Received 22 August 2023, accepted 13 October 2023, date of publication 17 October 2023, date of current version 23 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3325381

RESEARCH ARTICLE

A Novel Fuzzy Classifier Model for Cancer Classification Using Gene Expression Data

MAHMOOD KHALSAN[1,2], MU MU[1], (Member, IEEE), EMAN SALIH AL-SHAMERY[2], SURAJ AJIT[1], LEE R. MACHADO[3], AND MICHAEL OPOKU AGYEMAN[1], (Senior Member, IEEE)

[1]Advanced Technology Research Group, Faculty of Arts, Science and Technology, University of Northampton, NN1 5PH Northampton, U.K.
[2]Computer Science Department, College of Information Technology, University of Babylon, Babylon 51002, Iraq
[3]Centre for Physical Activity and Life Science, Faculty of Arts, Science and Technology, University of Northampton, NN1 5PH Northampton, U.K.

Corresponding author: Mahmood Khalsan (mahmood.khalsan@northampton.ac.uk)

The work of Mahmood Khalsan was supported in part by the Ministry of Higher Education and Scientific Research in the Republic of Iraq, in 2017, under Grant 22223.

ABSTRACT In the pursuit of better cancer classification, many studies have been conducted to identify the genes associated with cancer. However, the high dimensionality of gene expression data and the limited relevance of a few genes pose significant challenges to this endeavor. Existing gene selection methods yield divergent gene lists, further complicating the classification process. To overcome this issue, we developed a novel approach called Fuzzy Gene Selection (FGS), which combines the strengths of various gene selection methods in the field. FGS was developed using three feature selection techniques (Mutual Information, F-ClassIf, and Chi-squared) to rank genes based on their importance. These methods generated scores and rankings for each gene. Fuzzification and Defuzzification techniques were then applied to combine these scores into a single best score for each gene. This approach aids in identifying genes of significance in cancer classification, especially in multi-class scenarios. Classifiers often produce convergent decisions in such cases, where the predicted probabilities for different classes do not always correspond to the correct predicted class with the highest probability. To address this, we developed a novel Fuzzy classifier that leverages the contributions from each node's traditional deep classifier. This novel approach combines the strengths of traditional deep classifiers at individual nodes, enabling the Fuzzy classifier to make more robust and informed predictions. The Fuzzy classifier (FC) has demonstrated improvements in accuracy and the generalization of the proposed algorithm to accurately classify different cancer types.

INDEX TERMS Classifier methods, fuzzy gene selection technique, fuzzy classifier method.

I. INTRODUCTION

Cancer is a group of cells that arise from specific areas of the human body, that mostly, speedily spread to distant metastatic sites [1]. Cancer is the abnormal growth of cells by dividing cells uncontrolled. The major difference between cancerous cells and non-cancerous cells is that cancerous cells, keep growing even though there are no signals saying that, and neglect the signals that tell them to stop dividing or die (this process is called apoptosis). By contrast, non-cancerous cells, grow only when they received a signal for growth, and they stop dividing cells when they received signals inform that. The danger area of cancer is that it is speedily metastatic to other parts of the human body which make it too hard to control. One of the best solutions to treat cancer is by removing the part that is affected by a malignant tumor. Therefore, researchers are keen in their studies to provide early detection technology, which contributes to reducing the risk of its metastasizing to other parts that are difficult to control.

Although, the best technique to cure cancer is to detach the tumor from the portion that contains the tumor. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Marco Giannelli.

surgical therapy is not always accessible since, in certain sensitive areas, surgical removal might severely harm surrounding tissue, such as the spinal cord. Some tumours, such as those found in certain brain malignancies, can produce tiny tendrils that wrap through surrounding tissue and are hard to remove surgically without injuring the patient. This has prompted the researchers to conduct extensive studies on this sort of sickness since it is extremely harmful and there is currently no viable treatment for cancer. Early cancer identification aids in cancer treatment by eliminating the cancerous tissue. Based on that, researchers sought to analyze various forms of patient data in the hopes of discovering a method that aids in the detection of cancer at an early stage, allowing for the removal of the diseased section and preventing metastasis to other regions of the body. This endeavor began with CT scans and MRI pictures, which are used to detect whether or not a person has cancer, the location of the tumor, and the size of the tumor, as well as whether or not cancer has spread to other regions of the body. Nonetheless, these strategies produced positive outcomes in this discipline. However, they are expensive and have a detrimental influence on patients, particularly youngsters, because employing these procedures necessitates the use of a dosage of radiation, which may lead to the development of cancer in the future [2].

For these reasons, researchers are turning to alternate technologies that are more successful for early diagnosis, less costly, and have a lower risk of being affected by cancer or another disease in the future.By employing various measurement techniques, researchers have developed distinct methods to determine the levels of gene activity expressed in both malignant and non-cancerous cells. (Microarray and RNA-seq tools) are common measurement methods that have been developed to measure the expressed level of thousands of genes across hundreds/thousands of samples [3]. These approaches hold an advantage over previous methods such as CT scans and MRI, as they facilitate not only the early detection of cancer but also personalized therapy.

These datasets (Microarray and RNA-seq) contain noise, missing data, and duplicate data, which need the use of various approaches to address all of these issues in the gene expression data. As a result, most researchers analyze gene expression using strong techniques such as feature selection (FS) and machine learning (ML) approaches. The significant limitations of prior research are that gene expression is characterized by high dimensionality, which necessitates the use of a powerful gene selection technique to reduce dimensionality [21]. By choosing a modest number of genes to serve as identifiers for the training of the classifier algorithm. Furthermore, the accuracy obtained was inadequate for some datasets when using traditional classifiers. As a result, a new classifier capable of reliably classifying cancer is required. Another significant issue highlighted in previous studies is the challenge of developing a classifier that can effectively generalize and classify cancer across diverse datasets.

Our research makes important contributions described as follows.

1) Develops a novel fuzzy gene selection FGS method to select a significant subset of genes.
2) Improves the classifier performance by reducing classifier complexity, time spends on training, and overfitting issues.
3) Develops a novel fuzzy classifier method to enhance cancer classification accuracy.
4) Improves the results and increases the generalization of FC to continuously achieve the highest results rather than each classifier achieving the best results for certain datasets.

II. RELATED WORK

Matsubara et al. [4] classified lung cancer utilizing both protein interaction network data and gene expression data from 638 samples using CNN (combining spectral clustering information processing). The dataset may be obtained from NCBI GEO datasets (ID GSE66499). This study attained accuracy, recall, precision, and specificity of 81%, 88%, 78%, and 74%, respectively. This research, like the others listed, did not use validation data to test the model's efficiency. Furthermore, 190 of 487 cancer samples were picked at random, which explains the results obtained. (190) Cancerous samples were chosen at random, which would be ineffective. The most obvious drawback of this study is that the accuracy gained was not at the level of cancer disease sensitivity. Yuan et al. [5] used RF and SVM to classify two subtypes of lung cancer: (Adenocarcinomas (AC) and Squamous Cell Carcinoma (SCC)). They also applied Monte Carlo (MCSF) and incremental feature selection (IFS) methods to identify informative genes. The dataset was obtained from GEO (ID GSE43580). The study showed that when 1100 optimal features (genes) were selected for classification using an SVM classifier, higher accuracy was achieved compared with using 43 informative features (genes) obtained using an MCSF method. Accuracy decreased from 96% to 86% using SVM and 93% to 88% with RF.

From a classical DNN, a differential regulation network embedded deep neural network (DRE-DNN) strategy was designed and used to predict liver cancer (hepatocellular carcinoma) outcomes using three datasets (GEO GSE10143, GSE14520, and TCGA) [6]. The proposed model produced average AUC values of 86%, 74%, and 72% for the GSE10143, GSE14520, and TCGA datasets, respectively, which enhanced on standard DNN AUC values. The study validated and measured the performance of the suggested strategy using various data sources. It employed a sufficient dataset to train the DRE-DNN model, and while it was beneficial as a prediction tool, it did not achieve satisfactory classification results. It does, however, help to alleviate the model's overfitting issue.

ReliefF, a feature selection approach, and Random Forest (RF), a classification method for lung cancer, have both been proposed by Li et al. [7]. The datasets were obtained from TCGA. The research was compared to other classification algorithms like SVM and Nave Bayes. The suggested study employed ReliefF to pick 67 genes, which were then used to train the RF model. The RF classifier outperforms conventional classification algorithms. When ReliefF and RF are used together, the accuracy attained in this study is 83.6%. The study's major flaw is its low accuracy, which is inappropriate given the sensitivity of the cancer topic.

Xu et al. [8], utilized the maximum relevance minimum redundancy feature selection technique to select a small number of informative genes for training the k-Nearest Neighbors algorithm. The study was used to classify thyroid carcinoma. The dataset was retrieved from (GEO GSE33630) and contains 105 samples with 54,675 probes corresponding to 20,283 protein-coding genes. The study obtained 85.7% accuracy with the top ten genes. Even though the inquiry reduced the number of genes, the accuracy achieved was not at the level of the cancer sensitivity topic.

Hilal et al. [9], suggested a novel feature subset selection with an optimum adaptive neuro-fuzzy inference system (FSS-OANFIS) for cancer classification. The colon cancer dataset produced the best results, with 89.47%, 87.80%, 87.82%, and 87.82% for accuracy, sensitivity, specificity, and G-measure, respectively. When FSS-OANFIS was applied to the prostate dataset, the accuracy, sensitivity, specificity, and G-measure were 73.3%, 66.67%, 66.67%, and 70.47%, respectively. The study was assessed for various gene expression datasets; however, the findings were unsatisfactory when compared to the cancer topic's sensitivity. Another issue was identified when just the microarray gene expression dataset was used. As a result, the accomplishment accuracy and the number of selected genes require additional effort by developing new methodologies capable of properly identifying cancer with a small number of useful genes.

Rostami et al. [10], a novel social network analysis-based gene selection method was developed for selecting a limited number of informative genes that would be used for training different classifier algorithms. This method was developed by integrating node centrality and community detection concepts. The study was examined by using five microarray datasets and four classifier techniques. The highest average accuracy for the five datasets was 87.7% when the Extreme Learning Machine classifier was used. Although, the proposed model achieved better accuracy when compared to other gene selection methods. However, the accomplished results were not high compared to other studies in this field. Another disadvantage, the system was tested only using microarray datasets. It also used a multi-phase approach that is computationally expensive in high-dimensional datasets.

Hamraz et al. [23], employed the robust Fisher score method to effectively select a subset of genes tailored for binary classification. The investigation encompassed a comparative analysis of its efficacy across five distinct gene expression datasets, juxtaposed against six established feature selection techniques. These methodologies were rigorously evaluated utilizing three traditional classifiers SVM, kNN, and RF). The findings of this research were conspicuously compelling, with the Robust Fisher Score method consistently outperforming its counterparts across the majority of the utilized datasets. The training and testing protocol involved a 70% allocation for training and the remaining 30% for testing. However, it is noteworthy that while the results are promising, certain limitations merit consideration. Primarily, the scope of this study is confined exclusively to binary classification tasks. Furthermore, the adoption of a simple data split for training and testing may not yield a comprehensive performance assessment of the classifiers. In this regard, employing cross-validation would offer a more realistic evaluation of the model's capabilities. Another facet to contemplate is the study's reliance on relatively small datasets. Extending the analysis to encompass larger datasets might unveil variances in performance outcomes.

Bashir et al. [24] introduced a novel approach to feature selection aimed at identifying a subset of informative genes. They coupled this approach with an SVM-based evaluation to gauge its efficacy. The proposed method combine the top feature selection methods of three distinct methodologies: filter, wrapper, and embedded methods. This composite approach is followed by an intersection step, yielding a cohesive list of genes earmarked for training the SVM model. The outcomes of their study unveiled noteworthy results: achieving accuracy levels of 94%, 78.25% for sensitivity, 83.56% for specificity, and 80.9% for the F-measure. While the accuracy rates are commendable, it's worth noting that other evaluation metrics, such as sensitivity, exhibited comparatively lower values. This implies that although their method excelled in accurately predicting majority cases, there's room for enhancement in capturing the more nuanced instances, as indicated by the sensitivity and other related metrics.

In our previously published work [25], we delved into a broader range of studies that analyze gene expression datasets for cancer classification. These studies contribute to the rich landscape of research in the field.

III. DEVELOPED METHODOLOGY

A. FUZZY GENE SELECTION (FGS)

FGS aims to achieve a harmonious balance between a minimal number of genes and maximum accuracy metrics such as accuracy, precision, recall, and F1-score. FGS able to identify a subset of informative genes, which leads to classifier simplicity, and training time as well as improves accuracy, and mitigates overfitting. The FGS method assumes employing three filter feature selection methods (Mutual Information, F-ClassIf, and Chi-squared) that were evaluated and employed to obtain the score and rank for each gene. Therefore, three major steps will be done as follows.

1) VOTING PHASE

By employing the three-feature selection method each feature selection technique presents a different list of genes based on the Step Function (SF). Step function computed in the formula1 is intended to prevent a restricted number of selected genes, which may result in the omission of certain genes with the same score when employing a constant number of features, such as the top ten features. It also allows greater flexibility to the SF value when compared to constant values such as 0.3. If non- or minimally chosen features by a feature selection technique have scored equally to 0.3, we miss some essential features (genes) that may have been selected by other feature selection approaches.

$$SF = max(FSS) * 0.3 \tag{1}$$

where FSS is the feature selection method's score for all genes. While max is the highest possible score for all genes assessed by each feature selection technique. This step produces a list of genes with scores that are either equal to or larger than the previously calculated SF value. It usually produces a thousand or more genes. This is the initial filter procedure, which tries to reduce the number of chosen genes that will be utilized for further filtering in the subsequent phases.

2) FUZZIFICATION PHASE

This is the process of changing crisp data into fuzzy data utilizing membership functions, with the goal of transforming the crisp data into data spanning between [0-1]. There are several kinds of membership functions. The Triangular Membership Function was used in this study. In short, Triangular Membership Function has been used to unify the score for each gene for the three-feature selection methods between [0-1]. The equation below has been used for this purpose.

$$Mf = \frac{\mathrm{W_i} - a}{b - a} \tag{2}$$

where MF is the membership function. W is the gene's crisp value (score), a = lowest possible score (min), b = the highest possible score. This membership function was used for the three feature selection techniques, which are MF1, MF2, and MF3 in this study.

3) DEFUZZIFICATION PHASE

This stage involves the conversion of the output data to crisp data. This is the last process of the gene selection procedure used to choose important genes. These phases' identified genes were utilized as identifiers for training the classifier algorithms. It is calculated in the formal below.

$$ASG = \frac{\mathrm{MF_i} + \mathrm{MF_i} + \mathrm{MF_i}}{N} \tag{3}$$

where ASG is the Gene's Average Score using the three feature selection techniques. Each gene's membership function is denoted by MF. N is the total number of feature selection techniques used. In this piece, N is three. From the two processes above, it can be inferred that fuzzification and defuzzification have been used to achieve the goal of having a single score for each gene while filter feature selection approaches offer diverse scores for the same gene. As a result, employing a SF for deciding which genes are the optimal subset to utilize as markers for cancer classification, as illustrated in the equation below.

$$SF = max(FSS) * 0.5 \tag{4}$$

where FSS is the feature selection method's score for all genes. While max is the highest possible score for all genes assessed by each feature selection technique.

B. FUZZY CLASSIFIER METHOD (FC)

The major goal of the proposed FC is to improve cancer classification accuracy and raise an algorithm's generalization to be accurate with all provided datasets. More crucially, the fuzzy classifier approach enables the highest attained accuracy for many classifier methods for each dataset. FC posits that after applying three classifier approaches (Logistic regression (LR), SVM, and MLP) to a dataset, the probability of predicting a class label for each classifier is calculated. Therefore, using the three classifier techniques, find the maximum probability of the average for each class label. Consequently, the class label with the highest max of the average of the three classifier techniques is chosen as the predicted class that would be compared to the real class label, a process known as soft. Another technique, known as the majority, works as follows: if two classifier methods predict the class label as A and only one predicts the class label as B, the outcome will be class A because two out of three classifiers predicted A. FC proposed relying on these two processes in order to benefit from both. FC predicts class labels by combining soft and majority approaches. For example, if the predicted class label when using the soft method is A and the predicted class label when using the majority method is B, FC applied a member function that takes into account both methods (soft and majority) by adding 0.6 to the max average of the class label selected by the majority method and dividing by two. The result of this procedure is then compared to the max average; if the output is bigger, the max average is chosen as the predicted class; otherwise, the predicted class is the same predicted class by the max average (soft) method.

The objective function of this method is to achieve maximum accuracy and develop a classifier capable of consistently achieving high performance across different datasets, including normal vs. cancer, various cancer types, and cancer subtypes.

Support vector machine works by determining the optimum decision boundary (Hyperplane) to divide the input data into distinct areas. The SVM method seeks the hyperplane in an n-dimensional space that separates various data points [13].

Multilayer Perceptron translates the input to the output in a single data and computation direction. In general, it is composed of three perceptron or layers: an input layer, an output layer, and at least one in between known as a hidden layer [14]. In MLP, each layer is completely linked to the following layer. The input layer receives signals from the outside world and sends them to the network, hidden layers execute mathematical operations from the input layer to the output layer, and the output layer makes the judgment.

Logistic regression (LR) is a statistical approach for dealing with both classification and regression issues. It is classified as supervised learning. LR is based on the probability idea, which is determined using the Sigmoid function [15]. The process of proposing a novel fuzzy gene selection-fuzzy classifier method is described in Figure1.

C. MODEL SETTING

To achieve the best performance in the FC method, it is essential to determine the optimal settings for three different classifiers. Based on experimentation and analysis that used different configurations. These configurations such as MLP (1,2,3 hidden layer), 'lbfgs', 'sgd', and 'adam', as a solver. While with SVM kernel = 'poly', degree = 4, and kernel = 'linear' with c = 1,2) and LR using iterations 100,500,1000. Consequently, The following optimal settings have been identified:

1) MLP: An input layer, three hidden layers, and one output layer. Activation function: ReLU, Solver: Adam, and Maximum iterations: 200.
2) SVM: Kernel: Linear, Regularization parameter (C): 1, and Probability estimation enabled
3) LR: Multi-class approach: One-vs-Rest (ove), Maximum iterations: 1000, and Random state (for reproducibility)

IV. PROPOSED MODEL TOPOLOGY

The proposed model consists of a total of 20 layers, which includes one input layer, 18 hidden layers, and one output layer. 8 layers for the development FGS method, and 10 layers for the development of the FC classifier. In the FGS method, three vertical layers are employed for gene selection methods, and two additional vertical layers are dedicated to a voting stage to avoid time consumption. For the FC classifier, two vertical layers are utilized for classifiers, and two vertical layers are devoted to fuzzification. In summary, the proposed model consists of nine vertical layers and nine sequential layers, in addition to one input layer and one output layer. The number of neurons in each dataset varies based on the number of genes and samples in the dataset, ensuring flexibility and adaptability across different data types and sizes. The new topology is described in Figure2.

V. EXPERIMENTAL SETUP

The developed model (FGS-FC) is implemented using Python software with the Intel core i7-8565U processor and 32 GB RAM. A comprehensive evaluation was conducted using different cancer types to evaluate the effectiveness of the developed model. Specifically, thirteen datasets were employed, including nine microarray and four RNA-seq datasets. This wide range of datasets allowed for a thorough investigation into the performance and applicability of the developed model. The datasets were split into testing and training using a cross-valuation method for a better understanding of how well the model generalizes and reliable results are achieved. The Experimental setup is described in Figure3.

A. EMPLOYED DATASETS

Many popular sources offer cancer gene expression datasets containing both (Microarray and RNA-seq) data. Only the Gene Expression Omnibus (GEO) and the Cancer Genome Atlas (TCGA) were used in this study since they were widely used by other researchers. Both repositories accept Microarray and RNA-seq data, whereas GEO mostly focuses on Microarray data. GEO provides a total of 3,635,328 samples for various diseases [11], whereas TCGA provides 84,031 samples for 33 distinct cancer types [12]. Fourteen datasets consisting of nine microarray datasets, and five RNA-seq datasets were employed to train and evaluate the developed model. Table1 has a complete description of each dataset used, including the number of samples, genes, and classifiers for each dataset.

B. A CROSS-VALIDATION (CV)

CV is a statistical technique used in machine learning (ML) called cross-validation that seeks to reduce or completely eliminate overfitting problems in various classifier paradigms. With the use of the k cross-validation approach, a model may be trained on several training datasets as opposed to only one. By folding the dataset into k-folds and training the model on each fold. The model is able to generalize as a consequence, which is an indication of a robust model. It also helps to show a better indication of the performance of the algorithmic prediction. As illustrated in Figure4, the datasets are divided into k-folds, like k = 5.

C. EVALUATION PERFORMANCE

In general, four techniques are utilized to assess the performance of any classifier approach. These evaluation performances aim to determine how well a classifier is doing. These evaluation parameters began as follows:

Accuracy (AC) is an assessment metric used to identify which classifier is best for a certain dataset. In ML, AC is defined as the ratio of successfully predicted observations to total observations. It is computed mathematically as follows [16].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

where TP denotes True Positive, TN is True Negative, FP denotes False Positive, and FN denotes False Negative. A TP is a result that the model correctly predicts as a positive

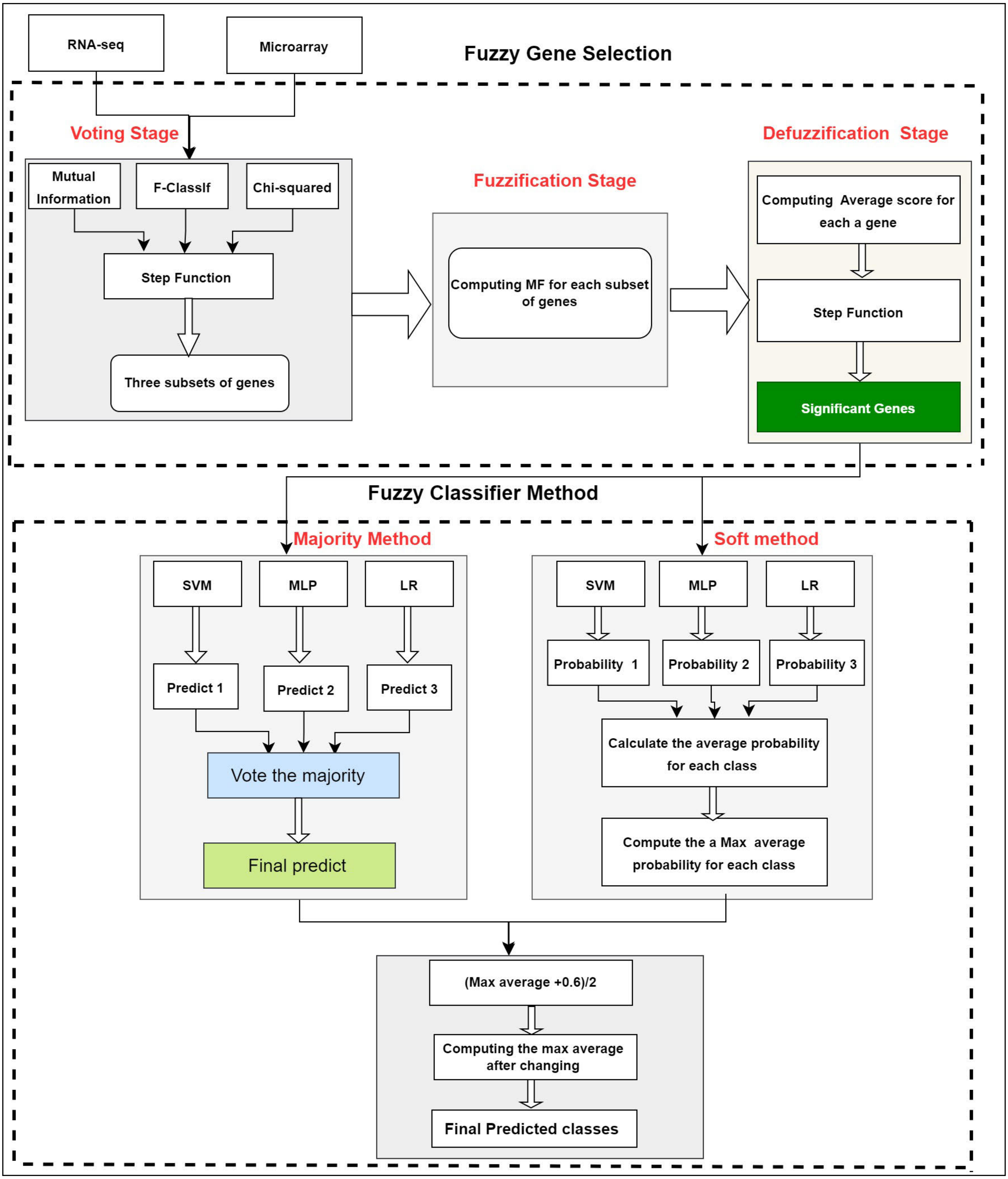

FIGURE 1. Flow Diagram of Proposing FGS-FC Model.

class. TN is an outcome in which the model accurately predicts a case as a negative class. Non-cancerous instances, for example, are classified and rightly classified as such by the model. FP is the incorrectly predicted positive class

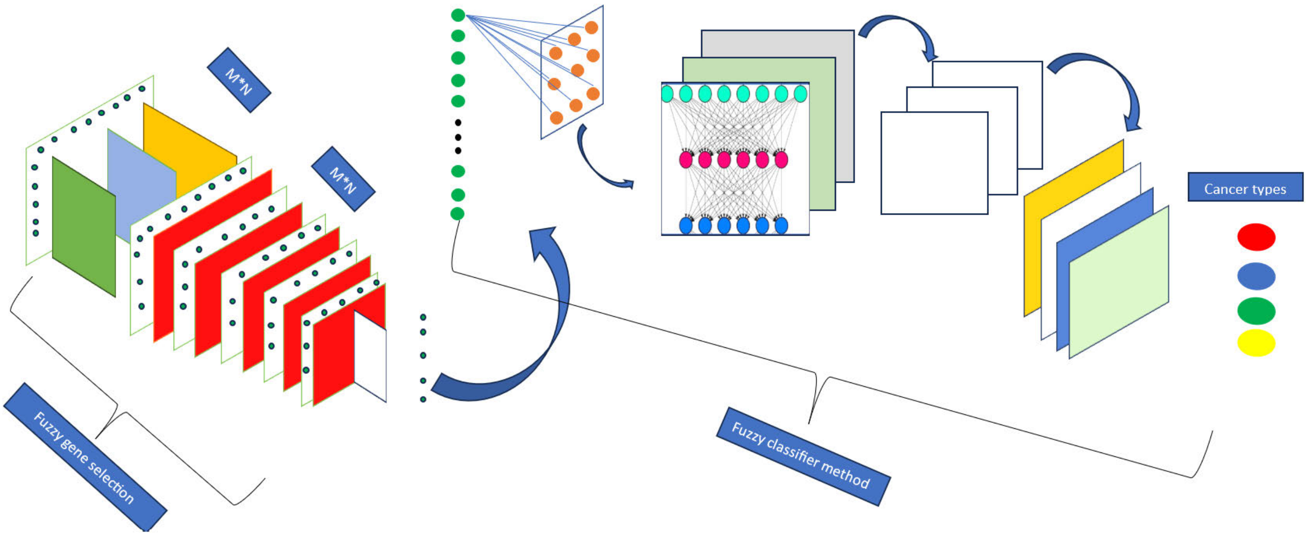

FIGURE 2. Topology of the Proposed Model.

outcome. FN indicates that the model identified the negative class incorrectly.

Precision is defined as the percentage of successfully predicted positive findings to total predicted positive observations is explained in [16]. Mathematically, it is illustrated in the formal below.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall is defined as the percentage of retrieved instances out of all relevant instances. It is sometimes referred to as sensitivity. The recall equation is shown as [16].

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

F1-score is defined as the weighted average of precision and recall, with a perfect F1 score of 1 and the poorest score of 0 [16]. In short, it's combined the precision and recall of a classifier algorithm into a single metric. Mathematically, it is described as follows:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{8}$$

VI. EXPERIMENT RESULTS

A. ACHIEVED OUTCOMES

This part compares the usage of a fuzzy classifier method and demonstrates the efficiency of using the developed FC approach using the thirteen datasets across five classifier algorithms. In this section, we evaluate the influence on cancer classification performance. The full details are presented in Table2, which includes the datasets used for training and testing the models, the gene number, the number of samples, and the achieved accuracy, precision, recall, and f1-score. The tables below compare the results obtained using traditional machine learning and the proposed Fuzzy classifier method. The findings demonstrate how well developing FC improves the accuracy across all datasets used. In summary, these tables provide a comparison of using FC and five traditional classifiers with different datasets when FGS is used. Furthermore, they support our prior findings that employing FGS improves accuracy when compared to not using FGS.

VII. DISCUSSION

The study's goal is to show how successful the suggested fuzzy classifier approach is when compared to other classifiers with and without FGS used. Although FGS has demonstrated its effectiveness in reducing the number of gene and enhancing the performance of traditional classifier algorithms across diverse datasets. However, it's essential to acknowledge that outcomes for certain datasets were substandard. Furthermore, no single classifier consistently achieved the highest accuracy for different datasets. This led to the proposal of the FC approach for dealing with these datasets. It also outperformed, even though some datasets got decent results when FGS was used with traditional classifier algorithms as illustrated in Table2. As a result, this section concentrates on the comparison of classic classifier techniques with FC when FGS is applied.

A comparison of five classifier algorithms and FC for classifying Gastric cancer using the FGS strategy in 5 kfold cross-validations is shown in the flowchart below (Figure 5). This dataset (GSE84437) is regarded as a challenging dataset because the accuracy obtained by the five applied classifiers, which ranged from 24% to 35.3%, was extremely low. Although using the FGS technique reduced the number of selected genes and somewhat improved accuracy in the classification of malignancies, the level of improvement was not proportionate with the sensitivity of the cancer topic where the accuracy was ranging between 35.3% and 37.89%. In order to classify cancer accurately given this

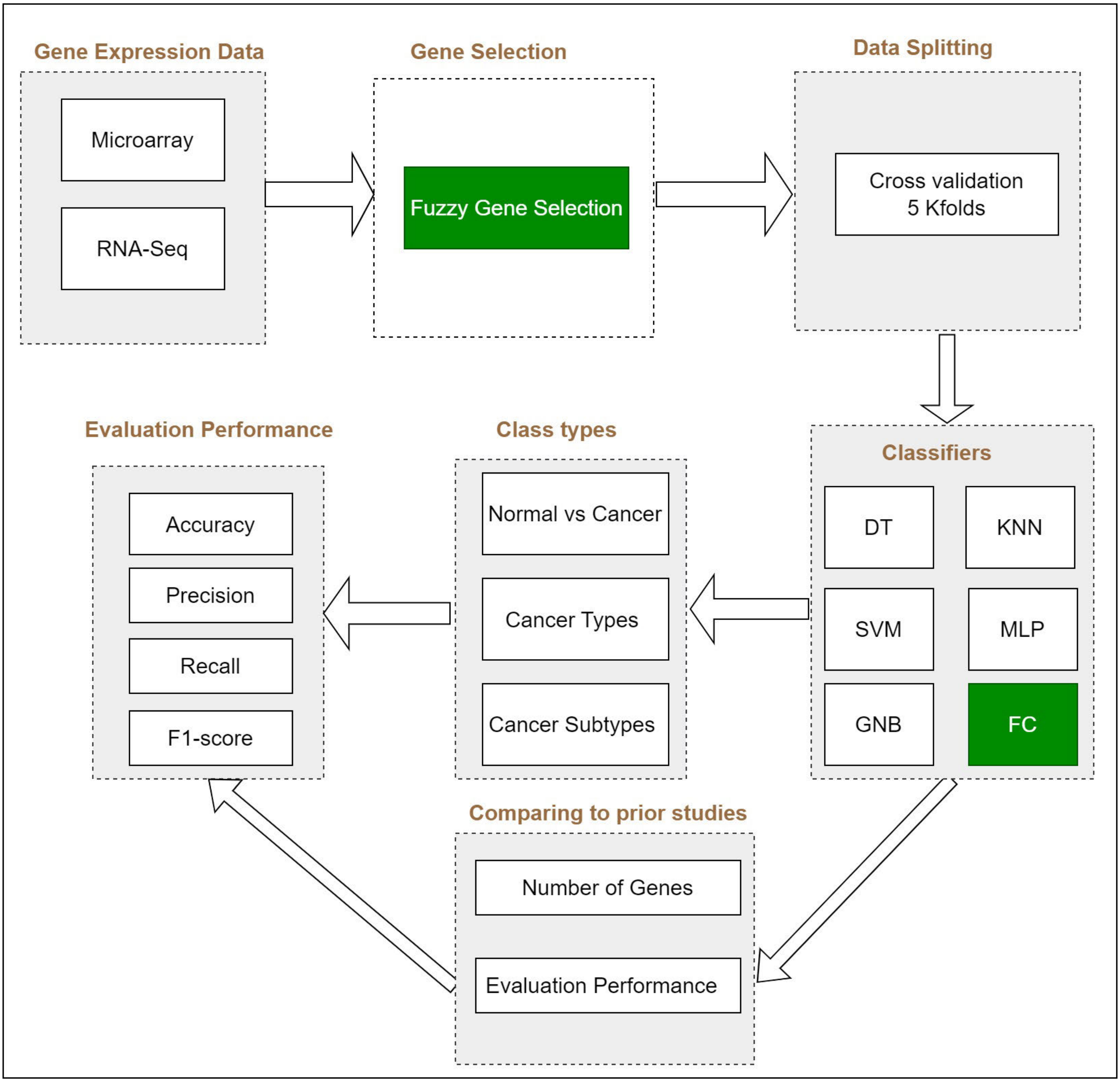

FIGURE 3. Experimental setup Process.

type of dataset, it becomes vital to create a fuzzy classifier algorithm. The significant level of enhancement in this set of data, which achieves 92.8% accuracy with only a tiny subset of selected genes chosen using the FGS method, made it clear how efficient the FC technique was. This leads to the conclusion that feature selection methods do not always improve classifier model performance, sometimes necessitating classifier improvement to make it fit particular datasets. This led us to create this method to enable the proposed algorithm to handle the majority of cases such that, even in the worst scenarios, it is still possible to acquire an accuracy that is proportionate to the volume of the sensitivity of the cancer topic.

The comparison of five well-known classifier techniques and the proposed model (FC) that attempt to classify lung cancer gene expression dataset in 5 kfolds cross-validation is shown in Figure6. The findings indicate that all the classifier models used had poor accuracy. Because no classifier approach has been able to effectively identify data utilising FGS and devoid using FGS, this dataset (GSE66499) is also regarded as a challenging dataset. The accuracy of the preserved data ranged from 64.7% to 73.8%. As a result,

TABLE 1. Detailed description of the datasets used to train and evaluate the proposed model.

Datasets ID	Measured	Samples Number	N-Genes	Cancer types	N-Classes	Reference
GSE45827	Microarray	155 (Basal 41,Her2 30, Luminal B 30, Luminal A 29, CellLine 14, Normal 11)	29873	Breast cancer subtypes	6	[17]
GSE66499	Microarray	680 (Cancer 490 , Normal 190))	33298	Lung cancer	2	[4]
GSE10072	Microarray	107 (Adenocarcinoma 58, Normal 49)	13298	Lung cancer	2	[17]
GSE43580	Microarray	150 (Adenocarcinomas 77, Squamous Cell Carcinomas 73)	54675	Lung Cancer	2	[5]
GSE14520	Microarray	445 (Cancer 227, Normal 218)	13425	Liver Cancer	2	[6]
GSE84437	Microarray	433 (T1 188,T2 132 T3 80,T4 33)	48710	Gastric Cancer	4	[17]
GSE53757	Microarray	144 (Cancer 72, Normal 72)	23516	Kidney Cancer	2	[17]
GSE77314	RNA-seq	100 (Cancer 50, Normal 50)	29087	Liver Cancer	2	[17]
GSE19804	Microarray	120 (Cancer 60, Normal 60)	45782	Lung Cancer	2	[17]
TCGA1	RNA-seq	2086 (BRCA 878, KIRC 537, UCEC 269, LUSC 240, LUAD 162)	972	Five cancer types	5	[12]
GSE33630	Microarray	105 (PTC 49, Normal 45, ATC 11)	23518	Thyroid	3	[8]
TCGA2	RNA-seq	964 (LumA 431,LumB 195, Basal 143, Normal 128, Her2 67)	20531	Breast cancer subtypes	5	[12]
TCGA3	RNA-seq	197 (Cancer 94, Normal 103)	768	Liver cancer	2	[6]
TCGA4	RNA-seq	592 (Cancer 533, 59)	56603	Lung cancer	2	[7]

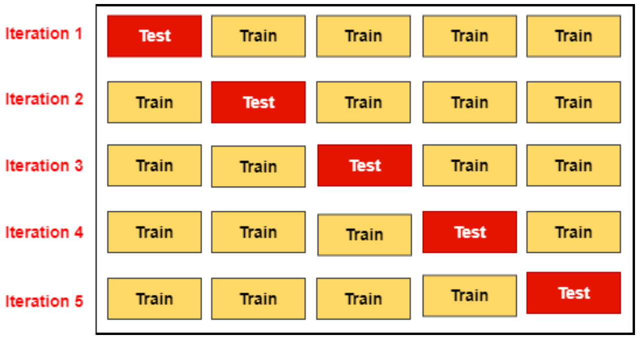

FIGURE 4. K-Fold Cross Validation Process with K = 5.

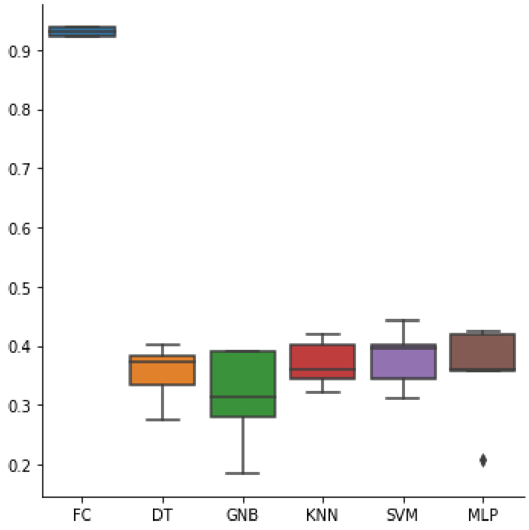

FIGURE 5. Accuracy score of gastric cancer (GSE84437) for five classifier and FC models in 5 kfolds.

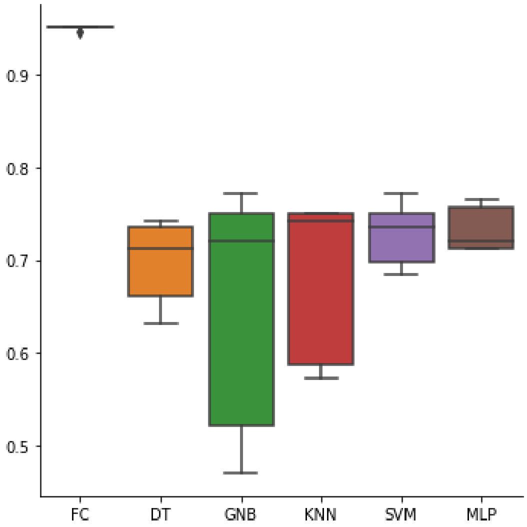

FIGURE 6. Accuracy scores of lung cancer (GSE66499) for the five classifiers and FC in 5 kfolds.

it was crucial to build a fuzzy classifier approach to address this problem. The FC produced promising results, scoring 95%, 95%, 92%, and 93% for accuracy, precision, recall, and f1-score, respectively. These outcomes demonstrate the technique can successfully classify cancer using a limited number of genes, which requires less time during the training phase and a simpler classifier algorithm. Additionally, it demonstrated the capacity to handle some unique datasets that cannot be resolved when using conventional machine learning techniques and FGS together.

Figure7 depicts the accuracy attained in five standard classifier algorithms and the suggested model (FC) when the FGS approach was used. When FGS was used to identify important genes, two classifiers (KNN and SVM) obtained the maximum accuracy. For accuracy, precision, recall, and f1-score, the results were 94%, 96%, 92.8%, and 93%, respectively. When the FC technique was used with FGS, the accuracy, precision, recall, and f1-score were all 100%. The findings showed that the proposed model outperformed

TABLE 2. A comparison of the outcomes of traditional classifier techniques and the FGS-FC.

Datasets	Gene Number	Gene Selection	Class Number	Classifiers	Accuracy %	Precision%	Recall %	F1-score %
GSE45827	29873	Without Feature selection	6	DT	85.8	83	82.6	81.5
				GNB	89	92.7	88.8	89
				KNN	85	88.4	87.7	86.9
				SVM	94.8	96.3	95.8	95.8
				MLP	20.6	6	17.9	7
GSE45827	68	FGS	6	DT	89.6	90.9	89.6	88.8
				GNB	91.6	94.5	92	92.8
				KNN	96.7	97.59	97.38	97.36
				SVM	97.4	98	97.66	97.75
				MLP	98	98.8	98	98.3
GSE45827	68	FGS	6	FC	100	100	100	100
GSE33630	23518	Without Feature selection	3	DT	87.6	77.6	81	79
				GNB	90.4	93.7	89.7	90
				KNN	91.4	87.7	86.5	86.3
				SVM	93.3	95.3	92	92.4
				MLP	72.3	55.6	64.5	58.5
GSE33630	76	FGS	3	DT	93.3	93.4	93.5	92.5
				GNB	92.3	88.3	89.8	88.8
				KNN	94	96	92.8	93
				SVM	94	96	92.8	93
				MLP	92.3	88.3	89.9	88.8
GSE33630	76	FGS	3	FC	100	100	100	100
GSE19804	45782	Without Feature selection	2	DT	89	89.9	88.3	88.9
				GNB	92.5	95	90	91.9
				KNN	90.8	88	95	91.3
				SVM	95.8	96.6	95	95.7
				MLP	50	20	40	26.6
GSE19804	36	FGS	2	DT	90.8	94.5	86.66	90
				GNB	95.8	96.7	95	95.7
				KNN	96.66	96.79	96.66	96.66
				SVM	96.66	96.79	96.66	96.66
				MLP	96.66	96.79	96.66	96.66
GSE19804	36	FGS	2	FC	100	100	100	100
TCGA1	972	Without Feature selection	5	DT	91	87	85.3	85.8
				GNB	94	89.7	92	90.7
				KNN	88	83.3	81.5	81.9
				SVM	93.6	91	88.9	89.8
				MLP	94	90.8	89.8	90
TCGA1	25	FGS	5	DT	91.7	88	87	86.5
				GNB	92.4	87.7	90.8	89
				KNN	93.6	89.4	90	89.6
				SVM	94	90.5	90.77	90.5
				MLP	95	92.3	91.6	91.6
TCGA1	25	FGS	5	FC	97	95	94	95
GSE77314	29087	Without Feature selection	2	DT	95	98	92	94
				GNB	84	100	68	80.3
				KNN	89	82	100	90
				SVM	99	98	100	99
				MLP	93	98	88	92
GSE77314	12	FGS	2	DT	98	98	98	98
				GNB	97	98	96	96.8
				KNN	99	98	100	99
				SVM	99	98	100	99
				MLP	99	98	100	99
GSE77314	12	FGS	2	FC	100	100	100	100
GSE14520	13425	Without Feature selection	2	DT	90	90.6	88.9	89.7
				GNB	95	95.6	94.4	94.8
				KNN	94	91	97.6	94
				SVM	97	96.4	97.6	97
				MLP	86.7	76.5	96.7	76.5
GSE14520	23	FGS	2	DT	95	95.4	94.9	95
				GNB	96.6	96	97	96.59
				KNN	96.6	96	97	96.59
				SVM	96.85	96	97.68	96.8
				MLP	95.5	95.6	95.3	95.3
GSE14520	23	FGS	2	FC	99	98	99	98
GSE53757	23516	Without Feature selection	2	DT	95.7	94.6	97	95.8
				GNB	95	95.6	94	95
				KNN	95.7	95.6	95.7	95.6
				SVM	95	94.8	96	95
				MLP	93.7	98.5	89	93.3

TABLE 2. *(Continued.)* A comparison of the outcomes of traditional classifier techniques and the FGS-FC.

Datasets	Gene Number	Gene Selection	Class Number	Classifiers	Accuracy %	Precision%	Recall %	F1-score %
GSE53757	78	FGS	2	DT	95.8	93.7	98.5	95.9
				GNB	97.8	97	98.5	97.8
				KNN	97.8	97	98.5	97.8
				SVM	97	96	98.5	97
				MLP	96.5	96	97	96.5
GSE53757	78	FGS	2	FC	100	100	100	100
GSE10072	13298	Without Feature selection	2	DT	94.3	93.5	96	94.3
				GNB	98	100	95.7	97.7
				KNN	97	95.3	100	97.3
				SVM	99	100	98	98.9
				MLP	50	18	4	25
GSE10072	52	FGS	2	DT	93.4	93.56	93.77	93
				GNB	98	100	96	97.8
				KNN	99	100	98	98.9
				SVM	98	100	96	97.89
				MLP	98	98	96	96.9
GSE10072	52	FGS	2	FC	100	100	100	100
GSE84437	48710	Without Feature selection	4	DT	33	20	23	19.7
				GNB	24.7	28.5	24.4	20.6
				KNN	35.3	19.7	23.7	19
				SVM	32.3	18.3	24	18.6
				MLP	27	10	23.4	12.6
GSE84437	105	FGS	4	DT	35.3	22.4	24.5	21
				GNB	31	34	42	29.4
				KNN	36.96	28.5	28.3	25.6
				SVM	37.89	33	29.7	26
				MLP	35.3	32.3	32.4	29.46
GSE84437	105	FGS	4	FC	92.8	95.6	81.4	85
GSE66499	33298	Without Feature selection	2	DT	66	39.5	34	36
				GNB	52.3	38	50	27.3
				KNN	67.8	42.4	17.8	19.7
				SVM	72.8	62	29	37
				MLP	68	59.3	20.5	19.3
GSE66499	150	FGS	2	DT	69.7	48	23	29.5
				GNB	64.7	53.4	56.8	44.69
				KNN	68	52	23.6	28.8
				SVM	72.79	55	37.36	43
				MLP	73.3	47.8	20.5	24.4
GSE66499	150	FGS	2	FC	95	95	92.4	93.4
TCGA2	20531	Without Feature selection	5	DT	83.7	83.4	80	81.3
				GNB	77	77.6	77.7	77
				KNN	73	76	64.5	66.3
				SVM	82.7	82	82.5	82
				MLP	84.4	84.4	83	82
TCGA2	116	FGS	5	DT	80.7	76.5	74.3	74.8
				GNB	81	76.4	80	77.56
				KNN	82.98	84	75	77.68
				SVM	86.3	87	82.6	84.3
				MLP	86.3	86	84.88	85
TCGA2	116	FGS	5	FC	97	98	97	97
TCGA3	768	Without Feature selection	2	DT	58	55.8	58.4	56.8
				GNB	57.3	54.5	69	60.5
				KNN	59.8	56.8	68	61.3
				SVM	62.8	63.3	56.3	59
				MLP	62.8	63	57.5	59
TCGA3	28	FGS	2	DT	58.34	55.9	59.8	56
				GNB	66.48	61	84	70.56
				KNN	63.94	59.87	75.73	66.46
				SVM	67.48	64	75.67	68.8
				MLP	65.44	63.8	64	63
TCGA3	28	FGS	2	FC	98	99	98	98
GSE43580	54675	Without Feature selection	2	DT	85.3	87.3	83.4	84.7
				GNB	84.6	91.5	75	82
				KNN	79.3	73	93	81.5
				SVM	83.3	84.3	80.6	82.3
				MLP	86	86.3	84.7	85
GSE43580	28	FGS	2	DT	79.3	77.6	82	79.4
				GNB	84.66	92.8	75	82.3
				KNN	83.3	91.95	72.57	80.46
				SVM	86.66	98	73	83.8
				MLP	75.3	75.55	74	74.6
GSE43580	28	FGS	2	FC	98	98	97	98

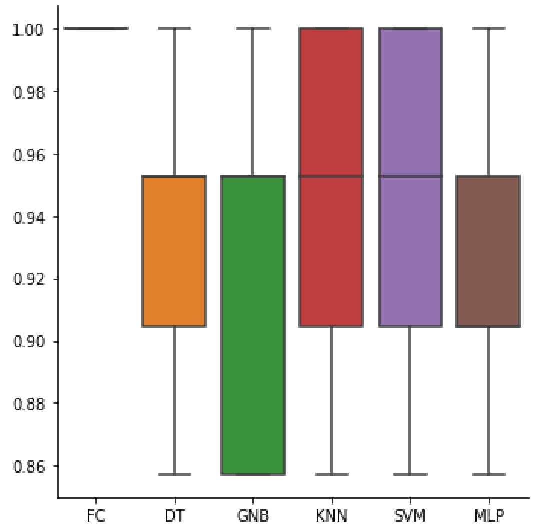

FIGURE 7. Accuracy scores of thyroid cancers (GSE33630) for five classifiers and FC in 5 kfolds.

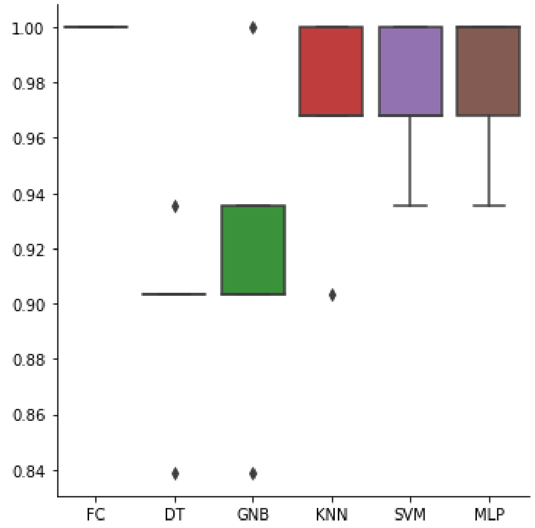

FIGURE 8. Accuracy scores of breast cancer subtypes (GSE45827) for five classifiers and FC in 5 kfolds.

other classifiers. FGS handles the dimensionality of gene expression data, where only 76 out of 23518 genes were chosen, whereas FC improved the accuracy for this dataset (GSE33630).

Figure8 shows a comparison of five common classifier techniques and FC in 5 kfolds to classify breast cancer subtypes when FGS employed. The highest results were 98%, 98.8%, 98% 98.3% when MLP was used. While FC achieved 100% accuracy, precision, recall, and f1-score. The proposed model proved its effectiveness in this dataset (GSE45827) by increasing the accuracy by 2%.

Figure9 displays the achievement accuracy of five classifiers and FC used to classify lung cancer. The accuracy is 90.8% DT, 95.8% GNB, and 96.66% for SVM, KKN, and MLP. While the proposed model achieved 100% as an accuracy score. Briefly, the improvement rate was 3.44% when compared to KNN, SVM, and MLP while 9.2% and 4.2% for DT and GNB respectively for (GSE19804) dataset when FC employed.

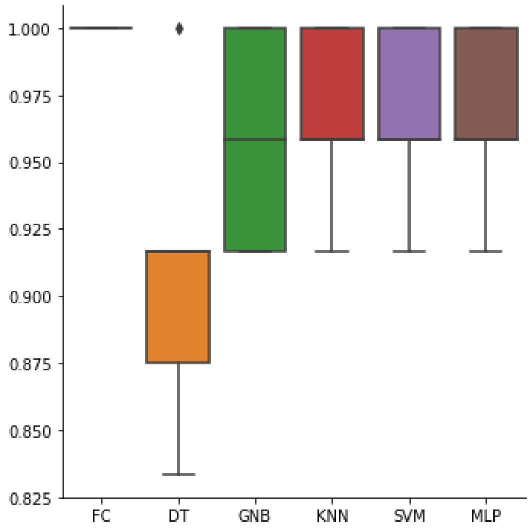

FIGURE 9. Accuracy scores of Lung cancer (GSE19804) for five classifiers and FC in 5 kfolds.

Figure10 illustrates the achieved accuracy scores of five classifier techniques and FC for five cancer types. When using the five classifier approaches with FGS, the acquired accuracy of the five classifiers was near to each other, with the lowest DT being 91.7% and the highest MLP being 94%. Even though the accuracy is 94%, the precision, recall, and f1-scores with MLP are 92.3%,91.6%, and 91.6%, respectively. The proposed model tried to increase not only the accuracy but also the other factors used to assess a model. FGS-FC improves all assessment parameters, including 97% accuracy, 95% for precision, recall, and f1-score. In short, the suggested model improved by 3% for accuracy, 2.7% for precision, and 3.4% for recall, and f1-score, when compared FC to the highest accuracy achieved by the five classifiers (MLP) and 5.3% when compared to DT classifier in the dataset (TCGA1).

Figure11 depicts a comparison of using several classifier approaches and FC for liver cancer gene expression datasets in 5 kfolds as cross-validation. The figure below represents the outcomes of applying ML algorithms when the FGS strategy was applied. Although, using the traditional classifier techniques and FGS for the (GSE77314) dataset accomplished satisfactory results. However, the proposed model demonstrates that it outperformed the highest accuracy by the five classifiers where FC achieved 100%.

Figure12 compares five classifier techniques and the FC method for classifying microarray gene expression liver cancer datasets in 5 kfolds as cross-validation. When utilising these classifier approaches, the obtained accuracy was 95%,

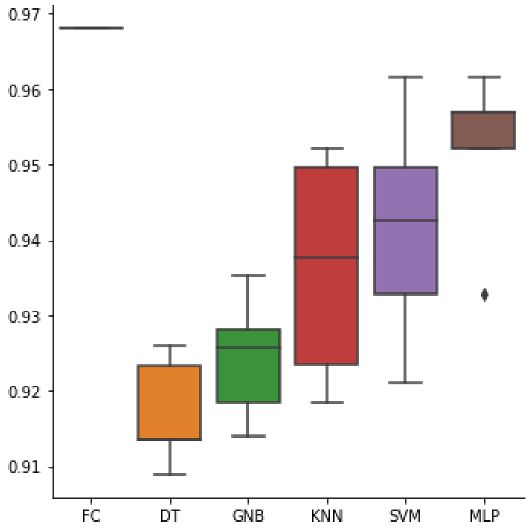

FIGURE 10. Accuracy scores of five cancer types (TCGA1) for five classifiers and FC in 5 kfolds.

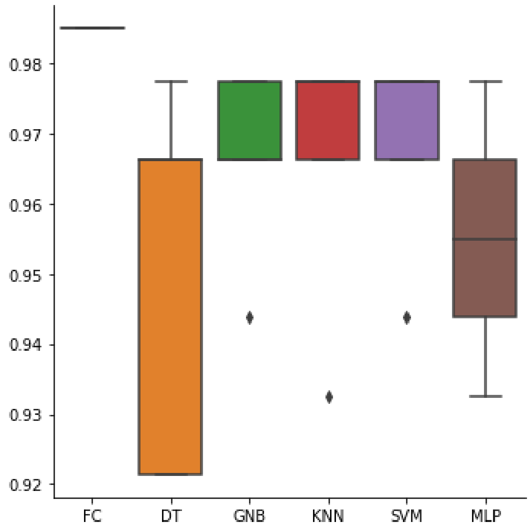

FIGURE 12. Accuracy scores of liver cancer (GSE14520) for five classifiers and FC in 5 kfolds.

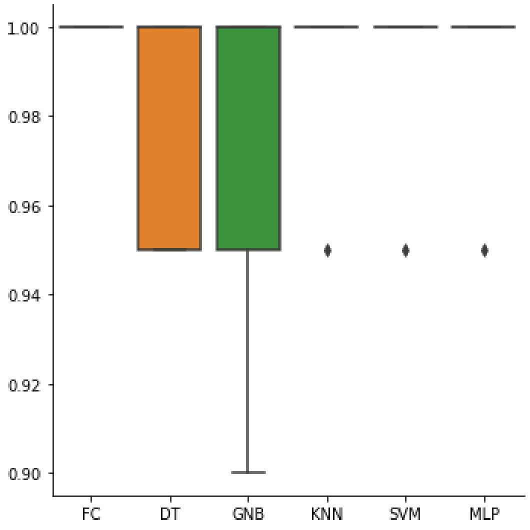

FIGURE 11. Accuracy scores of Liver cancer (GSE77314) for five classifiers and FC in 5 kfolds.

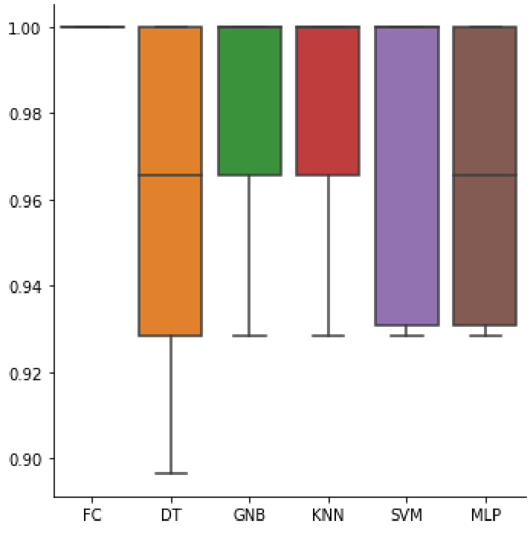

FIGURE 13. Accuracy scores of kidney cancer (GSE53757) for five classifiers and FC approaches in 5 Kfolds.

96.6%, 96.6%, 96.85%, and 95.5% for DT, GNB, KNN, SVM, and MLP respectively. While the suggested model demonstrated that achieved better accuracy compared to other classifier techniques by reaching 99% in this dataset (GSE14520). Although, the enhancement of proposing model was not high, however, FC always obtains the best accuracy in all given datasets rather than each classifier getting the best accuracy for a certain dataset.

Figure13 examined the accuracy scores of five commonly used classifiers and the FC method to classify the kidney cancer gene expression dataset (GSE53757) in 5 kfolds. The accuracy is shown in the flowchart when FGS technique applied for selecting significant genes (78 out of 23516 genes). Even though the accuracy scores were good, they ranged from 95.8% to 97.8%. In contrast, the proposed FC accuracy was 100% when using FGS. It can be said that the developing method FC has classified kidney cancer more accurately. It can conclude that no single classifier continuously achieved the highest accuracy for different datasets for example in this dataset KNN and GNB achieved the highest accuracy while in other datasets such as SVM or MLP and so on. By contrast, FC continuously achieved the highest accuracy for the different datasets as illustrated in this work.

Figure14 compares five classifier algorithms that are frequently employed for classification purposes and the proposed model FC in 5 kfolds for classifying the lung cancer dataset (GSE10072). The accuracy of the traditional

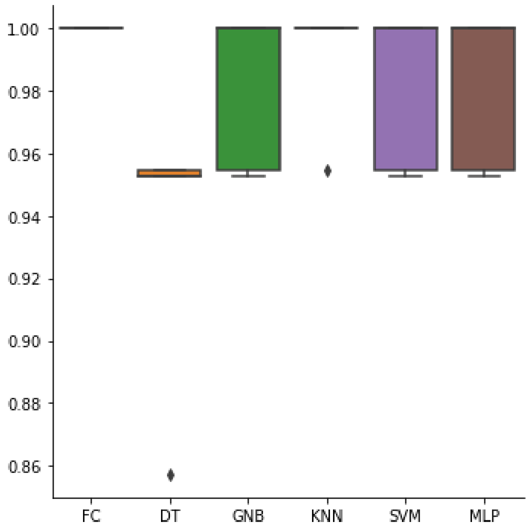

FIGURE 14. Accuracy scores for lung cancer (GSE10072) for five classifier and FC methods in 5 kfolds.

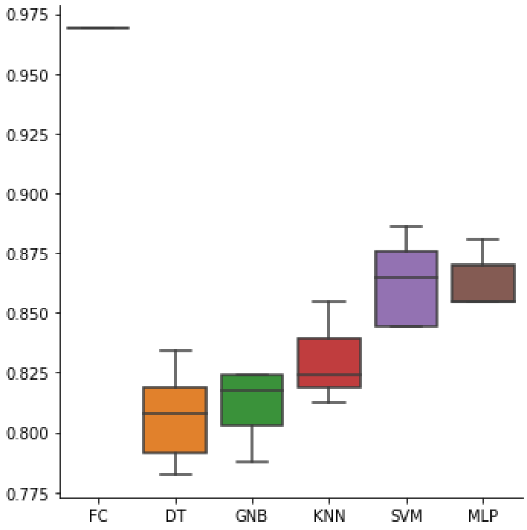

FIGURE 15. Accuracy score of Breast cancer subtypes (TCGA2) for five classifiers and FC in 5 kfolds.

classifiers ranged between 93.46% and 99% when FGS was used. As a result, the proposed model achieved 100% accuracy, precision, recall, and f1-score.

This is another challenging dataset (TCGA2) that was not classified accurately for breast cancer subtypes, whether using FGS or not, when classical classifier techniques were employed. This encourages the development of the FC model, which can provide reliable classification due to the sensitivity of the cancer topic. Figure15, provides a comparison of five classical classifier models and the FC method when FGS is employed. Although using FGS in this dataset marginally improved the accuracy of cancer classification, it still unsatisfactory. This discrepancy led to the development of a new approach capable of handling this type of data. An accuracy of 97% was attained with FC, whereas the accuracy ranged between 80.7% to 86.3% for the five classifier techniques. In brief, FC enhanced the accuracy score by 10.7% when compared to the highest accuracy achieved by the five classifier models and 16.3% when compared to the lowest accomplished accuracy among the five classifier approaches.

As previously stated, FC has been presented to handle datasets that haven't been successfully classified using FGS with traditional classifier techniques, achieving the highest accuracy across all used datasets. Since each classifier technique demonstrates the highest accuracy with specific data,we are perplexed about which one is the best choice for diverse data. Thus FC was developed to perform with various data consistently achieving the highest accuracy.

Figure16 compares the accuracy scores of five classifier techniques with FC when FGS is used. The findings demonstrate that integating the traditional classifiers and the FGS method achieved unsatisfactory accuracy when applied to liver cancer (TCGA) datasets. Traditional classifiers accomplished accuracy scores ranging between 58.34% and 67.48%. This dataset (TCGA3) is considered challenging as it has not been previously handled using FGS or other classifier approaches. FC has demonstrated its capacity to accurately classify this dataset, achieving 98% accuracy. In summary, the proposed model has increased the accuracy score by 30.52% when compared to the highest accuracy score achieved across the five classifier techniques.

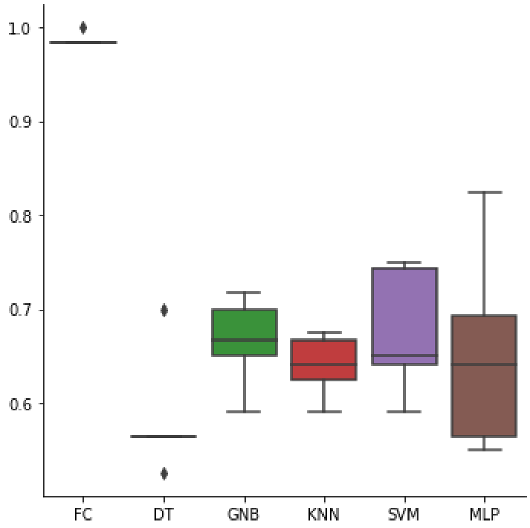

FIGURE 16. Comparing accuracy scores of Liver cancer (TCGA3) for five classifiers and FC in 5 kfolds.

Another challenge dataset was correctly classified as lung cancer using FGS and the traditional classifier methods. Despite the FGS technique reducing the number of identifying genes used to train the model, adequate accuracy was not reached. Consequently, employing the FC approach to deal with this data is critical. When SVM was applied,

TABLE 3. Comparing the proposed model (FGS-FC) to prior studies.

Datasets	Gene Number	Gene selection	Classifier Methods	Accuracy %	Precision %	Recall %	F1-score %	Reference
GSE66499	33298	No	CNN	81	88	78	74	[4]
	150	FGS	FC	95	95	92.4	93.4	The proposed model
TCGA4	67	ReliefF	RF	83.6	Unknown	Unknown	Unknown	[7]
	194	KL divergence	DNN	99	98	100	No	[22]
	10	FGS	FC	99	100	97	98	The proposed model
GSE19804	10	HLR	SVM	94.17	unknown	unknown	unknown	[16]
	36	FGS	FC	100	100	100	100	The proposed model
GSE14520	1253	No	DRE-DNN	82	83.3	95	88.9	[6]
	23	FGS	FC	99	98	99	98	The proposed model
TCGA3	768	NO	DRE-DNN	70	77.3	70.8	73.9	[6]
	28	FGS	FC	98.8	98.8	98.8	98.8	The proposed model
GSE33630	80	mRMR	KNN	93	unknown	unknown	unknown	[8]
	76	FGS	FC	100	100	100	100	The proposed model
GSE43580	43	MCSF and IFS	RF	88	82	97	89	[5]
	28	FGS	FC	98	98	97	98	The proposed model
GSE45827	38	Rough set	SVM	96.86	96.9	97.34	97.8	[19]
	68	FGS	FC	100	100	100	100	The proposed model
TCGA1	971	BPSO-DT	CNN	96	94.96	95	95	[20]
	25	FGS	FC	97	95	94	95	The proposed model

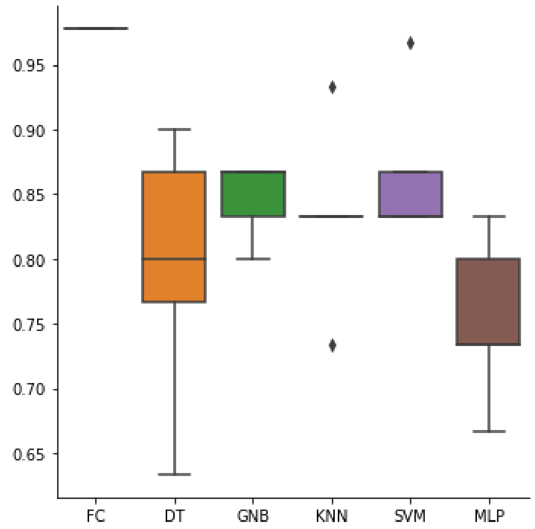

FIGURE 17. Accuracy score of lung cancer (GSE43580) for five classifiers and FC in 5 kfolds.

the best accuracy attained in the five classifier approaches was 86.66% with only 28 genes, while the suggested model achieved a 98% accuracy score. As shown Figure17 compares five classifier approaches and FC in 5 kfolds as cross-validation while using the FGS technique. FC proved its effectiveness by enhancing the accuracy by 11.34% compared to the highest accuracy achieved in the five classical classifier approaches.

In short, the fuzzy classifier approach was designed to improve cancer classification, particularly for datasets that encounter issues when a fuzzy gene selection technique is combined with the conventional classifiers commonly used in this field, such as SVM, MLP, etc. FC method demonstrated that can classifying cancer accurately with this challenge datasets that were shown low accuracy when FGS with traditional ML applied. It also showed successful cancer classification in most of the employed datasets where it is not improved but has not decreased the accuracy in all thirteen datasets that have been used. It can be concluded that the suggested model either greatly improves accuracy as seen in these datasets (GSE66499, GSE84437, TCGA1, TCGA2, TCGA3, and GSE43580) or somewhat improves slightly accuracy in the other datasets used with a small number of genes, as indicated in Table2.

In overall, the experimental outcomes unequivocally underscore the signficant superiority of the proposed model when contrasted with five exisiting classifiers. The performance of the proposed model, denoted as FC, yielding average scores of 98.2, 98.3, 96.8, and 97.2 for accuracy, precision, recall, and F1-score across all employed datasets in the expremint. In contrast, the existing classifiers exhibited a range of outcomes. Among them, the highest scores were achieved by the SVM classifier, reaching at 86.4 In summary, the experimental results form a compelling testament to the substantial performance leap offered by the proposed FC model, as evidenced by its superior metrics across a range of evaluation criteria. This comparison against existing classifiers underscores the efficacy of FC in enhancing accuracy, precision, recall, and F1-score, positioning it as a formidable contender in the realm of classification algorithms.

VIII. COMPARING FGS-FC TO EXISTING RESEARCH

The proposed model compared to ten previous published works that used the same datasets as our experiment used in terms of the number of genes and the evaluation performance. The results indicate that the proposed model outperforms the ten prior research in terms of evaluation performance and the number of genes that are used to train the classifier as shown in Table3. In summary, the developed model has greatly enhanced the results compared to previous studies [4], [5], [6], [7]. The findings achieved by the developed model demonstrated a significant reduction in the number of genes compared to prior studies. Specifically, our model successfully reduced the number of genes from 1253 (as reported in [6]) to a mere 23. Moreover, in comparison to the results in [20], where 971 genes were identified, our model accomplished a substantial decrease to only 25 genes. These outcomes demonstrate how well the proposed model significantly reduces the number of genes. Even though the proposed model accomplished results comparable to those of previous studies [20], [22], it stands out by using a significantly less number of genes. This key advantage over the published works is worth highlighting.

IX. CONCLUSION

This work involves the development of a novel fuzzy gene selection for selecting significant genes. Moreover, developing a novel fuzzy classifier method that significantly enhances the accuracy of cancer classification.This study aims to aid contributors in assisting biologists with the selection a subset of informative genes, crucial for delineating the specific type of cancer. This endeavor contributes to the early detection of cancer, thereby facilitating timely intervention. Furthermore, attaining the utmost classification accuracy. In short, this study and subsequent studies are attempting to identify the fewest genetic markers feasible to predict the kind of cancer. Consequently, this method decreases the complexity of the classification process while increasing accuracy.

The findings of this study, as noted above, show that the proposed approach outperformed commonly used classification techniques such as SVM, MLP, KNN, and others, as well as prior studies that employed the same data that was evaluated with the proposed technique. FGS was used as a gene selection technique for selecting informative genes that have a beneficial influence on classification. While FC has been proposed to gain accurate cancer classification. The new FC has addressed the issue that was not resolved when the FGS method and traditional machine learning approaches were employed together.

More crucially, the proposed model (FGS-FC) achieved accuracy in the thirteen utilized datasets ranging from 92.8% to 100% and using a small number of genes out of a large number of original accessible gene expression data. As a result, the high dimensionality issue of gene expression data has been handled in this work, indicating that the probability of overfitting occurring is low. Additionally, boost the classifier's generalization to be acceptable for multiple forms of cancer, where one classifier typically achieves the highest accuracy for a certain cancer dataset while another classifier can get the best accuracy with a different cancer type data. For example, the MLP classifier achieved the maximum accuracy with the GSE45827, TCGA1, GSE66499 and etc datasets, whereas the SVM classifier had the highest accuracy with the GSE43580, GSE14520 and etc datasets. The developed model was assessed using multi and binary-class datasets, as well as a small and high number of samples, to ensure that the model could properly classify cancer from different datasets.

Although, the proposed model has shown its effectiveness in cancer classification and gene selection. However, it has some limitations, compared with a limited number of prior studies. This work also focused on limited cancer types. Future work can compare the proposed model with a large number of previous studies. Additionally, using different cancer types to ensure that the proposed model has the ability to accurately classify the majority of cancer types. Future work also can focus on multi-omics integration such as gene expression, DNA methylation, and protein-protein interaction data, which can provide a more comprehensive understanding of biological processes. Future work can focus on developing machine learning techniques that effectively integrate and analyze multi-omics data to uncover complex relationships and mechanisms underlying gene expression regulation. Future work can explore the application of deep learning in determining robust gene expression biomarkers that can be used for early diagnosis, prognosis, and personalized treatment.

REFERENCES

[1] S. Shandilya and C. Chandankhede, "Survey on recent cancer classification systems for cancer diagnosis," in *Proc. Int. Conf. Wireless Commun., Signal Process. Netw. (WiSPNET)*, Mar. 2017, pp. 2590–2594.

[2] P. W. Wiest, J. A. Locken, P. H. Heintz, and F. A. Mettler Jr., "CT scanning: A major source of radiation exposure," *Seminars Ultrasound CT MR*, vol. 23, pp. 402–410, Oct. 2002.

[3] W. Dubitzky, M. Granzow, C. S. Downes, and D. Berrar, "Introduction to microarray data analysis," in *A Practical Approach to Microarray Data Analysis*, D. P. Berrar, W. Dubitzky, and M. Granzow, Eds. Boston, MA, USA: Springer, 2003.

[4] T. Matsubara, J. C. Nacher, T. Ochiai, M. Hayashida, and T. Akutsu, "Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles," in *Proc. IEEE 18th Int. Conf. Bioinf. Bioeng. (BIBE)*, Oct. 2018, pp. 151–154.

[5] F. Yuan, L. Lu, and Q. Zou, "Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms," *Biochimica et Biophysica Acta, Mol. Basis Disease*, vol. 1866, no. 8, Aug. 2020, Art. no. 165822.

[6] J. Li, Y. Ping, H. Li, H. Li, Y. Liu, B. Liu, and Y. Wang, "Prognostic prediction of carcinoma by a differential-regulatory-network-embedded deep neural network," *Comput. Biol. Chem.*, vol. 88, Oct. 2020, Art. no. 107317.

[7] J. Li, T. Ching, S. Huang, and L. X. Garmire, "Using epigenomics data to predict gene expression in lung cancer," *BMC Bioinf.*, vol. 16, no. S5, pp. 1–12, Dec. 2015.

[8] Y. Xu, Y. Deng, Z. Ji, H. Liu, Y. Liu, H. Peng, J. Wu, and J. Fan, "Identification of thyroid carcinoma related genes with mRMR and shortest path approaches," *PLoS ONE*, vol. 9, no. 4, Apr. 2014, Art. no. e94022.

[9] A. M. Hilal, A. A. Malibari, M. Obayya, J. S. Alzahrani, M. Alamgeer, A. Mohamed, A. Motwakel, I. Yaseen, M. A. Hamza, and A. S. Zamani, "Feature subset selection with optimal adaptive neuro-fuzzy systems for bioinformatics gene expression classification," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, May 2022.

[10] M. Rostami, S. Forouzandeh, K. Berahmand, M. Soltani, M. Shahsavari, and M. Oussalah, "Gene selection for microarray data classification via multi-objective graph theoretic-based method," *Artif. Intell. Med.*, vol. 123, Jan. 2022, Art. no. 102228.

[11] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, pp. 207–210, 2002.

[12] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013.

[13] A. Maharjan, "Machine learning approach for predicting cancer using gene expression," Comput. Sci. Theses, Dissertations, Professional Papers, Capstones, UNLV, 2020, p. 3922.

[14] R. Xie, J. Wen, A. Quitadamo, J. Cheng, and X. Shi, "A deep auto-encoder model for gene expression prediction," *BMC Genomics*, vol. 18, no. S9, p. 845, Nov. 2017.

[15] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," in *Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, vol. 4, Oct. 2019, pp. 135–139.

[16] C. A. Ul Hassan, M. S. Khan, and M. A. Shah, "Comparison of machine learning algorithms in data classification," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, Newcastle upon Tyne, U.K., Sep. 2018, pp. 1–6.

[17] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: Archive for functional genomics data sets–Update," *Nucleic Acids Res.*, vol. 41, pp. D991–D995, Jan. 2013.

[18] H.-H. Huang, X.-Y. Liu, and Y. Liang, "Feature selection and cancer classification via sparse logistic regression with the hybrid $L_{1/2+2}$ regularization," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0149675.

[19] S. Patil, K. R. Balmuri, J. Frnda, B. D. Parameshachari, S. Konda, and J. Nedoma, "Identification of triple negative breast cancer genes using rough set based feature selection algorithm & ensemble classifier," *Hum.-Centric Comput. Inf. Sci.*, vol. 12, p. 54, Nov. 2022.

[20] N. E. M. Khalifa, M. H. N. Taha, D. E. Ali, A. Slowik, and A. E. Hassanien, "Artificial intelligence technique for gene expression by tumor RNA-seq data: A novel optimized deep learning approach," *IEEE Access*, vol. 8, pp. 22874–22883, 2020.

[21] A. Yaqoob, R. M. Aziz, N. K. Verma, P. Lalwani, A. Makrariya, and P. Kumar, "A review on nature-inspired algorithms for cancer disease prediction and classification," *Mathematics*, vol. 11, no. 5, p. 1081, Feb. 2023.

[22] S. Liu and W. Yao, "Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection," *BMC Bioinf.*, vol. 23, no. 1, p. 175, Dec. 2022.

[23] M. Hamraz, Z. Khan, D. M. Khan, N. Gul, A. Ali, and S. Aldahmani, "Gene selection in binary classification problems within functional genomics experiments via robust Fisher score," *IEEE Access*, vol. 10, pp. 51682–51692, 2022, doi: 10.1109/ACCESS.2022.3172281.

[24] S. Bashir, I. U. Khattak, A. Khan, F. H. Khan, A. Gani, and M. Shiraz, "A novel feature selection method for classification of medical data using filters, wrappers, and embedded approaches," *Complexity*, vol. 2022, pp. 1–12, Aug. 2022.

[25] M. Khalsan, L. R. Machado, E. S. Al-Shamery, S. Ajit, K. Anthony, M. Mu, and M. O. Agyeman, "A survey of machine learning approaches applied to gene expression analysis for cancer prediction," *IEEE Access*, vol. 10, pp. 27522–27534, 2022, doi: 10.1109/ACCESS.2022.3146312.

MAHMOOD KHALSAN received the bachelor's degree in software engineering and the master's degree in computing from the University of Northampton, where he is currently pursuing the Ph.D. degree in computer science. Prior to doctoral studies, he was a Teaching Assistant with the University of Northampton. In addition, he was a Teaching Assistant with the University of Babylon. He received the fund from the Ministry of Higher Education, Iraq, for the bachelor's and master's degrees. His current research interest includes exploring the connection between cancer and genes through machine learning to assist with the early detection of cancer. He tries to develop an accurate machine-learning model capable of discovering the most probable genes associated with cancerous samples to advance the process of detecting cancer at early stages.

MU MU (Member, IEEE) is currently an Associate Professor with the University of Northampton, U.K. He has authored more than 50 peer-reviewed papers published in international conferences and journals. His current research interests include human factors in multimedia distribution, intelligent networks, and immersive media. He has been a principal investigator and a lead researcher of various research programs funded by the European Commission, the U.K. Research Councils, and other funding bodies. He has been the chair and a program committee member of renowned international conferences. He is an Associate Editor of *Multimedia Systems* (Springer) journal.

EMAN SALIH AL-SHAMERY received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Babylon, Iraq, in 1998, 2001, and 2013, respectively. After completing the M.Sc. degree, she was an Assistant Lecturer with the Department of Computer Science, University of Babylon. She is currently a Professor with the Software Department, University of Babylon. Her current research interests include artificial intelligence, bioinformatics, machine learning, neural networks, deep learning, and data mining.

SURAJ AJIT received the B.E. degree (Hons.) in computer science from Bangalore University, India, and the Ph.D. degree in computer science from the University of Aberdeen. He has five years of industry experience. He was a Research Scientist with BAE Systems, for three years. He was a Research Assistant in a prestigious advanced knowledge technologies project with the University of Aberdeen, for four years. He joined as a Software Engineering Lecturer with the University of Northampton, in 2011. He is currently an associate professor and the deputy subject leader overseeing the postgraduate courses in computing. His current research interests include software engineering, pedagogy (assessments and marking), constraints, and knowledge management. He is a fellow of the Higher Education Academy.

LEE R. MACHADO received the bachelor's degree from The University of Warwick and the Ph.D. degree in cancer studies from the University of Birmingham, U.K. He was a Postdoctoral Researcher with the Institute for Cancer Studies, Birmingham, the MRC Toxicology Unit, Leicester, and the Department of Genetics, Leicester, before working as a Senior Scientist with Cancer Vaccine Company and Scancell, Nottingham. He joined the University of Northampton, as a Lecturer, in 2013. He was an Interim Head of the Sport, Exercise and Life Sciences, from 2017 to 2018. He has three years of university board-level experience. He is currently a Professor of molecular medicine with the Division of Life Sciences; a Faculty Research and Enterprise Lead and the Co-Leader of the Molecular Biosciences Research Group, Physical Activity and Life Sciences Centre, University of Northampton; and an Honorary Research Fellow with the Department of Genetics and Genome Biology, University of Leicester. His current research interests include employing cellular and molecular genetic strategies to address how the host immune system responds to pathogens and cancer. This work aims to increase our understanding of human health and disease and develop rational therapeutic approaches to harness the exquisite specificity and sensitivity of the immune systems.

MICHAEL OPOKU AGYEMAN (Senior Member, IEEE) received the Ph.D. degree from Glasgow Caledonian University, U.K. He is currently a Professor and a Program Leader of computer systems engineering with the University of Northampton (UoN), U.K. He represents the Research Community of UoN at the University Senate. He is the Postgraduate (PGR) Lead of the Faculty of Arts Science and Technology and co-chairs the University's PGR Supervisory Forum. He has more than ten years of experience in embedded systems engineering. Previously, he was a Research Fellow with the Intel Embedded System Research Group, The Chinese University of Hong Kong (CUHK). He is the author of books, two book chapters, and more than 80 publications in major journals and conference proceedings. His current research interests include embedded systems and high-performance computing, such as VLSI SoC design, computer architecture, reconfigurable computing, wired and wireless NoCs, smart rehabilitation solutions, embedded systems, and the Internet of Things (IoT); business administration, such as neuromarketing, advertising, and market research; and pedagogy. He is a fellow of the Higher Education Academy, U.K. He is a Technical Committee Member of several conferences, such as IEEE ICCSN, IEEE ICBDA, and IEEE ICCT. He is a Chartered Engineer (C.Eng.) of IET and a Chartered Manager (C.Mgr.) of CMI. His work on wireless NoC has attracted two best paper awards in the 2016 IEEE/IFIP EUC and the 2016 Euromicro DSD 2016. He was a recipient of the 2018 International Changemaker of the Year Award in the First U.K. Ashoka U Changemaker Campus. He serves as a Reviewer of several conferences and journals, including IEEE Access. He has been a Guest Editor of the *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*.