IJCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Vol. 8, Issue.10, October 2020 E-ISSN: 2347-2693

A Survey on Feature Selection in Microarray Data: Methods, Algorithms and Challenges

Khadija Abdullah Uthman[1*], Fadl Mutaher Ba-Alwi[2], Suad Mohammed Othman[3]

[1,2,3]Dept. of Computer Science, Faculty of Computer & Information Technology (FCIT), Sana'a, University, Sana'a, Yemen

[*]*Corresponding Author: Khadejacso@gmail.com, Tel.: +967-738482998*

DOI: https://doi.org/10.26438/ijcse/v8i10.106116 | Available online at: www.ijcseonline.org

Received: 04/Oct/2020, Accepted: 12/Oct/2020, Published: 31/Oct/2020

Abstract**—** In biomedical researches a massive amount of data are produced day after day, using machine learning algorithms to discover the knowledge is very important in early diagnosis, prevention and treatment, as well as drug development. Biomedical data like DNA microarray suffers from curse of dimensionality phenomenon, since there are a huge number of features (genes) with high ambiguity. Feature selection is still a hot topic which cares about reducing the high of dimensionality by applying different techniques. Different contributions are conducted with new models, frameworks, methodologies and algorithms aiming to dissolve the curse of dimensionality problem and produce more meaningful and reliable data. The objective of this study is to explain the concept of feature selection, its methods, the algorithms and techniques that have been recently used in microarray data and the most popular microarray datasets were used. Moreover, the challenges that can appear when selecting more informative and non-redundant features from high dimensional datasets.

***Keywords* —** Feature Selection, Filter Method, Wrapper Method, Hybrid Method, DNA microarray, Metaheuristic.

I. INTRODUCTION

There is an urgent need for researchers in machine learning and bioinformatics community to get reliable and accredited data as a prerequisite step before analyzing these data ,whereas retrieving meaningful information from high dimensional datasets is a challenging task[1] in fields like detection of fraud, finance, prediction of diseases at early stages, intrusion detection system ..etc.[2][3]. Pre-processing step come into being, it has gain a great importance in solving some problems in datasets such as noisy instances and class-imbalance. Moreover, sometimes there is a need to reduce the dimensionality by getting rid of irrelevant and redundant features which can mislead the learning algorithms and reduce the classification performance[4][5].

Dimensionality reduction can be achieved by feature extraction or feature selection, by applying these methods the high dimensional data transform into a meaningful presentation [6]. In Feature extraction a new feature space will be constructed with low features of stronger discriminating power[7]. In applications such as image processing and information retrieval feature extraction is an idealistic choice.

On the other hand, Feature selection is one of the popular technique in the face of curse of dimensionality problem which has come up as a result of advancement of data collection[1]. The aim of feature selection is to obtain the most discriminable feature subset by deleting irrelevant and redundant features[4], Figure.1 explains the process of FS. It is a typical choice where the original features are important in knowledge extraction process in applications like gene expressions data, media, image processing. The main challenge with these datasets is the modest number of instances compared with the huge number of features[8]. Both feature selection and feature extraction have the capability of improving performance, increasing computational efficiency, decreasing memory storage, and building better generalization models [9].

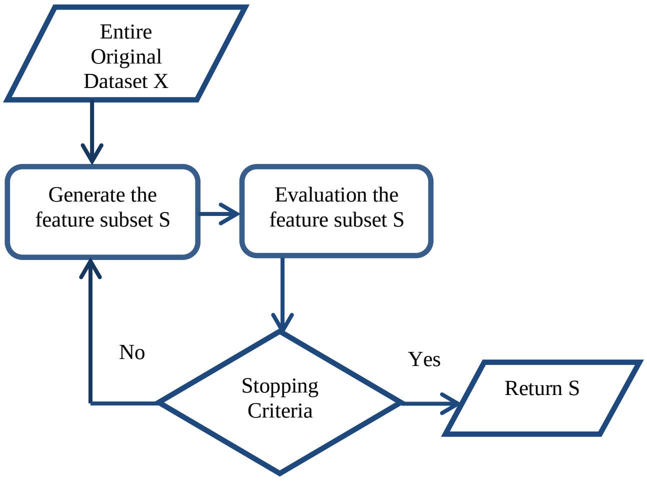
Entire Original Dataset X
Generate the feature subset S
Evaluation the feature subset S
Stopping Criteria
No
Yes
Return S

Figure.1: The process of feature selection.

The algorithms that are used in feature selection can be categorized into supervised and unsupervised learning, in supervised method class labels work as a clear guidance to the feature selection process which make the supervised methods more reliable. However, unsupervised method become more challenging task due to the absence of class label[10]. The advantage of unsupervised classification in microarray data is possibility to find new tumor subtypes, but the disadvantage is that the existing tumor type could be not utilize, The advantage of supervised classification is being able to learn from existing cases[11].

Features can be categorized into: 1)- strong relevant features which always be in the final optimal feature subset 2)-weak relevant non-redundant features which can be included in the final optimal subset 3)-Irrelevant features, these features should be discarded from the optimal subset because they do not add any useful information in the final model, finally there are 4)-redundant features, with this type of features no extra information can be added than the currently selected features[9][12], Figure.2 shows the types of features.

By applying feature selection algorithms relevant feature should be included, either they are strong or weak relevant , however redundant and irrelevant features should be excluded[13]. The main task in microarray gene expression data is to discover the most relevant genes which can cause the disease as well remove redundancy, a small number of genes in datasets can play a major role in predicting type of cancer[2].

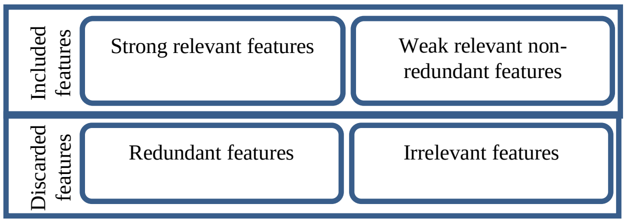

Figure.2: Types of features in any dataset

Before applying FS techniques some dataset especially real datasets need to handle the missing values issues, remove noise and discretization[1]. The performance of learning algorithms can be degrade in presence of redundant and irrelevant features, since these type of features will increase the size of search space then make generalization more complicated[6]. Feature selection methods can be broadly categorized into three main methods filter, wrapper and embedded in addition to hybrid and ensemble methods[4].

The rest of this paper is organized as follows: In section II
Some related work will be discussed. Section III gives more information about DNA microarray data and the most common datasets that have been used in the literature. Describing the FS methods with recent contributions are presented in Section IV. In section V metaheuristic algorithms in FS are highlighted. Section VI gives more discussion about related work. Finally section VII concludes the paper with future works.

II .RELATED WORK

Machine learning and biological researchers have conducted several researches in feature (gene) selection to obtain optimal subset of features with non-redundant and relevant features which can improve the classification algorithms and increase the accuracy.

While conventional feature selection methods need to be enhanced or integrated with other method to obtain better results some researchers have combined between filter and wrapper method. Two phase classification model was developed in[14], most relevant features were selected by integrating ReliefF with a wrapper based approach called Recursive Binary Gravitational Search Algorithm(RBGSA) using multinomial Naïve Bayes classifier, at each iteration(RBGSA) transforms a very raw feature space to an optimized one until reaching specific criteria, ReliefF-RBGSA-MNB proposed model has better accuracy comparing with ReliefF-RBGSA method in all tested datasets. Authors in[15] have developed a wrapper-filter based called Recursive Memetic Algorithm(RMA) model for gene selection, the performance of the proposed method RMA has measured on seven microarray datasets, by testing the accuracy using RMA model the results were classifiers independent.

New studies have been proposed by combining different methods and algorithms , The researchers in [16] proposed a gene selection method for data classification via adaptive hypergraph embedded dictionary learning, the learned dictionary was used to represent original genes with a reconstruction coefficient matrix, instead of using the original feature space, the proposed method was compared with other state-of-the-art gene selection methods to verify it and appeared better result with multi class datasets. Whereas in[5] they proposed a method by combining maximal information entropy (MIE) to determine the dependencies among features and the maximal information coefficient (MIC) to scale the correlation between a feature subset and the class, then two searching strategy were applied using the binary particle swarm optimization algorithm BPSO and sequential forward selection SFS as search strategies, the outcomes have proven that the mMIE-mMIC algorithm showed a strong advantage over other FS algorithms such as relief and CHI, however the proposed method was not tested using high dimensional datasets.

A novel feature selection method was proposed by[17] to discriminate the prominent features utilizing both sparse representation which aim to determine a small number of features to preserve the classification accuracy and information theoretic dependence analysis, the method was tested in one high dimensional datasets and showed intractable computational performance. Moreover, Three feature selection techniques were used in[18] to pick up the most informative features ROC, T -

test and Wilcoxon, the authors proposed a novel error correcting output code(ECOC) algorithm for classifying multiclass microarray data which is considered as a hard task ,the proposed method was more insensitive when changing the feature subset sizes.

Further, a novel neighborhood rough sets and entropy measure-based for gene selection with Fisher score for tumor classification was proposed in [19], the Fisher score method carry out preliminary dimensionality reduction by discard irrelevant genes to reduce the complexity of computation, for handling the uncertainly and noisy of gene expressions some neighborhood entropy-based uncertainty measures are investigated, in most cases the proposed method outperformed other methods after modified some parameters, but in Leukemia datasets the method need to be modified. Recent works on feature selection in microarray data will be discussed in the following sections.

III. DNA MICROARRAY DATA

Humans have many cells, each cell contains a complete copy of genome which is encoded into Deoxyribo Nucleic Acid (DNA)[14]. DNA microarray emerging technology is used recently to collect information from tissue and cell samples to produce valuable amount of data[15]. While direct research investigation of this data in laboratory is very expensive and tough, computational techniques have become an apparent need[16] see Figure.3.

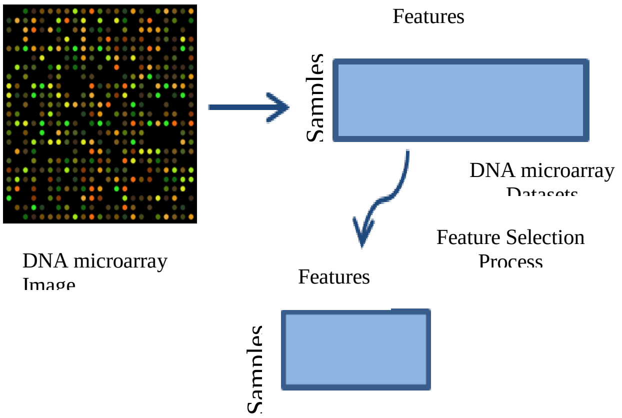

Figure.3: Feature Selection process Applying toDNA microarray

classification or clustering process is very important for diagnosing disease or identifying distinctive tumor types as well as identifying genes that can be targeted by drugs[12], it is difficult to classify microarray data regarding the small size of samples and high number of features with potential to contain irrelevant and redundant features, in addition to have noise and variability features due to the experimental complications[14].

Building a classification model in bioinformatics and machine learning domains depending on analysis of tumor gene expression data has become a hot topic[11],the characteristic of microarray dataset which are high dimensionality, small sample size and imbalanced distribution are considered substantial challenges when using this type of dataset.

Biologist with microarray data have two fundamental tasks the first one is to separate healthy patient with benign tumor from cancer patients which known as binary approach, while the second is to distinguish between different types of tumors which called multiclass approach[7].

A. Microarray Datasets:-

Microarray dataset is generally represented as N by M matrix, where N is the number of samples (rows) and M is the number of features (columns)[17]. The main characteristic of Microarray dataset is the large number of features and few number of samples (in most cases less than 100), in addition to complex interaction between genes[18]. Handling such a large number of features for just a few samples is a challenge for machine learning and bioinformatics researchers since the possibility of occurrence the “false positives” that can be raised by chance[3], Table.1 shows the most common datasets that have been used in the literature .

Table.1: The most used datasets in the literature

Dataset	Samples#	Features#	classes#	Where used
SRBCT	83	2308	4	[11] [19] [20] [21] [22]
DLBCL	77	5469	2	[11] [21][22] [23] [24]
Prostate	102	12600	2	[11] [19] [25] [26] [27]
Lung	203	12600	5	[11] [20] [25] [26] [27]
Lymphoma	62	4026	2	[11] [18] [24] [25] [27]
Ovarian	253	15,154	2	[14] [18] [27] [28] [29]
Colon	62	2,000	2	[21][27] [29][30] [31]
Breast	97	24,481	2	[18][28] [29][32] [33]
Leukemia	72	7129	2	[25][31] [33][34] [35]

Most microarray datasets suffer from imbalanced data which is widely common in applications related to fields like bioinformatics, face recognition, fraud detection, gene expressions, text and image classification. Data has been called imbalanced when data is highly-skewed distribution between its classes[35] it happens when the number of instances of one class is much higher than the number of other classes[37]. Based on filter method a new method was proposed in[38] to handle high dimensional gene expression data, the authors have proved that imRelief can correct the bias towards to the dominant classes, and take into account the scattered distributional characteristic of minority class samples, the proposed method has been applied using four microarray gene expression datasets ,it achieved better result with different measures such as Recall, Precision and Acc comparing with other methods. While in[35] they proposed the fused case-control screening for ultrahigh dimensional imbalanced data, the proposed method was model-free with low computational cost**.**

IV. CATEGORIZE OF FS METHODS

There are mainly three traditional methods for feature selection: Filter, Wrapper and embedded method. Furthermore, two other methods can be defined as a Hybrid method which means combining between more than one method and an ensemble method which generate different subsets before obtaining the final optimal subset[4][9][13], see figure.4. Comparison of traditional Feature Selection Methods is shown in Table.2.

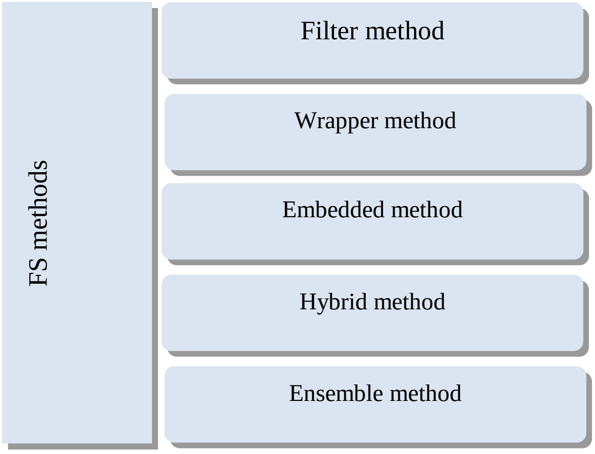

Figure.4:Catogerize of FS methods

1 .Filter method

The optimal features are selected according to some evaluation criteria without considering any learning approach[23], Figure.5 shows the process of filter method.

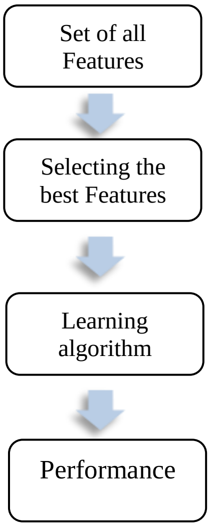

Figure.5: The process of Filter method

The advantages lie in its simplicity and light computations cost. Filter method uses independent feature evaluation measures like distance, dependency (or probability), consistency and information (or uncertainty)[39], this method can consider each feature(gene) individually known as univariate filter or group of genes which known as multivariate filter. There are variety of techniques or algorithms in filter method like minimum redundancy maximum relevancy(mRMR)[40], correlation based feature selection(CFS), and feature ranking techniques like symmetrical uncertainty(SU)[29], ReliefF[20], information gain and chi-square [39].

Graph theoretic and community detection concept have been utilized in[29] depending on filter based FS using graph technique, they have used symmetric uncertainty concept for visualizing the feature search space as a graph, the classification accuracy has increased when applying this method comparing with whole features in the selected microarray datasets and other FS methods except some datasets like lung and Colon cancer. Otherwise, researchers in[41] have conducted comparative study between some filters which are F test, T test, Signal to noise ratio (S/R), ReliefF and Pearson product-moment correlation coefficient(CC), they have evaluated the result using (KNN),(SVMs),(LDA),(DTC) and (NV) classifiers, the result proved that using S/R selection method and the KNN classifier have given the highest accuracies for different datasets. Further, in [32] the authors proposed a heuristic filter feature selection methods called Xvariance and Mutual Congestion, Xvariance depends on the internal specifications of features like mean and variance to classify labels, however Mutual Congestion is frequency based ,the experimental result has showed that Mutual Congestion increased the accuracy in high dimensional datasets while Xvariance worked well with standard datasets .

2 .Wrapper method

Evaluating the candidate subset based only on the classification results, more detail in Figure.6.

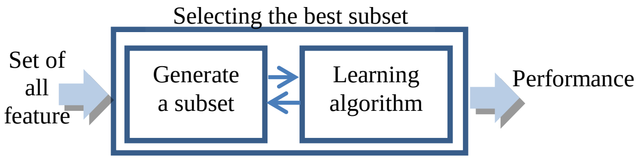

Figure.6: The process of wrapper method.

In other words, classification is performed for each candidate subset, which can achieve the most effective result in term of classification performance while enduring a heavy computational burden[25]. Any classification technique can be integrated with the wrapper, since wrapper method considers a classification algorithm as a black box[22]. Iterative search, heuristic search and random search techniques[12] are considering within wrapper based methods regarding the fact that searching for optimal subset of features is very hard and time consuming especially with huge number of features, wherefore metaheuristic algorithms are used to solve FS problem which will be discussed later in section five. A wrapper based approach using Artificial Bee Colony(ABC) was proposed in[42], the result of selecting genes from DNA microarray using the proposed method with KNN and DT classifiers has increased the accuracy performance comparing with GA and PSO in Lung and Lymphoma datasets, while achieved the same accuracy with PSO in Leukemia dataset.

New wrapper based feature selection method has been proposed in[43] called binary teaching learning based optimization(FS-BTLBO), in the proposed algorithm common controlling parameters and a number of generation are needed to obtain the optimal subset of features, the proposed method improved the classification accuracy when compared with original features and GA algorithm ,but authors did not test their method with high dimensional datasets. Data set with large search space make the optimization process harder which motivated authors in[25] to design a wrapper based feature selection method based on the binary variant of Dragonfly Algorithm (BDA), the authors have got good result by using DA comparing with their literature ,while BGSA algorithm outperform the proposed method in Brain_Tumor1 and Prostate_Tumor dataset. A novel framework was proposed in [44] for gene selection and classification in microarray data, ANOVA statistics was used to select the relevant genes and sort them according to their p-values, then an evolutionary wrapper-based approach using the principles of enhanced Jaya(EJaya) algorithm and forest optimization algorithm (FOA) was proposed, the classification results obtained by applying the proposed method with binary class datasets achieved 100% accuracy rate with Ovarian cancer dataset and 99.87% with multi class Lymphoma-3 dataset.

3. Embedded method

Combining the feature selection process and training of the classifier, Figure.7 explains the process, this method perform feature selection in the process of training and is usually specific to a given learning machines[45]. That is, FS is accomplished automatically during the training of the classifier.

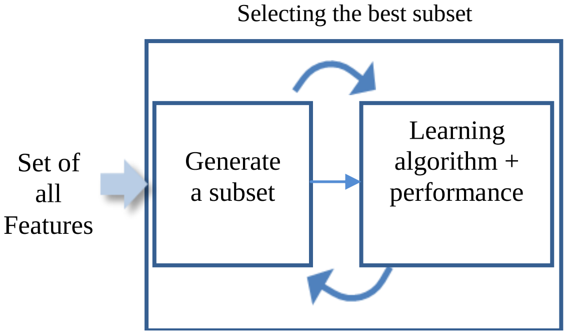

Figure.7: The process of Embedded method

An embedded FS method based on SVM called (1-norm SVMSL) was proposed in[46] using four microarray datasets, the authors have used Recall measure to evaluate the performance of the proposed method , the SVM-RFE method outperforms the proposed method in classification performance ,but it consumes more time in ranking genes while the proposed method has an acceptable performance comparing with other methods. Also, based on embedded method researchers in [47] improved Support vector machine recursive feature elimination(SVM-RFE) which suffers from high computational complexity by merging feature clustering, the proposed FCSVM-RFE consists of three stages: clustering, representation and ranking genes,it reduced the running time ,but still need to enhance the accuracy classification in colon tumor and DLBCL datasets.

In order to classify tumor a novel method has been proposed in[11] called (SGL-SVM) utilizing Spare Group Lasso and support vector machine as a classifier, they applied Kruskal-Wallis rank sum test, then a spare group lasso has been used for further selection, the experimental result explained the relationship between the number of features and the classification accuracy which increase when the number of features increase, the proposed method need to be enhanced with less number of features.

We compare between filter, wrapper and embedded method in Table.2 using different factors (for simplicity, computational cost and accuracy we use high/low criteria while Yes/No for interact with classifier and risk of overfitting, finally practical or not practical when dealing with large datasets are used with scalability).

Table 2: Comparison of FS methods

	Filter	Wrapper	Embedded
Simplicity	high	low	low
Computational cost	low	high	high
accuracy	low	high	high
Scalability	practical	Not practical for large dataset	
Interact with classifier	No	Yes	Yes
Risk of overfitting	No	Yes	Yes

4 .Hybrid method

Hybrid methods combine two or more algorithms to find a new strategy to solve a specific problem[2]. Most researchers in the literature have used hybrid method because they can obtain better performance by using more than one method. In [31] they proposed a novel approach by combining Spearman's Correlation (SC) and distributed filter FS methods, from the fact that sometimes when applying conventional methods some irrelevant ranked features that could perform well regarding classification accuracy can be neglected, the authors have used Spearman's Correlation(SC) first to rank top features, then they have used similarity based, information theory based and statistical based as a filter method, the proposed method with SVM has exceeded other methods in the study regarding classification accuracy with low execution time. A hybrid method was proposed by[20]combining Relief and stacked autoencoder approaches for dimension reduction, then they used support vector machine(SVM) and convolutional neural networks (CNN) for classify samples using microarray datasets, by using Relief in this model the accuracy was increased and achieved better result comparing with SAE methods.

Another study was proposed in[2] using a Parallelized hybrid feature selection (HFS) method to improve the classification accuracy utilizing parallel programming frameworks, The proposed hybrid method combined between parallelized correlation feature selection CFS as a filter method and rank-based feature selection(RFS) methods as a wrapper method, the final results show that using parallel spark has given better results than using wekaspark. In[24] they have proposed a hybrid filter-wrapper method called rMRMR-MBA, they have combined Robust Minimum Redundancy Maximum Relevancy (rMRMR) as filter to select the more relevant genes (features) with modified bat algorithm (MBA) as wrapper for searching of informative genes, they have evaluated MBA with and without TRIZ optimization operators.

5 .Ensemble method

Different FS techniques are run in ensemble method, each technique generates a separate feature subset, then the ensemble method will combine the resulting feature subset to form the final feature subset[8]. An ensemble-based feature selection technique was proposed by[8] based on a Nested Genetic Algorithm by combining the information from two type of microarray data Gene Expression data and DNA methylation data, they have used two filter and wrapper methods, in filter method they have used t_test as a preprocessing step then a nested genetic algorithm with two genetic algorithm have been used, one with a support vector machine SVM and the other with a neural network, they have been used as a wrapper feature selection method, eventually they used incremental FS as an ensemble approach, the proposed approach showed that the selected subset of features improved the classification performance when they compared their result with other approaches. Another study based on parameter free greedy ensemble attribute selection method is proposed utilizing the concept of rough set theory[1] from medical datasets, this research aimed to combine multiple subsets produced by different rough set filter then producing optimal subset as a final result, they have used accuracy, precision and recall measurements to evaluate the proposed method, despite achieving good result in most cases but the datasets were not high dimensional. An ensemble method for diagnosis of breast and lung cancer was proposed in[48], for feature extraction Principal Component Analysis (PCA) was used, Pearson Correlation Coefficient (PCC) and Chi Square (Chi2) have been considered using set union operation, then Naive Bayes, K Nearest Neighbours, Decision Tree classifier were used, the proposed model suffered from complexity which can make the classification process more difficult when dealing with high dimensional datasets. Moreover, an ensemble of filter methods were developed in[39] by considering the union and intersection of the top-n features of filter methods, ReliefF, chi-square and symmetrical uncertainty, in the next step they have used Genetic Algorithm, three classifiers were applied multi-layer perceptron(MLP), support vector machine (SVM), and K-nearest neighbour (K-NN), the highest accuracy has achieved with SVM classifier except lung dataset which scored highest accuracy when using MLP classifier. Finally in[49] authors proposed a novel filter–wrapper hybrid ensemble feature selection approach based on the weighted occurrence frequency and the penalty scheme, in this study the researchers have modified some parameters to get high accuracy result from the selected features.

We have summarized in Table 3 some recent works on feature selection in microarray data including: methods, FS techniques, classifiers and tools that they have used, as well published year and the highest classification accuracy they have achieved.

Table 3: Related works on feature selection in microarray data

No.	Method	Year	Feature Selection Techniques	Classifiers	Tools	Highest accuracy
[25]	wrapper-based	2020	Binary Dragonfly Algorithm	KNN	MATLAB 2017a	94.1%
[20]	hybrid methods	2020	Relief and stacked autoencoder approaches	(SVM) and convolutional neural networks (CNN)	Matlab R2017b + Weka	96.14%
[1]	ensemble	2020	rough set concept / kNN imputation method	Naıve Bayes, decision trees , random forest	JAVA + WEKA	96.77%
[2]	hybrid	2019	parallelized correlation/ rank-based feature selection methods	Decision tree , Random forest	spark	96.77%
[19]	Hybrid	2019	Fisher score method, a joint neighborhood entropy-based	KNN, C4.5 and LibSVM	Matlab + WEKA	98.40%
[31]	Hybrid	2019	Spearman's Correlation / distributed filter FS methods	support vector machine, naïve Bayes, k-nearest neighbor, and decision tree	MATLAB 2016	98.98%
[8]	ensemble	2019	t-test, Genetic Algorithm	SVM N-Net(NNW)	unknown	99.9%
[50]	hybrid	2019	artificial bee colony , The stochastic diffusion search	SVM	unknown	89.02%
[48]	hybrid	2019	Pearson Correlation Coefficient , Chi2	NB, DT, KNN, SVM	WEKA	97.8%
[38]	filter	2019	imRelief	kNN	unknown	99.29%
[27]	Hybrid	2019	artificial bee colony algorithm and genetic algorithm	SVM	Matlab R2016a	98.91%
[44]	Hybrid	2019	ANOVA, enhanced Jaya (EJaya) algorithm, forest optimization algorithm (FOA)	SVM	MATLAB 2015a	99.87%
[32]	Wrapper _based	2019	Xvariance , Mutual Congestion	SVM NB DT	unknown	95%
[24]	hybrid filter/ wrapper	2019	robust Minimum Redundancy Maximum Relevancy (rMRMR) / modified bat algorithm (MBA)	SVM	Java with weaka tool and Matlab	100%
[28]	Hybrid	2018	graph regularized subspace learning method for gene selection / projection matrix	Support Vector Machine, Random Forest, and k-Nearest Neighbor	Matlab R2013b	98.14 %
[36]	Hybrid	2018	CMIM / Adaptive genetic algorithm	Extreme learning machine / SVM /Knn	unknown	96.96%
[34]	Hybrid	2018	multi-layer approach and f-score approach	(SVM) and Naïve Bayes (NB)	unknown	99.03%
[41]	filter	2018	F test, T test, Signal to noise ratio (S/R), ReliefF and Pearson product-moment correlation coefficient (CC)	K Nearest Neighbors (KNN), Support Vector Machines (SVMs), Linear Discriminant Analysis (LDA), Decision Tree for Classification (DTC) and Naïve Bayes classifier (NV)	unknown	100%
[52]	Hybrid	2017	Markov Blanket models	Naïve Bayes , SVM ,MB	unknown	97.8%

V. METAHEURISTIC ALGORITHM IN FS

While finding an optimal subset of features is considering an NP-hard problem[39] metaheuristic algorithms are used to solve many real world problem like feature selection[25],it applies to cope the drawback of non-metaheuristic algorithms like trapping in local or weak solutions[12]. From the main popular metaheuristic algorithms are genetic algorithm[49], particle swarm optimization(PSO)[53], artificial bee colony(ABC)[27], ant colony optimization (ACO),Bacterial Foraging Optimization (BFO), and Gravitational search algorithm (GSA)[14], teaching learning based optimization(TLBO),Sequential Forward Search (SFS), and Sequential Backward Search (SBS). In gene expression metaheuristic algorithms are used to analyze and interpret data[24]. Two hybrid algorithms were used in [54] an enhanced firefly algorithm based meta-heuristic(EFA) and adaptive neuro neutrosophic inference system(ANNIS) classifier , the (EFA) has been used to distinguish prescient genes for breast cancer prediction, then the authors showed that(ANNIS) classifier produced higher exactness consequences Comparing with other classifier such as SVM, BLASSO, FNN, and ANFIS. A hybrid bio-inspired framework was proposed in[23] using two powerful metaheuristic approaches (C-HMOSHSSA), researchers have used multi-objective spotted hyena optimizer(MOSHO) and salp swarm algorithm (SSA), the calculation of (MOSHO) is used for maintaining the necessary information wherefore it requires low computational efforts, while (SSA) maintains diversity. From the most used metaheuristic algorithms in the literature are:-

1-Genetic algorithm (GA)

It is a bio-inspired metaheuristic belonging to the category of evolutionary algorithms[39]. A two-stage MI-GA Gene Selection algorithm for selecting informative genes was proposed in [33], Mutual Information-based gene selection is applied first which selects only the genes that have high information related to the class, then the output of first stage will be the input to the second stage which applied the Genetic algorithm to identify and select the final optimal set, the highest classification accuracy appeared in Ovarian Cancer dataset. Moreover, nested genetic algorithms was used in[8] (outer and inner), the Outer GA with SVM worked on Microarray gene expression, while the Inner GA worked on DNA Methylation data. To reduce the computational complexity authors in [55] have used Genetic algorithm with CFS in feature selection stage, they have observed that the quality of search methods improved when the highly correlated features were ignored ,by applying this method the classification accuracy has improved ,but the number of features that they have used were limited.

2-Particle Swarm Optimization (PSO)

It is a swarm-intelligence algorithm that simulate the swarm behaviors[25], the PSO algorithm provides a global search method, but this does not guarantee that the result converges to the global best[11]. Different versions of PSO were used to solve feature selection problem in [5] they combined binary particle swarm optimization BPSO algorithm with other algorithm as a search strategy,using BPSO as a search strategy has achieved good classification accuracy. Also, a recursive PSO approach for gene selection[56] was proposed, different filter based ranking strategies were combined, the average accuracy values using a RPSO was better than using whole features in colon and breast cancer dataset . In[57] they applied PSO to tackle the curse of dimensionality problem by selecting subset of features and used the evolutionary outlay aware deep belief network to classify the datasets , the result has shown that the proposed method achieved high prediction accuracy with all datasets comparing with other methods, also it has taken less time for the classification, the proposed method need to be tested in high dimensional microarray datasets .

3- Artificial bee colony algorithm (ABC)

The ABC is a swarm intelligence inspired by insect behavior, the idea is that every source (position) of food will be a possible solution for any problem that needs to be optimized and the amount of nectar will correspond to the fitness of the solution [50]. A novel feature selection algorithm based on artificial bee colony algorithm and genetic algorithm was proposed in [27], the novel algorithm was applied using six public datasets and scored better classification accuracy than using GA and ABC separately. Another study[53] have used PSO with SVM to eliminate inefficient genes, then they compared the result with Artificial Bee Colony(ABC) with SVM which was more efficient.

VI. DISCUSSION

In most biological studies feature (gene) selection is a preliminary stage, hence selecting a discriminate subset of genes which can improve the classification process[20] is becoming a mandatory step. A small number of genes can play a vital role in biological studies which can help in tumor classification and early diagnosis of disease. The conventional methods in feature selection are not enough to deal with datasets like DNA microarray, because this type of data have a huge number of features comparing with small samples which can affect the classification process in addition to ambiguity interacting between genes, all these problems are motivated researches to broaden the circle of search to obtain the optimal subset of genes by creating new methods or combining existing modified methods. In this study researches in DNA microarray data or gene expression data are included. The study interested in new research which applied new methods trying to select the optimal or near optimal subset which can later improve the classification accuracy. In the literatures the researches have used different tools to implement their studies, most of them have used one tool or programming language, but some of them have used more than one tool to improve their results[18]. There are different tools for analyzing microarray data which contain variety of packages were written to handle this type of data [15](Figure.8 has more details) .

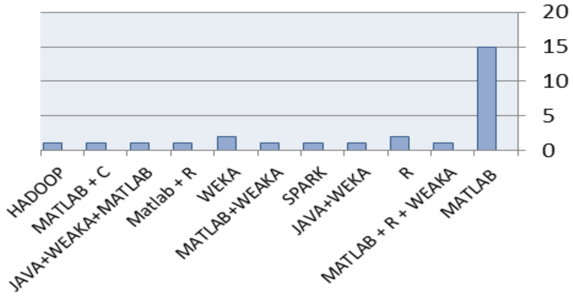

Figure.8: The tools used in the literature

To evaluate the performance of classification task different measured are used like Accuracy, specificity, sensitivity and Matthews Correlation Coefficient (MCC)[31].

When dealing with Microarray datasets some challenges have to be taken into consideration:

1- Microarray data are imbalanced data, since most samples (instances) belong to healthy cases which can affect the classification, prediction or clustering problems. Some researches in the literature do not mention how they deal with this problem.
2- Microarray data can be presented as a structured data like group, graph and tree which can increase the understanding of interaction between genes. Dealing with structured data needs to enhance traditional methods or investigate novel methods.
3- Microarray data suffer from noise data, which need to apply some preprocessing techniques before applying Feature selection, since the number of samples is low, deleting any samples to get rid of noise can affect the learning process.

VII. CONCLUSION

In some fields the characteristics of data make the process of selecting features more complex and time consuming which need to enhance or modify new models or techniques .Building simpler and more comprehensive models in machine learning and biomedical researches depends on the quality of data which will be used in learning task. This data should be cleaned, informative and non-redundant. Therefore data quality is a critical subject in DNA microarray data classification, prediction and clustering. Feature selection is a process that can produce high quality of data by deleting redundant and irrelevant features and including most relevant and informative data. In this paper a study on feature selection in DNA microarray is presented starting with the traditional methods and the recent studies related to these methods. It presents Metaheuristic Algorithms as a search strategy to obtain optimal subset and cope the complexity of finding more reliable features. Moreover, it describes the DNA microarray technology and some popular datasets that have been used in the literatures. Finally some challenging are presented especially those which are related to finding optimal features in data which are noise, huge and have different structured. As a future work working in parallel environment by combining between other methods can achieve better classification accuracy and decrease the number of features, also more studies can be conducted in FS methods on structured data such as graph, group and tree.

REFERENCES

[1] R. K. Bania and A. Halder, "R-Ensembler: A greedy rough set based ensemble attribute selection algorithm with kNN imputation for classification of medical data," *Computer Methods and Programs in Biomedicine,* vol. **184**, p. **105122**, **2020**.

[2] L. Venkataramana, S. G. Jacob, R. Ramadoss, D. Saisuma, D. Haritha, and K. Manoja, "Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data," *Genes & genomics,* vol. **41**, pp. **1301-1313**, **2019**.

[3] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in biology and medicine,* p. **103375**, **2019**.

[4] K. Tadist, S. Najah, N. S. Nikolov, F. Mrabti, and A. Zahi, "Feature selection methods and genomic big data: a systematic review," *Journal of Big Data,* vol. **6**, p. **79**, **2019**.

[5] K. Zheng, X. Wang, B. Wu, and T. Wu, "Feature subset selection combining maximal information entropy and maximal information coefficient," *Applied Intelligence,* pp. **1-15**, **2019**.

[6] N. Sánchez-Maroño, O. Fontenla-Romero, and B. Pérez-Sánchez, "Classification of Microarray Data," in *Microarray Bioinformatics*, ed: Springer, **2019**, pp. **185-205**.

[7] A. Alonso-Betanzos, V. Bolón-Canedo, L. Morán-Fernández, and B. Seijo-Pardo, "Feature Selection Applied to Microarray Data," in *Microarray Bioinformatics*, ed: Springer, **2019**, pp. **123-152**.

[8] S. Sayed, M. Nassef, A. Badr, and I. Farag, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets," *Expert Systems with Applications,* vol. **121**, pp. **233-243**, **2019**.

[9] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, *et al.*, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR),* vol. **50**, p. **94**, **2018**.

[10] E. Hoseini and E. G. Mansoori, "Unsupervised feature selection in linked biological data," *Pattern Analysis and Applications,* vol. **22**, pp. **999-1013**, **2019**.

[11] Y. Huo, L. Xin, C. Kang, M. Wang, Q. Ma, and B. Yu, "SGL-SVM: A novel method for tumor classification via support vector machine with sparse group Lasso," *Journal of Theoretical Biology,* vol. **486**, p. **110098**, **2020**.

[12] A. K. Shukla, D. Tripathi, B. R. Reddy, and D. Chandramohan, "A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges," *Evolutionary Intelligence*, pp. **1-21**, **2019**.

[13] N. AlNuaimi, M. M. Masud, M. A. Serhani, and N. Zaki, "Streaming Feature Selection Algorithms for Big Data: A Survey," *Applied Computing and Informatics,* **2019**.

[14] X. H. Han, D. A. Li, and L. Wang, "A Hybrid Cancer Classification Model Based Recursive Binary Gravitational Search Algorithm in Microarray Data," *Procedia Computer Science,* vol. **154**, pp. **274-282**, **2019**.

[15] M. Ghosh, S. Begum, R. Sarkar, D. Chakraborty, and U. Maulik, "Recursive memetic algorithm for gene selection in microarray data," *Expert Systems with Applications,* vol. **116**, pp. **172-185**, **2019**.

[16] X. Zheng, W. Zhu, C. Tang, and M. Wang, "Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning," *Gene,* vol. **706**, pp. **188-200**, **2019**.

[17] Y. Zhang, Q. Zhang, Z. Chen, J. Shang, and H. Wei, "Feature assessment and ranking for classification with nonlinear sparse representation and approximate dependence analysis," *Decision Support Systems,* vol. **122**, p. **113064**, **2019**.

[18] M. Sun, K. Liu, Q. Wu, Q. Hong, B. Wang, and H. Zhang, "A novel ECOC algorithm for multiclass microarray data classification based on data complexity analysis," *Pattern Recognition,* vol. **90**, pp. **346-362**, **2019**.

[19] L. Sun, X.-Y. Zhang, Y.-H. Qian, J.-C. Xu, S.-G. Zhang, and Y. Tian, "Joint neighborhood entropy-based gene selection method with fisher score for tumor classification," *Applied Intelligence,* vol. **49**, pp. **1245-1259**, **2019**.

[20] S. Kiliçarslan, K. Adem, and M. Çelik, "Diagnosis and Classification of Cancer Using Hybrid Model Based on ReliefF and Convolutional Neural Network," *Medical Hypotheses,* p. **109577**, **2020**.

[21] G. Agapito, "Computer Tools to Analyze Microarray Data," in *Microarray Bioinformatics*, ed: Springer, **2019**, pp. **267-282**.

[22] J. Chaki and N. Dey, "Pattern analysis of genetics and genomics: a survey of the state-of-art," *Multimedia Tools and Applications,* pp. **1-32**, **2019**.

[23] A. Sharma and R. Rani, "C-HMOSHSSA: Gene selection for cancer classification using multi-objective meta-heuristic and machine learning methods," *Computer methods and programs in biomedicine,* vol. **178**, pp. **219-235**, **2019**.

[24] M. A. Al-Betar, O. A. Alomari, and S. M. Abu-Romman, "A TRIZ-inspired bat algorithm for gene selection in cancer classification," *Genomics,* **2019**.

[25] M. Mafarja, A. A. Heidari, H. Faris, S. Mirjalili, and I. Aljarah, "Dragonfly algorithm: theory, literature review, and application in feature selection," in *Nature-Inspired Optimizers*, ed: Springer, **2020**, pp. 47-67.

[26] C. Kang, Y. Huo, L. Xin, B. Tian, and B. Yu, "Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine," *Journal of theoretical biology,* vol. **463**, pp. **77-91**, **2019**.

[27] J. Ge, X. Zhang, G. Liu, and Y. Sun, "A Novel Feature Selection Algorithm Based on Artificial Bee Colony Algorithm and Genetic Algorithm," in *2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)*, **2019**, pp. **131-135**.

[28] C. Tang, L. Cao, X. Zheng, and M. Wang, "Gene selection for microarray data classification via subspace learning and manifold regularization," *Medical & biological engineering & computing,* vol. **56**, pp. **1271-1284**, **2018**.

[29] S. Bakhshandeh, R. Azmi, and M. Teshnehlab, "Symmetric uncertainty class-feature association map for feature selection in microarray dataset," *International Journal of Machine Learning and Cybernetics,* vol. **11**, pp. **15-32**, **2020**.

[30] S. A. Medjahed, T. A. Saadi, A. Benyettou, and M. Ouali, "Kernel-based learning and feature selection analysis for cancer diagnosis," *Applied Soft Computing,* vol. **51**, pp. **39-48**, **2017**.

[31] A. K. Shukla and D. Tripathi, "Identification of potential biomarkers on microarray data using distributed gene selection approach," *Mathematical biosciences,* vol. **315**, p. **108230**, **2019**.

[32] M. Alirezanejad, R. Enayatifar, H. Motameni, and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets," *Genomics,* **2019**.

[33] M. J. Rani and D. Devaraj, "Two-Stage Hybrid Gene Selection Using Mutual Information and Genetic Algorithm for Cancer Data Classification," *Journal of medical systems,* vol. **43**, p. **235**, **2019**.

[34] A. K. Shukla, P. Singh, and M. Vardhan, "A hybrid gene selection method for microarray recognition," *Biocybernetics and Biomedical Engineering,* vol. **38**, pp. **975-991**, **2018**.

[35] J. Xie, M. Hao, W. Liu, and Y. Lin, "Fused variable screening for massive imbalanced data," *Computational Statistics & Data Analysis,* vol. **141**, pp. **94-108**, **2020**.

[36] A. K. Shukla, P. Singh, and M. Vardhan, "A two-stage gene selection method for biomarker discovery from microarray data for cancer classification," *Chemometrics and Intelligent Laboratory Systems,* vol. **183**, pp. **47-58**, **2018**.

[37] P. Lopez-Garcia, A. D. Masegosa, E. Osaba, E. Onieva, and A. Perallos, "Ensemble classification for imbalanced data based on feature space partitioning and hybrid metaheuristics," *Applied Intelligence,* vol. **49**, pp. **2807-2822**, **2019**.

[38] Y. He, J. Zhou, Y. Lin, and T. Zhu, "A class imbalance-aware Relief algorithm for the classification of tumors using microarray gene expression data," *Computational Biology and Chemistry,* vol. **80**, pp. **121-127**, **2019**.

[39] M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, S. Begum, and R. Sarkar, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods," *Medical & biological engineering & computing,* vol. **57**, pp. **159-176**, **2019**.

[40] S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, V. Bolón-Canedo, J. M. Benítez, F. Herrera, *et al.*, "Fast-mRMR: Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data," *International Journal of Intelligent Systems,* vol. **32**, pp. **134-152**, **2017**.

[41] S. H. Bouazza, K. Auhmani, A. Zeroual, and N. Hamdi, "Selecting significant marker genes from microarray data by filter approach for cancer diagnosis," *Procedia Computer Science,* vol. **127**, pp. **300-309**, **2018**.

[42] W. Andaru, I. Syarif, and A. R. Barakbah, "Feature selection software development using artificial bee colony on dna microarray data," in *2017 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)*, **2017**, pp. **6-11**.

[43] M. Allam and M. Nandhini, "Optimal feature selection using binary teaching learning based optimization algorithm," *Journal of King Saud University-Computer and Information Sciences,* **2018**.

[44] S. K. Baliarsingh, S. Vipsita, and B. Dash, "A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm," *Neural Computing and Applications,* pp. **1-18**, **2019**.

[45] M. A. Basir, M. S. Hussin, and Y. Yusof, "Ideal Combination Feature Selection Model for Classification Problem based on Bio-Inspired Approach," in *Computational Science and Technology*, ed: Springer, **2020**, pp. **585-593**.

[46] L. Zhang, W. Zhou, B. Wang, Z. Zhang, and F. Li, "Applying 1-norm SVM with squared loss to gene selection for cancer classification," *Applied Intelligence,* vol. **48**, pp. **1878-1890**, **2018**.

[47] X. Huang, L. Zhang, B. Wang, F. Li, and Z. Zhang, "Feature clustering based support vector machine recursive feature elimination for gene selection," *Applied Intelligence,* vol. **48**, pp. **594-607**, **2018**.

[48] M. S. M. Prince, A. Hasan, and F. M. Shah, "An Efficient Ensemble Method for Cancer Detection," in *2019 1st*

International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), **2019**, pp. **1-6**.

[49] T. Gangavarapu and N. Patil, "A novel filter–wrapper hybrid greedy ensemble approach optimized using the genetic algorithm to reduce the dimensionality of high-dimensional biomedical datasets," *Applied Soft Computing*, vol. **81**, p. **105538**, **2019**.

[50] T. Raguntharand S. Selvakumar, "A wrapper based feature selection in bone marrow plasma cell gene expression data," *Cluster Computing*, vol. **22**, pp. **13785-13796**, **2019**.

[51] P. Singh, A. Shukla, and M. Vardhan, "Hybrid approach for gene selection and classification using filter and genetic algorithm," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, **2017**, pp. **832-837**.

[52] K. Passi, A. Nour, and C. K. Jain, "Markov blanket: Efficient strategy for feature subset selection method for high dimensional microarray cancer datasets," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, **2017**, pp. **1864-1871**.

[53] D. Utami and Z. Rustam, "Gene selection in cancer classification using hybrid method based on Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) feature selection and support vector machine," in *AIP Conference Proceedings*, **2019**, p. **020047**.

[54] T. Umamaheswari and P. Sumathi, "Enhanced firefly algorithm (EFA) based gene selection and adaptive neuro neutrosophic inference system (ANNIS) prediction model for detection of circulating tumor cells (CTCs) in breast cancer analysis," *Cluster Computing*, vol. **22**, pp. **14035-14047**, **2019**.

[55] S. Singla, P. Ghosh, and U. Kumari, "Breast Cancer Detection using Genetic Algorithm with Correlation based Feature Selection: Experiment on Different Datasets," *International Journal of Computer Sciences and Engineering*, vol. **7**, pp. **406-410**, **2019**.

[56] Y. Prasad, K. Biswas, and M. Hanmandlu, "A recursive PSO scheme for gene selection in microarray data," *Applied Soft Computing*, vol. **71**, pp. **213-225**, **2018**.

[57] V. J. M. Praveena, "Particle Swarm Optimization based Feature Selection with Evolutionary Outlay-Aware Deep Belief Network Classifier (PSO-EOA-DBNC) for High Dimensional Datasets," *International Journal of Computer Sciences and Engineering*, vol. **7**, pp. **61-69**, **2019**.

AUTHORS PROFILE

Khadija Uthman, she held a B.S degree in Computer Science, Sana'a University and has finished a preliminary master degree from Faculty of Computer Science and information technology, Sana'a University in 2018, she is interested in data mining, data Science and data analysing .She recently works at statistical organization as a trainer and data cleaner .

Prof. Fadl M.M Ba-alwi, Professor in Artificial Intelligence (AI), presently working as a Vice President of the Council for Accreditation & Quality Assurance- Ministry of higher Education & Scientific Research, Yemen. In addition, he is working as a professor in Faculty of Computer and Information Technology at Sana'a University. He held a Ph.D. in Artificial Intelligence (AI) Data Mining field Computer Science –JNU- New Delhi- India. He completed his master's degree in computer Application, Jawaharlal Nehru University, also has master degree in Technology (M-Tech).

Suad.M. Othman is a lecturer in the faculty of Computer and Information Technology at Sana'a University .She held a B.S in Information System from the faculty of C&IT Sana'a University Republic of Yemen .She has held her master degree from the faculty of C&IT in intrusion detection System. In addition she published some researches related to big data and data science.