Identification of potential biomarkers on microarray data using distributed gene selection approach." Mathematical biosciences 315 (2019): 108230.
66. Gangeh, M. J., Zarkoob, H., &Ghodsi, A. (2017). Fast and scalable feature selection for gene expression data using hilbert-schmidt independence criterion. IEEE/ACM transactions on computational biology and bioinformatics, 14(1), 167-181.
67. Mazumder, D. H., &Veilumuthu, R. (2019). An enhanced feature selection filter for classification of microarray cancer data. ETRI Journal, 41(3), 358-370.
68. Wang, A., An, N., Yang, J., Chen, G., Li, L., &Alterovitz, G. (2017). Wrapper-based gene selection with Markov blanket. Computers in biology and medicine, 81, 11-23.
69. Hasri, N. N. M., Wen, N. H., Howe, C. W., Mohamad, M. S., Deris, S., &Kasim, S. (2017). Improved support vector machine using multiple SVM-RFE for cancer

classification. International Journal on Advanced Science, Engineering and Information. Technology, 7(4-2), 1589-1594.

70. Shanab, A. A., Khoshgoftaar, T. M., & Wald, R. (2014, November). Evaluation of wrapper-based feature selection using hard, moderate, and easy bioinformatics data. In 2014 IEEE International Conference on Bioinformatics and Bioengineering (pp. 149-155). IEEE.
71. Mishra, S., & Mishra, D. (2015). SVM-BT-RFE: An improved gene selection framework using Bayesian T-test embedded in support vector machine (recursive feature elimination) algorithm. Karbala International Journal of Modern Science, 1(2), 86-96.
72. Zhang, Y., Deng, Q., Liang, W., & Zou, X. (2018). An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. BioMed research international, 2018.
73. Zare, M., Eftekhari, M., &Aghamollaei, G. (2019). Supervised feature selection via matrix factorization based on singular value decomposition. Chemometrics and Intelligent Laboratory Systems, 185, 105-113.
74. Chinnaswamy, A., & Srinivasan, R. (2016). Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data. In Innovations in bio-inspired computing and applications (pp. 229-239). Springer, Cham.
75. Alshamlan, H. M., Badr, G. H., &Alohali, Y. A. (2015). Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational biology and chemistry, 56, 49-60.
76. Liao, B., Jiang, Y., Liang, W., Zhu, W., Cai, L., & Cao, Z. (2014). Gene selection using locality sensitive Laplacian score. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(6), 1146-1156.
77. Shukla, A. K., Singh, P., &Vardhan, M. (2018). A hybrid gene selection method for microarray recognition. Biocybernetics and Biomedical Engineering, 38(4), 975-991.
78. L Sun, X Kong, J Xu, R Zhai, S Zhang (2019) A hybrid gene selection method based on ReliefF and ant colony optimization algorithm for tumor classification. Scientific Reports, 9, Article 8978
79. A Sharma, R Rani (2019) C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods. Computer Methods and Programs in Biomedicine, 178, 219-235
80. Ghosh, M., Adhikary, S., Ghosh, K. K., Sardar, A., Begum, S., & Sarkar, R. (2019). Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods. Medical & biological engineering & computing, 57(1), 159-176.
81. Seijo-Pardo, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2016, April). Using a feature selection ensemble on DNA microarray datasets. In ESANN.
82. Xu, J., Sun, L., Gao, Y., & Xu, T. (2014). An ensemble feature selection technique for cancer recognition. Bio-medical materials and engineering, 24(1), 1001-1008.
83. Yang, J., Zhou, J., Zhu, Z., Ma, X., & Ji, Z. (2016). Iterative ensemble feature selection for multiclass classification of imbalanced microarray data. Journal of Biological Research-Thessaloniki, 23(1), 13.

84. Brahim, A. B., &Limam, M. (2018). Ensemble feature selection for high dimensional data: a new method and a comparative study. Advances in Data Analysis and Classification, 12(4), 937-952.
85. Boucheham, A., Batouche, M., &Meshoul, S. (2015, April). An ensemble of cooperative parallel metaheuristics for gene selection in cancer classification. In International Conference on Bioinformatics and Biomedical Engineering (pp. 301-312). Springer, Cham.
86. Dashtban, M., &Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics, 109(2), 91-107.
87. Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., &Maulik, U. (2019). Recursive memetic algorithm for gene selection in microarray data. Expert Systems with Applications, 116, 172-185.
88. Guo, S., Guo, D., Chen, L., & Jiang, Q. (2017). A L1-regularized feature selection method for local dimension reduction on microarray data. Computational biology and chemistry, 67, 92-101.
89. Wang, H., Jing, X., &Niu, B. (2017). A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. Knowledge-Based Systems, 126, 8-19.
90. Maldonado, S., &López, J. (2018). Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification. Applied Soft Computing, 67, 94-105.
91. ZY Algamal, MH Lee (2015) Penalized logistic regression with the adaptive LASSO for gene selection in high-dimensional cancer classification. Expert Systems with Applications, 42(23), 9326-9332
92. Solorio-Fernández, S., Martínez-Trinidad, J. F., & Carrasco-Ochoa, J. A. (2017). A new unsupervised spectral feature selection method for mixed data: a filter approach. *Pattern Recognition*, *72*, 314-326.
93. Liaghat, S., &Mansoori, E. G. (2016). Unsupervised selection of informative genes in microarray gene expression data. International Journal of Applied Pattern Recognition, 3(4), 351-367.
94. Tabakhi, S., Najafi, A., Ranjbar, R., &Moradi, P. (2015). Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing, 168, 1024-1036.
95. Liu, J., Cheng, Y., Wang, X., Zhang, L., & Wang, Z. J. (2018). Cancer characteristic gene selection via sample learning based on deep sparse filtering. Scientific reports, 8(1), 1-13.
96. Xu, G., Zhang, M., Zhu, H., & Xu, J. (2017). A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. Gene, 604, 33-40.
97. Solorio-Fernández, S., Carrasco-Ochoa, J. A., &Martínez-Trinidad, J. F. (2016). A new hybrid filter–wrapper feature selection method for clustering based on ranking. Neurocomputing, 214, 866-880.
98. Manbari, Z., AkhlaghianTab, F., &Salavati, C. (2019). Hybrid fast unsupervised feature selection for high-dimensional data. Expert Systems with Applications, 124, 97-118.

99. Li, J., Tang, J., & Liu, H. (2017, August). Reconstruction-based Unsupervised Feature Selection: An Embedded Approach. In IJCAI (pp. 2159-2165).
100. Elghazel, H., &Aussem, A. (2015). Unsupervised feature selection with ensemble learning. Machine Learning, 98(1-2), 157-180.
101. Ang, J. C., Haron, H., &Hamed, H. N. A. (2015, June). Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 468-477). Springer, Cham.
102. Li, Z., Liao, B., Cai, L., Chen, M., & Liu, W. (2018). Semi-supervised maximum discriminative local margin for gene selection. Scientific reports, 8(1), 1-11.
103. Rajeswari, R., &Gunasekaran, G. Semi-Supervised Tumor Data Clustering via Spectral Biased Normalized Cuts.
104. Chakraborty, D., &Maulik, U. (2014). Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning. IEEE journal of translational engineering in health and medicine, 2, 1-11.
105. Jiang, B., Wu, X., Yu, K., & Chen, H. (2019, July). Joint semi-supervised feature selection and classification through Bayesian approach. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 3983-3990).
106. Liang, Y., Chai, H., Liu, X. Y., Xu, Z. B., Zhang, H., & Leung, K. S. (2016). Cancer survival analysis using semi-supervised learning method based on cox and aft models with l 1/2 regularization. BMC medical genomics, 9(1), 11.
107. M Dashtban, M Balafar, P Suravajhala (2018) Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics, 110(1), 10-17
108. OA Alomari, AT Khader, MA Al-Betar, LM Abualigah (2017) Gene selection for cancer classification by combining minimum redundancy maximum relevancy and bat-inspired algorithm. International Journal of Data Mining and Bioinformatics, 9(1), 32–51
109. I Jain, VK Jain, R Jain (2018) Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Applied Soft Computing, 62, 203-215
110. Y Prasad, KK Biswas, M Hanmandlu (2018) A recursive PSO scheme for gene selection in microarray data. Applied Soft Computing, 71, 213-225
111. S Kar, KD Sharma, M Maitra (2015) Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Systems with Applications, 42(1), 612-627
112. F Han, C Yang, YQ Wu, JS Zhu, QH Ling, YQ Song, DS Huang (2015) A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(1), 85-96
113. SS Shreem, S Abdullah, MZA Nazri (2014) Hybridising harmony search with a Markov blanket for gene selection problems. Information Sciences, 258, 108-121

114. FV Sharbaf, S Mosafer, MH Moattar (2016) A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics, 107(6), 231-238

115. CM Lai, WC Yeh, CY Chang (2016) Gene selection using information gain and improved simplified swarm optimization. Neurocomputing, 218, 331-338
116. L Gao, M Ye, X Lu, D Huang (2017) Hybrid method based on information gain and support vector machine for gene selection in cancer classification. Genomics, proteomics & bioinformatics, 15(6), 389-395
117. T Nguyen, A Khosravi, D Creighton, S Nahavandi (2015) Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification. PLoS ONE, 10(3), e012036
118. H Motieghader, A Najafi, B Sadeghi, A Masoudi-Nejad (2017) A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata. Informatics in Medicine Unlocked, 9, 246-254
119. Y Chen, Z Zhang, J Zheng, Y Ma, Y Xue (2017) Gene selection for tumor classification using neighborhood rough sets and entropy measures. Journal of biomedical informatics, 67, 59-68
120. Y Xiao, TH Hsiao, U Suresh, HIH Chen, X Wu, SE Wolf, Y Chen (2014) A novel significance score for gene selection and ranking. Bioinformatics, 30(6), 801–807
121. L Sun, XY Zhang, YH Qian, JC Xu, SG Zhang, Y Tian (2019) Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. Applied Intelligence, 49, 1245–1259
122. MB Kursa (2014) Robustness of Random Forest-based gene selection methods. BMC bioinformatics, 15, Article 8
123. KH Chen, KJ Wang, ML Tsai, KM Wang, AM Adrian, WC Cheng, TS Yang, NC Teng, KP Tan, KS Chang (2014) Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm. BMC Bioinformatics, 15, Article 49
124. H Deng, G Runger (2013) Gene selection with guided regularized random forest. Pattern Recognition, 46(12), 3483-3489
125. L Cleofas-Sánchez, JS Sánchez, V García (2019) Gene selection and disease prediction from gene expression data using a two-stage hetero-associative memory. Progress in Artificial Intelligence 8(1), 63-71

126. C Tang, L Cao, X Zheng, M Wang (2018) Gene selection for microarray data classification via subspace learning and manifold regularization. Medical & Biological Engineering & Computing, 56, 1271–1284

127. H Chen, Y Zhang, I Gutman (2016) A kernel-based clustering method for gene selection with gene expression data. Journal of Biomedical Informatics, 62, 12-20
128. SS Shreem, S Abdullah, MZA Nazri (2014) Hybridising harmony search with a Markov blanket for gene selection problems. Information Sciences, 258, 108-121
129. H Cai, P Ruan, M Ng, T Akutsu (2014) Feature weight estimation for gene selection: a local hyperlinear learning approach. BMC Bioinformatics, 15, Article 70

130. A Zibakhsh, MS Abadeh (2013) Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function. Engineering Applications of Artificial Intelligence, 26(4), 1274-1281
131. V García, JS Sánchez (2015) Mapping microarray gene expression data into dissimilarity spaces for tumor classification. Information Sciences 294, 362-375
132. Kotsiantis, S. B., Zaharakis, I., &Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.
133. Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., ... & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, *97*(1), 262-267.
134. Czajkowski, M., &Kretowski, M. (2019). Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. *Expert Systems with Applications*, *137*, 392-404.
135. Arevalillo, J. M., & Navarro, H. (2013). Exploring correlations in gene expression microarray data for maximum predictive–minimum redundancy biomarker selection and classification. Computers in biology and medicine, 43(10), 1437-1443.
136. Chan, W. H., Mohamad, M. S., Deris, S., Zaki, N., Kasim, S., Omatu, S., ...& Al Ashwal, H. (2016). Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme. Computers in biology and medicine, 77, 102-115.
137. Li, J., & Wang, F. (2016). Towards unsupervised gene selection: a matrix factorization framework. IEEE/ACM transactions on computational biology and bioinformatics, 14(3), 514-521.
138. V García, JS Sánchez, L Cleofas-Sánchez, HJ Ochoa-Domínguez, F López-Orozco (2017) An insight on the 'large G, small n' problem in gene-expression microarray classification. In Proc. of 8th Iberian Conference on Pattern Recognition and Image Analysis, pp. 483-490, Faro (Portugal)
139. Mramor, M., Leban, G., Demšar, J., &Zupan, B. (2005, July). Conquering the curse of dimensionality in gene expression cancer diagnosis: tough problem, simple models. In *Conference on Artificial Intelligence in Medicine in Europe* (pp. 514-523). Springer, Berlin, Heidelberg.
140. Abdulla, M., &Khasawneh, M. T. (2020). G-Forest: An Ensemble Method for Cost-sensitive Feature Selection in Gene Expression Microarrays. *Artificial Intelligence in Medicine*, 101941.
141. Lakshmanan, B., &Jenitha, T. (2020). Optimized Feature Selection and Classification in Microarray Gene Expression Cancer Data. *Indian Journal of Public Health Research & Development*, *11*(1), 347-352.
142. Rouhi, A., &Nezamabadi-Pour, H. (2020). Feature Selection in High-Dimensional Data. In *Optimization, Learning, and Control for Interdependent Complex Networks* (pp. 85-128). Springer, Cham.

1003

1004 **Table 1.** Comparison of Existing Reviews with the Current Survey

Reference	Description	Shortcomings
[14]	The survey focuses on the Supervised Gene Selection methods on Cancer Microarray dataset.	Concentrates only on Supervised Gene Selection methods.
[29]	The work discusses various works done in the Semi-Supervised Gene Selection methods, and the hierarchical structure of semi-supervised methods is also focused.	Concentrates only on Semi-Supervised Gene Selection methods.
[32]	The work focuses on the gene selection methods from a search strategy perspective.	Concentrates on search strategies in the feature selection methods
[42]	A survey on the filter-based feature selection techniques	Concentrates on filter-based techniques in cancer microarray data.
[46] [51]	The work concentrates on various feature selection methods in microarray data.	In general, focus on the feature selection methods did not categorize as supervised, unsupervised, or semi-supervised.
[55]	A survey on the hybrid-based gene selection techniques	Concentrates only on hybrid approach based gene selection methods.
Current Survey	Our survey on the existing literature focuses on the works mentioned above, categorizing into Supervised, Unsupervised, and Semi-Supervised Learning. Also, it discusses the performance of the existing gene selection methods.	

1005

1006 **Table 2.** Filter-based Supervised Gene Selection

Reference	Ideology	Gene Selection Algorithm	Classifier	Dataset	Performance Evaluation Metrics
[64]	The informative genes are selected with the help and Mutual Information, which are then used to train the	Mutual Information	SVM (Linear, Quadratic, RBE and Polynomial), KNN, ANN	• Colon • Cancer • Lymphoma	• Error Rate • LOOCV

	classifier.				
[65]	The Spearman Correlation and Distributed Filters have been used to select the most significant genes.	Spearman Correlation and distributed filter	Naïve Bayes, Decision Tree, SVM, and kNN	• Breast Cancer • Colon Cancer • DLBCL • SBRCT • Prostate Cancer • Lung Cancer	• Accuracy • Precision • Sensitivity • FMeasure • ROC
[66]	The proposed method is based on the Hilbert Schmidt Independence Criterion, and it achieves scalability to large datasets and high computational speed.	Sparse Hilbert-Schmidt Independence Criterion (SHS)	SVM and kNN	• Lymphoma • Leukemia • Brain Tumour • 11_Tumors • SRBCT • Lung	• Classification Accuracy
[67]	In this study, a new method of Normalised Mutual Information called Joe's Normalised Mutual Information (JNMI) had been developed and evaluated with five classifiers.	Joe's Normalized Mutual Information	Naïve Bayes, Radical Function Network, Instance-based Classifier, Decision-based Table and Decision Tree	• Leukemia • Lymphoma • CNS • MLL • SRBCT	• Accuracy AUC

Table 3. Wrapper-based Supervised Gene Selection

Reference	Ideology	Gene Selection Algorithm	Classifier	Dataset	Performance Evaluation Metrics
[68]	Aims to improve the evaluation time	Wrapper-based Sequential	kNN, Naïve Bayes, C4.5	• Colon • SRBCT • Leukemia	• Classification Accuracy • Wilcoxon

	with the help of Markov Blanket with Sequential Forward Selection.	Forward Selection with Markov Blanket	Decision Tree	• DLBCL • Prostate • Bladder • Gastric • Tox • Blastoma	signed-rank test
[69]	The proposed method Multiple Support Vector Machine – Recursive Feature Elimination is an enhancement of SVM-RFE for improving the accuracy in selecting the informative features.	MSVM-RFE	Random Forest, C4.5 Decision Tree	• Leukemia • Lung Cancer	• Classification Accuracy
[70]	A wrapper-based feature selection technique has been developed with Naïve Bayes by using the real-world high dimensional data in terms of difficulty due to noise.	Naïve Bayes-Wrapper	Naïve Bayes, MLP, 5NN, SVM and Logistic Regression	• Ovarian • ALL AML Leukaemia • CNS • Prostate MAT • Lymphoma • Lung Cancer	• AUC
[71]	This method aims to gather the relevant genes to distinguish the biological facts. The method is an extension of SVM-T-RFE,	SVM-Bayesian T-Test –RFE (SVM-BT-RFE)	SVM-RFE, SVM-T-RFE	• Colon • Leukemia • Medulla Blastoma • Lymphoma • Prostate	• Classification Accuracy

	where instead of a t-test, a Bayesian t-test has been used for better results.				
[72]	In this study, three wrapper based feature selections are implemented, and the results show that SVM-RFE-PSO performs better in selecting informative features than the other two.	SVM-RFE-GS, SVM-RFE-PSO, and SVM-RFE-GA	SVM	• Breast Cancer • TGCA	• AUC • Accuracy • Precision • Recall • F-Score

Table 4. Hybrid Supervised Gene Selection

Reference	Ideology	Gene Selection Algorithm	Classifier	Dataset	Performance Evaluation Metrics
[73]	Addresses the linear independence to find informative features with the help of matrix factorization and SVD.	Matrix Factorization based on SVD	Naïve Bayes, C4.5, and SVM	• Brain • CNS • Colon • DLBCL • GLI • Ovarian • SMK • Breast • Prostrate	• Cross-Validation (5-Fold and DOB-SCV) • Sensitivity • Specificity • Accuracy • G-Mean
[74]	The correlation coefficient is used as the attribute evaluator and PSO as a search strategy to select the necessary features.	Correlation Coefficient and PSO	ELM, J48, Random Forest, Random Tree,	• SRBCT • Lymphoma • MLL	• Classifier Accuracy

			Decision Stump, and Genetic Programming		
[75]	The Genetic Bee Colony combines the benefits of the Genetic Algorithm and Artificial Bee Colony. The method is evaluated using SVM.	Genetic Bee Colony	SVM	• Colon • Leukemia • Lung • SRBCT • Lymphoma	• Classification Accuracy • LOOCV
[76]	Two-stage feature selection methods involve the Laplacian Score and wrapper approach (SFS and SBS) to select the superior genes. Also, it considers the variance information.	Locality Sensitive Laplacian Score, Sequential Forward Selection and Sequential Backward Selection	SVM	• Acute Lymphoma • Lung Cancer • DLBCL • Prostrate • MLL Leukaemia • SRBCT	• Accuracy • Precision • Recall • F-Score • AUROC
[77]	This hybrid method targets at improving the classification accuracy with a two-stage method. It comprises the EGS (multi-layer and F-Score approach) as the first stage to reduce the noise and redundant features; in the second stage, AGA is used as a wrapper to select the informative genes used SVM and NB as fitness functions.	Multi-Layer Ensemble Gene Selection (EGS) and Adaptive Genetic Algorithm (AGA)	SVM and Naïve Bayes	• Breast • Colon • DLBCL • SBRCT • Lung • Leukemia	• Accuracy • FMeasure • Sensitivity
[78]	A hybrid gene selection method combining the ReliefF and the Ant Colony Optimization is proposed. It is a filter-	ReliefF-Ant Colony Optimization	RFACO-GS	• Colon • Leukemia • Lung • Prostrate	• Classification Accuracy

	wrapper based gene selection.				

Table 5. Ensemble-based Supervised Gene Selection

Reference	Ideology	Gene Selection Algorithm	Classifier	Dataset	Performance Evaluation Metrics
[80]	The three filter methods are made into an ensemble with the Union and Intersection of top n features, which are then further fine-tuned using the Genetic Algorithm.	Relief F, Chi-Square, and Symmetrical Uncertainty.	KNN, MLP, and SVM	• Colon • Lung • Leukemia • SRBCT • Prostrate	• Accuracy
[81]	The proposed method combines different individual rankings with various aggregation methods. The methods used are Chi-Square, InfoGain, mRMR, and ReliefF.	Ranker Ensemble	SVM-RBF Kernel	• Colon • DBCL • CNS • Leukemia • Lung • Prostate • Ovarian	• Error Rate
[82]	The Correlation based feature selection incorporating the Neighbourhood Mutual Information (NMI) and Particle Swarm Optimization (PSO) are combined into an ensemble (NMICFS-PSO) for cancer recognition.	NMICFS – PSO	SVM	• Breast • DLBCL • Leukemia • Lung • SRBCT	• LOOCV • Classification Accuracy
[83]	The authors have designed an ensemble based feature selection for a multi-class	Iterative Ensemble Feature Selection	SVM and kNN	• GLM • Lung • ALL • ALL-AML-4	• AUC

	classification problem. The study aims to show that balanced sampling and feature selection together assists in improving the results.	(IEFS)		• ALL-AML-3 • Thyroid	
[84]	A robust aggregator technique has been proposed by combining the reliability assessment and classification performance based on the expert algorithms' outputs.	Reliability Assessment-based Aggregation	kNN	• DLBCL • Bladder • Lymphoma • Prostate • Breast • CNS • Lung	• k-Fold Cross-Validation (k = 10)
[85]	A two-staged wrapper-based ensemble gene selection method has been implemented to identify the gene expression data's biomarkers. A filter-based approach and parallel metaheuristics were performed at every stage in the ensemble.	Ensemble of Co-operative Parallel Metaheuristics	-	• 9_tumours • 11_tumours • Prostate • Colon • Leukemia • Ovarian • DLBCL • SRBCT • Brain Tumour	• Accuracy • Jaccard Index • Kuncheva Index

Table 6. Embedded-based Supervised Gene Selection

Reference	Ideology	Feature Selection Algorithm	Classifiers	Datasets	Performance Evaluation Metrics
[86]	The IDGA uses Laplacian and Fisher score as ranking measures and a genetic algorithm to select the informative	Intelligent Dynamic Genetic Algorithm (IDGA)	KNN, SVM, Naïve Bayes	• SRBCT • Breast • DLBCL • Leukemia • Prostrate	• LOOCV

	features.				
[87]	The wrapper approach is embedded in the RMA algorithm to find informative features.	Wrapper based Recursive Memetic Algorithm	SVM, MLP, and KNNS	• AMLGSE2191 • Colon • Leukemia • MLL • SRBCT • Prostrate	• Accuracy • 5-Fold Cross-Validation • LOOCV
[88]	This study's embedded method is a two-stage method with feature selection and feature extraction, L1 regularization as the feature selection method, and Partial Least Square (PLS) as the feature extraction.	L1 Regularization	LDA	• GCM • MLL • GLIOMA • Lung • SRBCT • NCI60 • Breast • CLL-SUB-111 • GLA-BAR-180 • DLBCL	• Classification Accuracy • CPU Time • Sensitivity
[89]	This method targets minimizing the computational cost and maximizing the performance by selecting a minimal number of necessary genes. This method distinguishes the features by their occurrence frequency and classification performance.	Weighted Bacterial Colony Optimization	Sequential Minimal Optimization (SMO) and kNN	• Breast Cancer Wisconsin • CNS • Colon • Leukemia • 9_Tumours • 11_Tumours • Brain • SRBCT • Prostate • DLBCL	• Classification Error Rate • Classification Accuracy

[90]	With the scaling factors approach's help, the embedded strategy proposed in this study penalizes the feature cardinalities.	Kernel Penalised-Support Vector Data Description (KP-SVDD) and Kernel Penalised-Cost Sensitive Support Vector Machine (KP-CSSVM)	SVM	• GORDAN • GLIOMA • SRBCT • BHAT • CAR • BULL	• Classification Accuracy
[91]	The embedded approach proposed implements the adaptive LASSO, which focuses on solving the initial weight uncertainty issue	Adaptive LASSO (APLR)		• Colon • Prostrate • DLBCL	• AUC • Misclassification Error

Table 7. Filter-based Unsupervised Gene Selection

Reference	Ideology	Gene Selection Algorithm	Classifier	Dataset	Performance Evaluation Metrics
[92]	A new filter based unsupervised gene selection method, which can be used for numerical and non-numerical data, has been proposed. It is a combination of a spectrum based feature evaluation and a kernel.	Unsupervised Spectral Feature Selection Method (USFSM)	SVM, kNN, and Naïve Bayes	• Heart • Liver • Dermatology Thoracic	• AUROC • Accuracy K-Fold (k=5)
[93]	HSIC is a framework for unsupervised gene	Hilbert-Schmidt	Gap-Statistics	• Several microarray datasets	• Accuracy

	selection that considers the dependency maximization among the similarity matrices after eliminating a gene.	Independence Criterion (HSIC)	and k-Means		
[94]	The Ant Colony Optimization is used as a Filter approach to maximize the relevance scores among the genes and minimize the redundancy.	Microarray Gene Selection based on Ant Colony Optimization (MGSACO)	SVM, Naïve Bayes, and Decision Tree	• Colon • Leukemia • SRBCT • Prostate • Lung Cancer	• Classification Error Rate
[95]	An unsupervised gene selection by implementing the sparse filtering and sample learning as a filter approach was proposed. It takes into consideration deep structures, which helps in obtaining improved results.	Sample Learning based on Deep Sparse Filtering (SLDSF)	-	• DLBCL • Lung Cancer • Leukemia • Esophageal Cancer (ESCA) • Squamous cell Carcinoma Head and Neck (HNSC)	• p-Values of GO terms

Table 8. Wrapper-based Unsupervised Gene Selection

Reference	Ideology	Gene Selection Algorithm	Classifier	Dataset	Performance Evaluation Metrics
[96]	In this study, a standard Support Vector Machine – Recursive Feature Elimination was performed on microarray data to distinguish low-risk and high-risk colon cancer patients.	SVM-RFE	SVM	• GSE38832 • GSE17538 • GSE28814 • TGCA	• AUROC • Accuracy • K-Fold (k=5)

Table 9. Hybrid Unsupervised Gene Selection

Reference	Ideology	Gene Selection Algorithms	Classifier	Dataset	Performance Evaluation Metrics
[97]	A filter-wrapper based hybrid gene selection method having the properties of spectral feature selection, Laplacian Score Ranking, and enhanced Calinski-Harabasz Index.	Laplacian Score Ranking – Weighted Normalised Calinski Harabasz (LS – WNCH)	k-Means	• Lymphoma • Tumors • Leukemia	• Jaccard Index
[98]	To solve the issue of high-dimension and the search space, a filter-wrapper based hybrid gene selection	Feature Selection based on Binary Ant System	SVM, kNN, and Naïve Bayes	• Colon • Leukemia	• Accuracy • FMeasure • Recall • Precision

	has been proposed with a clustering and improved Binary Ant System.	(FSCBASM)

Table 10. Ensemble and Embedded Unsupervised Gene Selection

Reference	Ideology	Gene Selection Algorithms	Classifier	Dataset	Performance Evaluation Metrics
[99]	A reconstruction based gene selection has been proposed to perform a data independent filter-based gene selection embedded in the approach. (Embedded)	Reconstruction-based Unsupervised Feature Selection (REFS)	k-Means	• Lung • GLIOMA	• Accuracy • Normalized Mutual Information
[100]	The RCE was constructed with a random set of features different bootstrap samples at each partition. The out-of-bag feature importance was calculated from every ensemble partition. (Ensemble)	Random Cluster Ensemble (RCE)	k-Means	• Leukemia • Ovarian Lung	• Accuracy • Normalized Mutual Information (NMI)

1029 **Table 11.** Semi-Supervised Gene Selection Approaches

Reference	Ideology	Gene Selection Algorithm	Classifier	Dataset	Performance Evaluation Metrics
[101]	An SVM-based semi-supervised gene selection technique has been proposed. The results show better performance in terms of accuracy and process time than other standard Supervised gene selection techniques. (Wrapper)	Semi-Supervised SVM-based RFE (S^3VM-RFE)	SVM	• Lung Cancer	• K-fold Cross-Validation (k = 10)
[102]	A filter-based feature selection called SemiMM was proposed, which handles mutual information, local structure, and variance at the same time. It is a combination of mutual information and spectral graph. (Filter)	Semi-Supervised Maximum Discriminative Local Margin (SemiMM)	SVM	• DLBCL • Prostate • Tumor • Leukemia2 • SRBCT • Lung Cancer	• Accuracy • Precision • Recall • FMeasure • AUC
[103]	The authors have proposed an ensemble-based framework aiming to improve the quality of the clustering model. The double selection cluster ensemble feature selection assists in selecting the most relevant genes. (Ensemble)	Modified Double Selection-based Semi-Supervised Cluster Ensemble (MDSVM-SSCE)	PC-K-means Clustering approach.	• Tumors	• Normalized Mutual Information (NMI)
[104]	The SVM model has been combined with Fuzzy Rough Set as a Semi-Supervised approach to select the informative	Kernalised Fuzzy Rough Set (KFRS) S^3VM	Transductive SVM (TSVM)	• SRBCT • DLBCL • Leukemia • MicroRNA	• T-Statistics • Wilcoxon Signed-Rank test • AUC • FMeasure

	features. The proposed algorithm proves to be capable of selecting useful biomarkers from the datasets. (Hybrid)				
[105]	An embedded method, with the Bayesian approach. It automatically chooses the informative features and also trains the classifier. (Embedded)	Joint Semi-Supervised Feature Selection and Algorithm (JSFS)	Bayesian approach to select and classify	• Prostate • Colon	• Accuracy
[106]	The most widely adopted for high and low-risk classification. The L ½ regularization has been embedded in these models to select appropriate and relevant genes to enhance the models' performance. (Embedded)	L ½ Regularization	CoX and AFT Models	• Tumour	• Precision

Table 12. Performance Analysis of Prostate Dataset

Category	Literature	Performance Analysis	Type pf Metric Used	Selected No. of Genes
Supervised Feature Selection	[65]	99.81%	Accuracy	-
	[68]	88.3%	Accuracy	12
	[71]	99.64%	Accuracy	20
	[73]	92%	Accuracy	25
	[76]	86.76%	Accuracy	52

	[80]	98.03%	Accuracy	7
	[81]	2.94	Error Rate	89
	[84]	82%	10-Fold CV	20
	[86]	100%	LOOCV	18
	[87]	95.1%	5-Fold CV	5
Unsupervised Gene Selection	[94]	26.85	Error Rate	20
Semi-Supervised Gene Selection	[102]	90%	Accuracy	150
	[105]	91%	Accuracy	30

Table 13. Performance Analysis on Leukaemia Dataset

Category	Literature	Performance Analysis	Type pf Metric Used	Selected No. of Genes
Supervised Feature Selection	[66]	98.61%	Accuracy	1000
	[68]	95.5%	Accuracy	4
	[75]	100%	LOOCV	200
	[76]	97.79%	Accuracy	-
	[77]	94.34%	Accuracy	13
	[80]	100%	Accuracy	12
	[81]	14.71	Error Rate	5
	[82]	99%	10-Fold CV	15

	[85]	100%	Accuracy	25
	[86]	94.1%	Accuracy	15
	[87]	96.1%	5-Fold CV	5
Unsupervised Gene Selection	[94]	23.07	Error Rate	20
	[98]	94.8%	FMeasure	40
	[109]	97.2%	Accuracy	3
Semi-Supervised Gene Selection	[100]	95%	Accuracy	150
	[104]	98%	Accuracy	20

Table 14. Performance Analysis of Colon Dataset

Category	Literature	Performance Analysis	Type pf Metric Used	Selected No. of Genes
Supervised Feature Selection	[64]	11	Error Rate	200
	[65]	99.1%	Accuracy	13
	[68]	82.9%	Accuracy	1000
	[71]	99.5%	Accuracy	25
	[73]	93%	Accuracy	15
	[75]	96.7%	Accuracy	5
	[77]	83.54%	Accuracy	5
	[80]	100%	Accuracy	-

	[81]	20	Error Rate	15
	[87]	100%	5-Fold CV	12
	[89]	98.25%	Accuracy	4
Unsupervised Gene Selection	[94]	23.63	Error Rate	20
	[98]	95%	Accuracy	40
Semi-Supervised Gene Selection	[105]	87%	Accuracy	120

Table 15. Acronyms

Acronyms

EGS-AGA	Multi-Layer Ensemble Gene Selection (EGS) and Adaptive Genetic Algorithm (AGA)
LSLS-SFS	Locality Sensitive Laplacian Score, Sequential Forward Selection
PSO	Particle Swarm Optimization
SVD	Singular Vector Decomposition
SVM-RFE-GS	Support Vector Machine-Recursive Feature Elimination-Grid Search
SVM-BT-RFE	Support Vector Machine-Bayesian T test-Recursive Feature Elimination
NB	Naïve Bayes
MSVM-RFE	Multiple Support Vector Machine-Recursive Feature Elimination

SFS-MB	Sequential Forward Selection-Markov Blanket
NMI	Normalized Mutual Information
IDGA	Intelligent Dynamic Genetic Algorithm
KP-SVDD	Kernel Penalised-Support Vector Data Description
BCO	Bee Colony Optimization
RMA	Recursive Memetic Algorithm
RCE	Random Cluster Ensemble
REFS	Reconstruction-based Unsupervised Feature Selection
FSCBASM	Feature Selection based on Binary Ant System
LS-WNCH	Laplacian Score – Weighted Normalized Calinski-Harabasz
SLDSF	Sample Learning based on Deep Sparse Filtering
MGSACO	Microarray Gene Selection based on Ant Colony Optimization
HSIC	Hilbert-Schmidt Independence Criterion
USFSM	Unsupervised Spectral Feature Selection Method
MDSVM-SSCE	Modified Double Selection based Semi-Supervised Cluster Ensemble
JSFS	Joint Semi-Supervised Feature Selection
KFRS-S3VM	Kernalised Fuzzy Rough Set
Semi-MM	Semi-Supervised Maximum Discriminative Local Margin

1044

Figure 01.JPEG

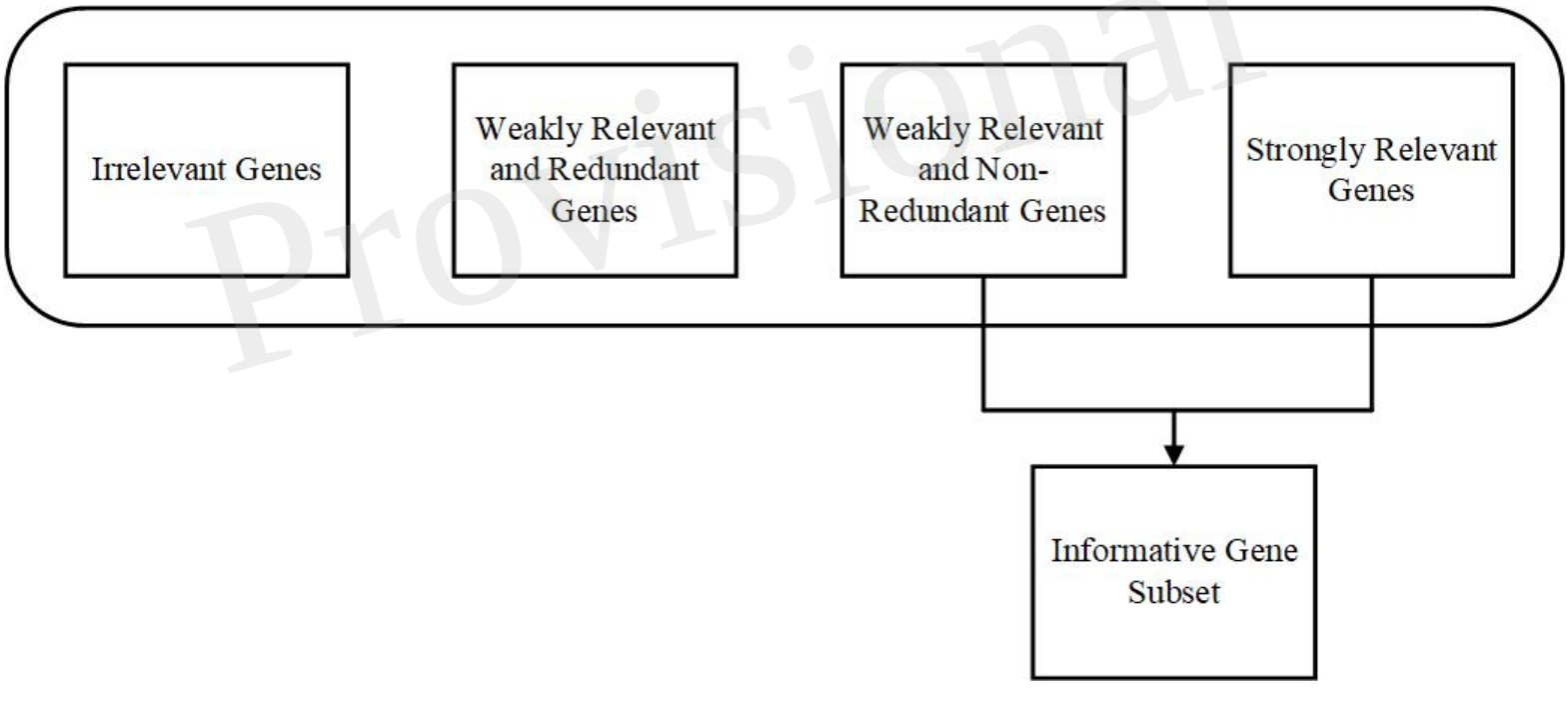

Figure 02.JPEG

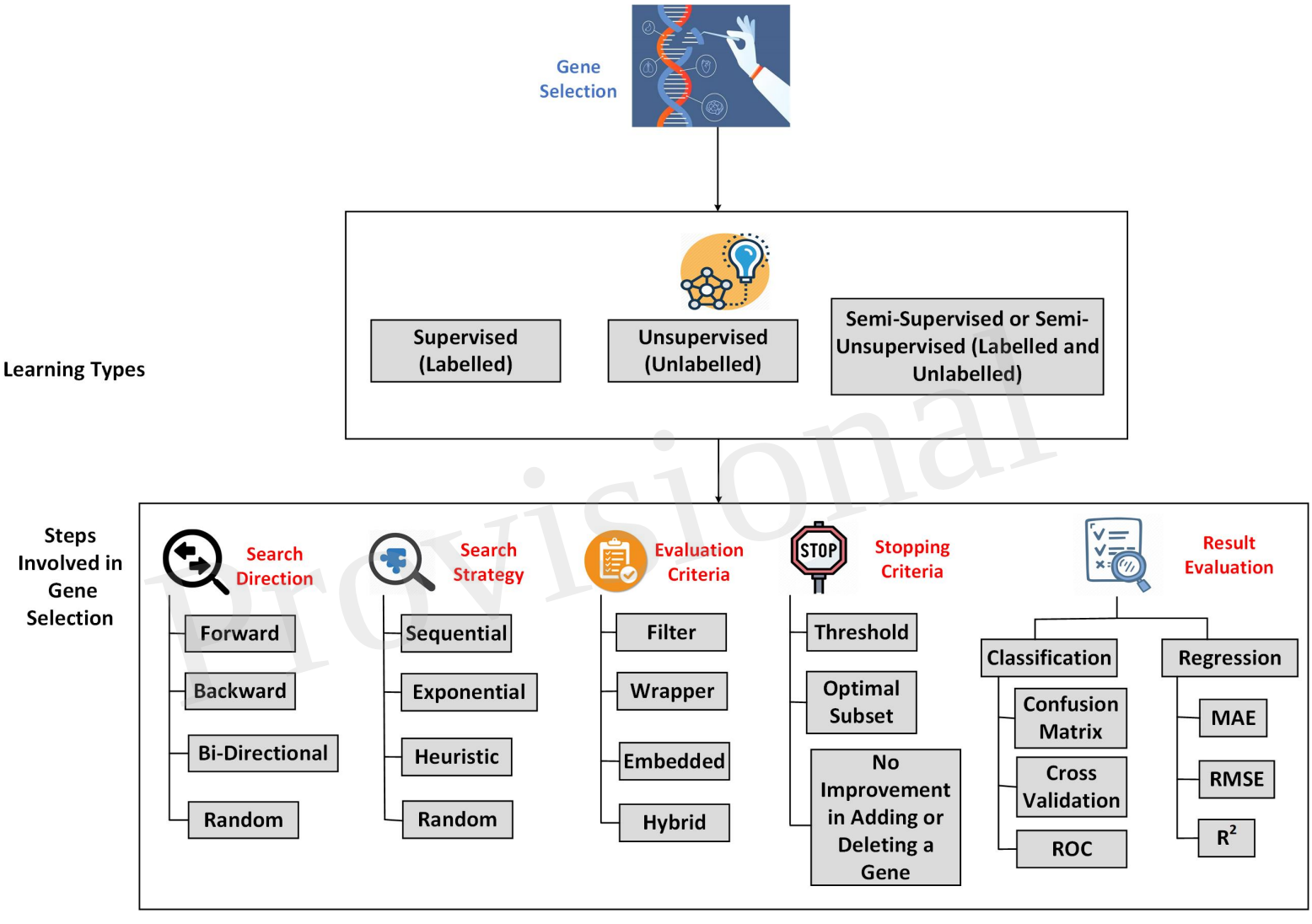

Figure 03.JPEG

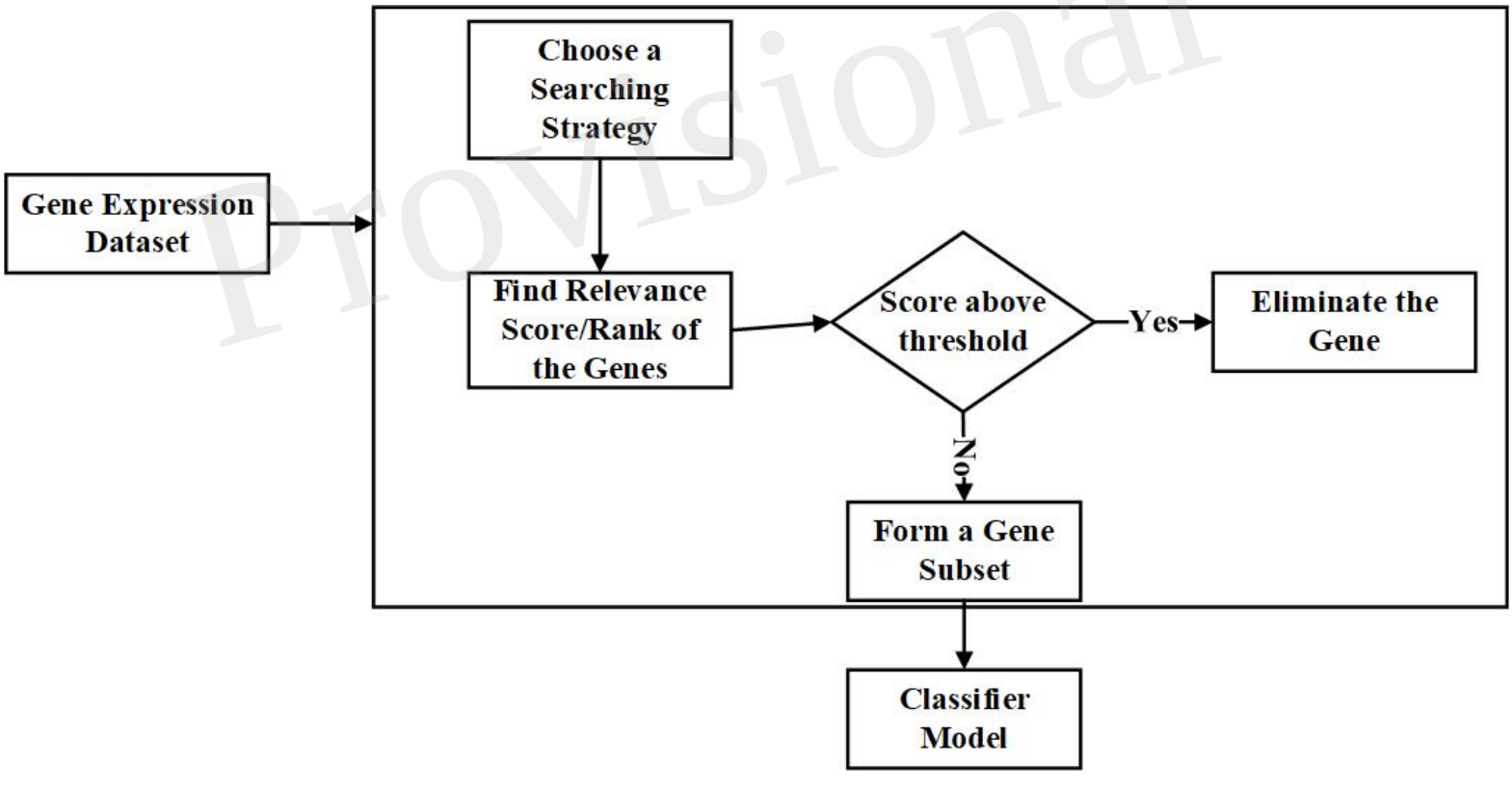

Figure 04.JPEG

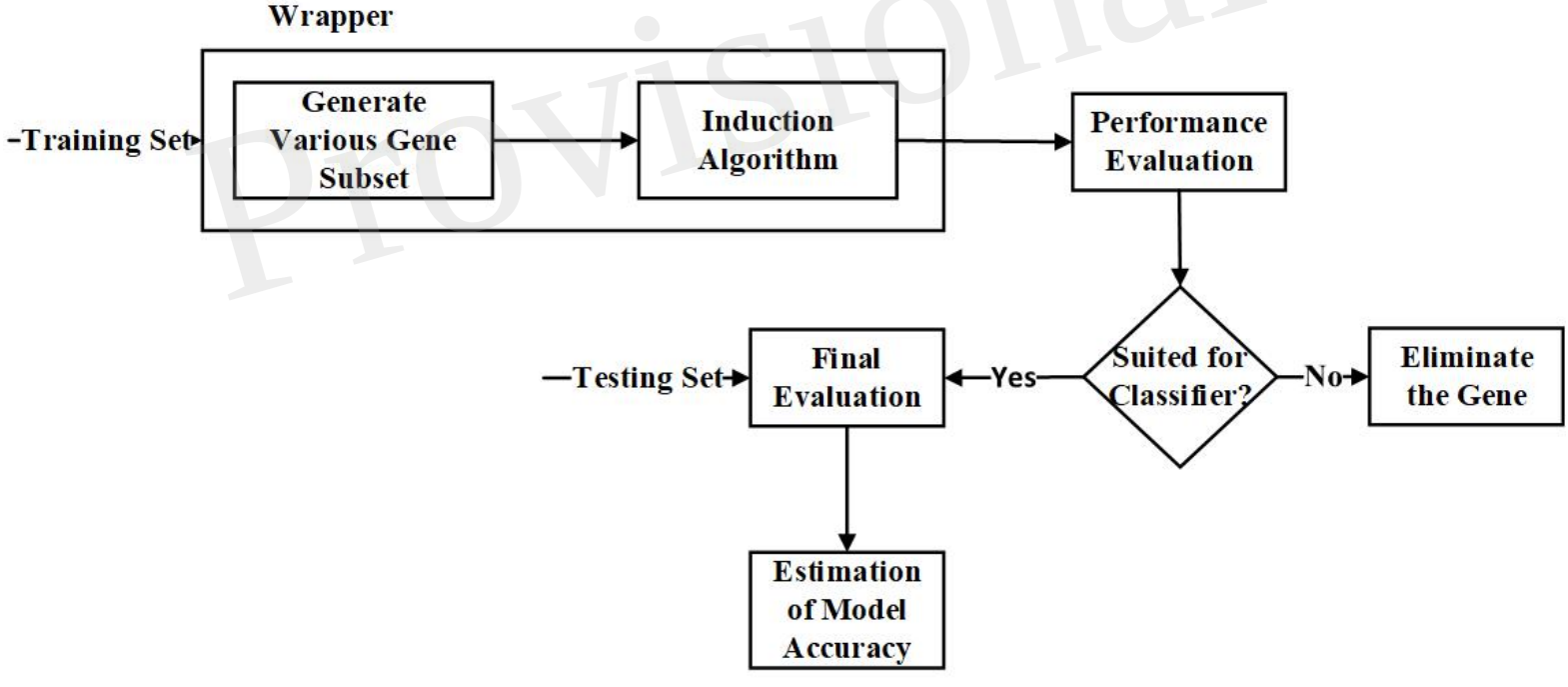

Figure 05.JPEG

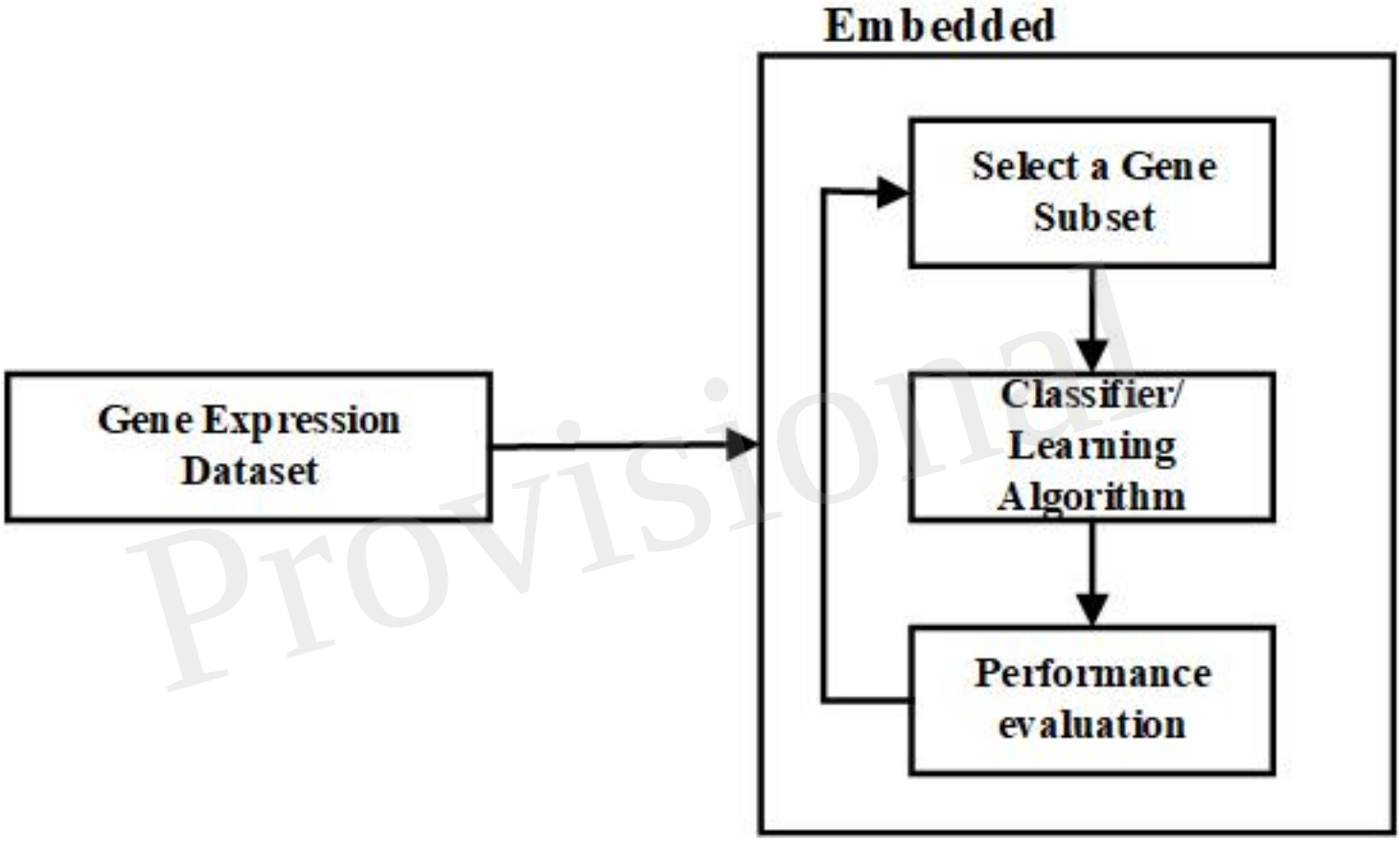

Figure 06.JPEG

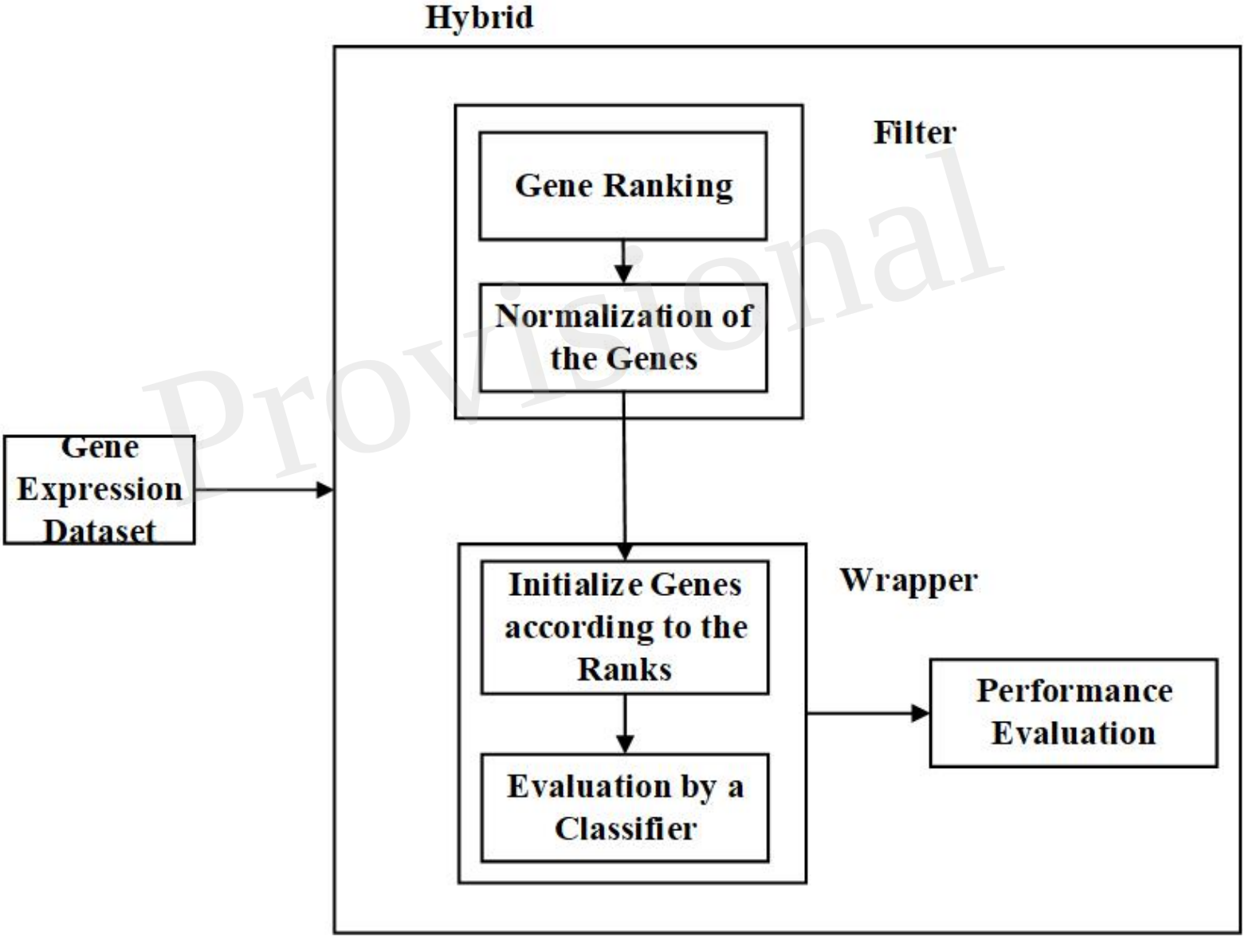

Figure 07.JPEG

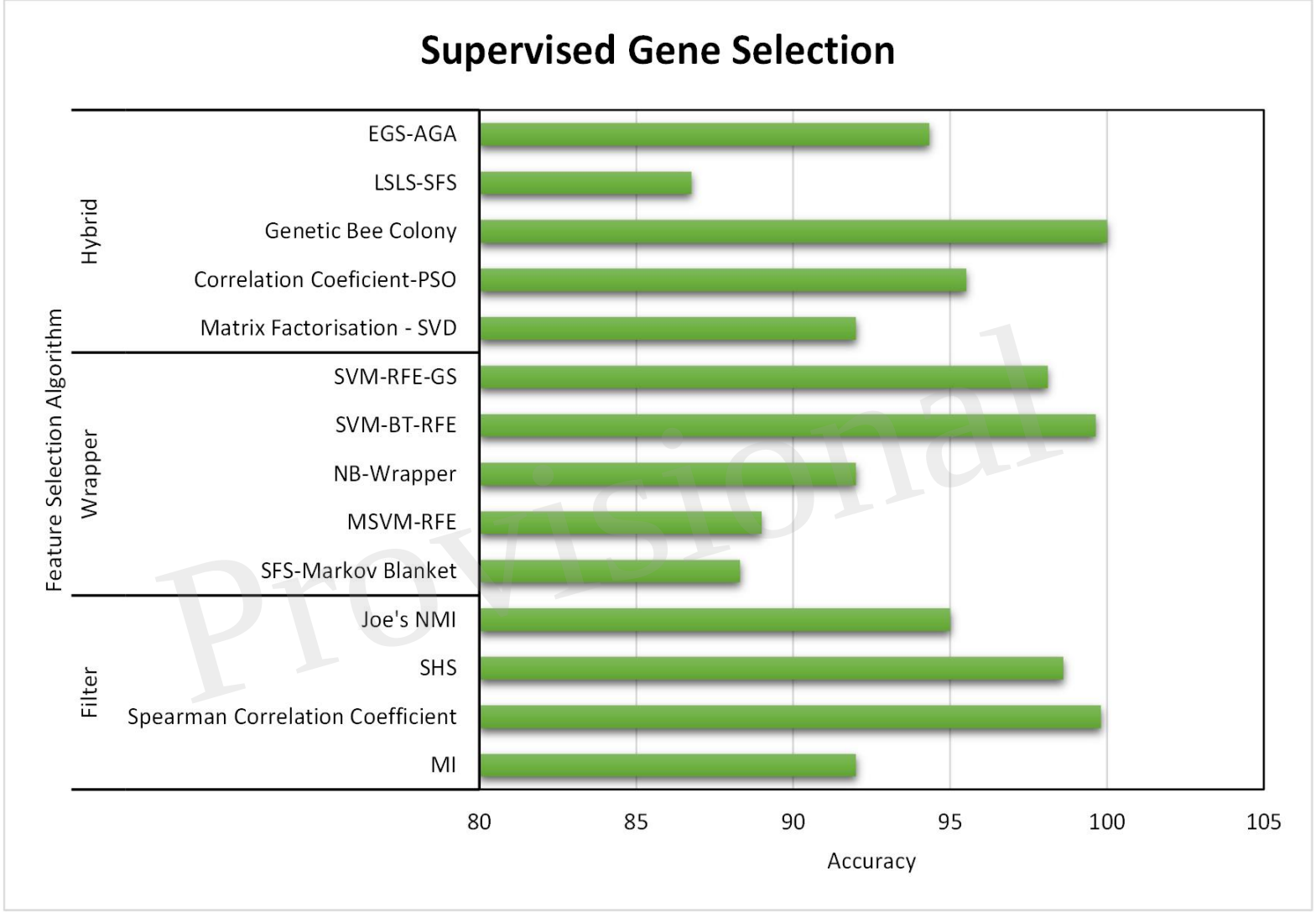

Figure 08.JPEG

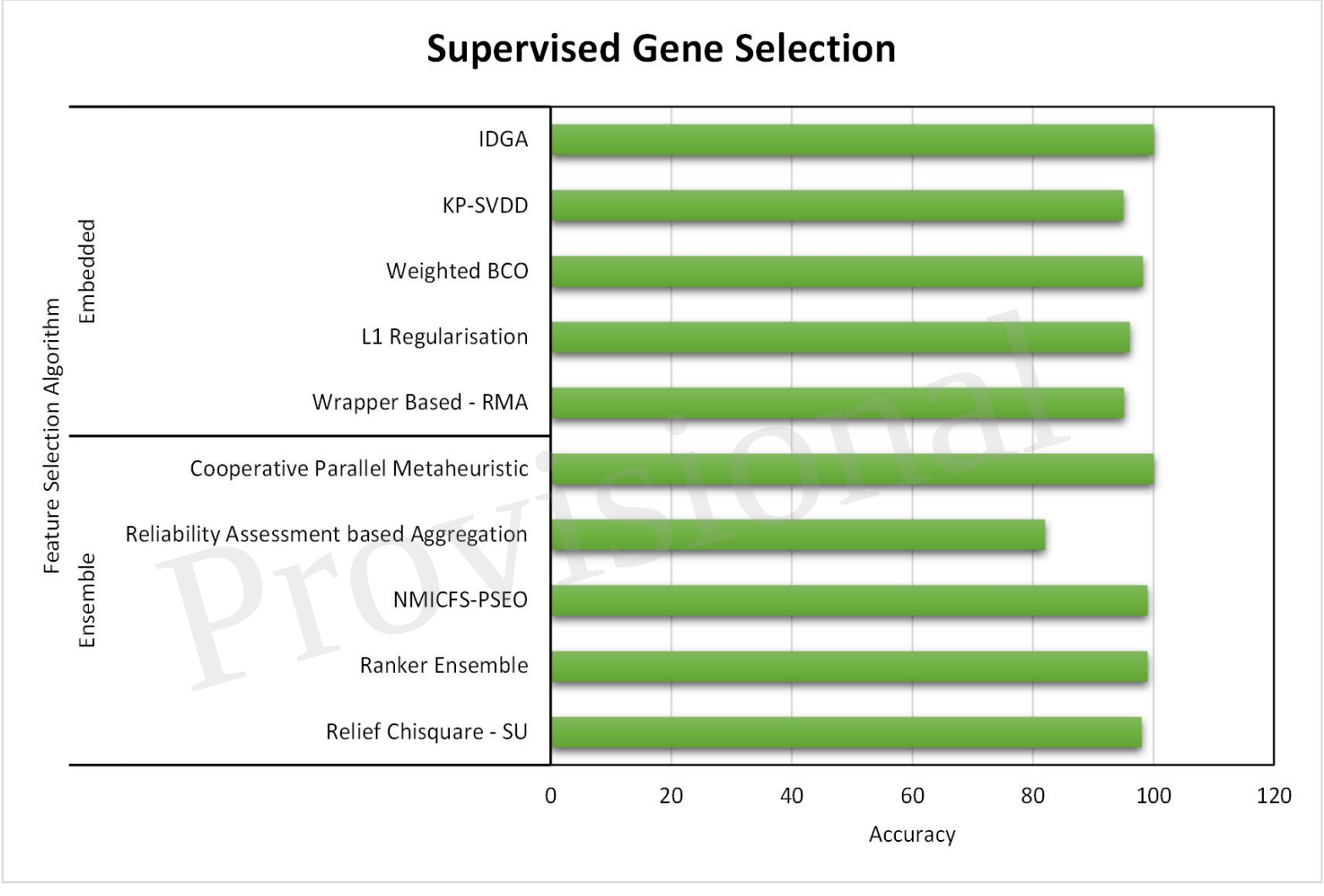

Figure 09.JPEG