Human-centric Computing and Information Sciences

November 2022 | Volume 12

www.hcisjournal.com

KIPS Korea Information Processing Society

KIPS CSWRG Korea Information Processing Society Computer Software Research Group

HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2022) 12:54
DOI: https://doi.org/10.22967/HCIS.2022.12.054
Received: January 21, 2022; Accepted: April 25, 2022; Published: November 30, 2022

Identification of Triple Negative Breast Cancer Genes Using Rough Set Based Feature Selection Algorithm & Ensemble Classifier

Sujata Patil[1], Kavitha Rani Balmuri[2], Jaroslav Frnda[3,4,*], Parameshachari B.D.[5], Srinivas Konda[6], and Jan Nedoma[4]

Abstract

In recent decades, microarray datasets have played an important role in triple negative breast cancer (TNBC) detection. Microarray data classification is a challenging process due to the presence of numerous redundant and irrelevant features. Therefore, feature selection becomes irreplaceable in this research field that eliminates non-required feature vectors from the system. The selection of an optimal number of features significantly reduces the NP hard problem, so a rough set-based feature selection algorithm is used in this manuscript for selecting the optimal feature values. Initially, the datasets related to TNBC are acquired from gene expression omnibuses like GSE45827, GSE76275, GSE65194, GSE3744, GSE21653, and GSE7904. Then, a robust multi-array average technique is used for eliminating the outlier samples of TNBC/non-TNBC which helps in enhancing classification performance. Further, the pre-processed microarray data are fed to a rough set theory for optimal gene selection, and then the selected genes are given as the inputs to the ensemble classification technique for classifying low-risk genes (non-TNBC) and high-risk genes (TNBC). The experimental evaluation showed that the ensemble-based rough set model obtained a mean accuracy of 97.24%, which is superior related to other comparative machine learning techniques.

Keywords

Ensemble Classifier, Machine-Learning Technique, Microarray Data, Robust Multi-Array Average Technique, Rough Set Theory, Triple Negative Breast Cancer

1. Introduction

Cancer is one of the most serious health problems in the world, which begins in the cells of the human body. Cancer is defined as an uncontrolled growth of abnormal cells anywhere in the human body, and these cells are termed malignant, tumor, or cancer cells, where these cells affect normal body tissues [1]. Cancer develops from a series of genetic mutations that stop checking normal cell growth, with these

***Corresponding Author:** Jaroslav Frnda (jaroslav.frnda@fpedas.uniza.sk)
[1]Department of ECE, KLE Dr. M. S. Sheshgiri College of Engineering and Technology, Belgaum, India
[2]Department of Information Technology, CMR Technical Campus, Hyderabad, Telangana, India
[3]Department of Quantitative Methods and Economic Informatics, Faculty of Operation and Economics of Transport and Communication, University of Zilina, Zilina, Slovakia
[4]Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Ostrava-Poruba, Czech Republic
[5]Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bengaluru, India
[6]Department of Computer Science Engineering, CMR Technical Campus, Kandlakoya, Hyderabad, India

cells continuing to grow, divide, and develop into cancer [2, 3]. Cancer that develops in the breast tissues is called breast cancer, and usually develops in the lobules or inner lining of milk dust that is responsible for milk supply in the ducts [4, 5]. In recent decades, breast cancer is a crucial reason for a respective high female mortality rate. Around 10%–15% of breast cancer patients have pain in the breast [6]. Breast cancer symptoms are swelling, dimpling of the skin surface, and skin with an orange appearance, skin irritation, nipple discharge, and tenderness nipple inversion [7, 8]. Sometimes, the overly growing cancer cells dilate the veins on the breast surface. The characteristics of cancer cells depends on the structure of nucleus, cell outline, capacity to metastasize, and its shape. Generally, the human detection of triple negative breast cancer (TNBC) utilizing microarray data is effective, but will not be accurate in all circumstances. In addition, the human detection consists of two major concerns, such as consuming more time for classifying TNBC and non-TNBC genes and being only suitable for minimum data [9, 10]. To address the above-stated concerns, numerous machine-learning techniques are developed by researchers that nearly made a huge impact in TNBC detection. In this manuscript, a new model is implemented for effective TNBC detection using microarray data. The major contributions of this work are listed as follows:

- Firstly, microarray data related to TNBC/non-TNBC are acquired from six datasets, namely GSE45827, GSE76275, GSE65194, GSE3744, GSE21653, and GSE7904.
- Next, the outlier samples of TNBC/non-TNBC are eliminated using robust multi-array average (RMA) technique. The RMA technique summarizes the perfect matching genes through media polish, which is robust, and it behaves based on the number of analyzed samples. In microarray data classification, the RMA technique includes three major advantages like quality control, spot filtering, and background correction.
- After normalizing the TNBC/non-TNBC samples, gene selection is carried out by using the rough set-based feature selection algorithm on individual gene expression omnibus IDs.
- After selecting optimal genes in every gene expression omnibus ID, the ensemble classifier (combination of a k-nearest neighbor [kNN] and support vector machine [SVM]) is applied to classify the low-risk genes (non-TNBC) and high-risk genes (TNBC). The motivation behind an ensemble classifier is to learn a set of classifiers (combination of a kNN and SVM) and vote for the best results using soft voting, which obtains better results compared to individual classifiers. Soft voting predicts the class with the highest summed probability from a kNN and SVM.
- The proposed ensemble-based rough set model's effectiveness is tested in terms of the related Matthews correlation coefficient (MCC), F-measure, precision, recall, and accuracy.

This manuscript is organized as follows. A few articles on the topic "microarray data classification" are reviewed in Section 2, and the problem statement with motivation is given in Section 3. The theoretical description and experimental evaluation of an ensemble-based rough set model is represented in Sections 4 and 5, while the conclusion of this manuscript is specified in Section 6.

2. Related Work

Li et al. [11] introduced a new machine-learning model for identifying TNBC-related genes. In this literature study, seven gene expression datasets, namely GSE15852, GSE45255, GSE32646, GSE20271, GSE20194, GSE9574, and GSE31519 were utilized for experimental evaluation. The KEGG pathway examination showed that 54 genes were related to viral carcinogenesis, and a gene ontology investigation indicated that the organic cyclic compound in the cellular response influences the onset of breast cancer. Additionally, a machine learning technique uses a SVM to predict the high-risk of breast cancer. The experimental examination showed that the presented model significantly identifies cancer-related genes, and assists physicians in medical diagnosis. However, a SVM classifier performs only binary-class classification, but it was inappropriate for multi-class classification. Cai et al. [12] used Dijkstra's algorithm for finding the genes that mediate bone cancer metastasis to breast cancer. Many putative genes were determined using Dijkstra's algorithm from large networks, and then a protein-to-protein interaction

(PPI) was constructed using the selected genes of breast and bone cancers. Overall, eighteen putative genes were determined utilizing Dijkstra's algorithm, with the experimental result confirming that these putative genes participates in metastasis. However, the Dijkstra's algorithm does a blind search that is time-consuming in finding the unnecessary resources.

Sarkar et al. [13] used a random forest classifier with a seven-feature selection algorithm such as joint mutual information, minimum redundancy maximum relevance, double input symmetrical relevance, conditional infomax feature extraction, mutual information maximization, interaction capping and conditional mutual information maximization to predict any breast cancer subtype miRNA biomarkers. Additionally, a cox regression-based survival investigation was carried out for finding important miRNAs for breast cancer detection. However, the computational time of the developed model was high by implementing several feature selection algorithms in this study. Mahapatra et al. [14] combined an extreme gradient boosting classifier and deep neural network to predict PPIs. In this literature, three sequence-based features such as a local descriptor, conjoint triad composition, and amino acid composition were given as the input to a hybrid classifier. The experimental analysis showed that the hybrid classifier effectively predicts the inter-species and intra-species PPIs, and the developed hybrid classifier obtained a better classification accuracy on the independent test sets, which represent that it could be used for cross-species prediction. Pan et al. [15] combined the fast Walsh-Hadamard transform and random forest classifier for predicting plant PPIs. The introduced model performance was tested on the three plant's PPI datasets such as Arabidopsis thaliana, maize, and rice. However, the random forest classifier consumes more computational power and resources for predicting plant PPIs which was considered a major concern in this literature study.

Naorem et al. [16] used a correlation-based feature selection algorithm and naïve Bayes classifier for classifying non-TNBC and TNBC samples from GSE45827, GSE21653, GSE76275, GSE7904, GSE3744, and GSE65194 datasets. The experimental investigations suggested that the selected key candidate genes were a therapeutic target for TNBC treatment. The implemented model was more appropriate in structured data, but obtained limited performance with unstructured data. Wang et al. [17] implemented a new computational model that integrates a random forest, rough set-based rule learning, and a Monte Carlo feature selection algorithm for identifying the genes that were related to original human tumors and breast cancer. Among 831 breast tumors, 32 optimal genes were determined for constructing a prediction model. The presented model experiences a class imbalance problem in a few circumstances that was a major issue in this literature study. Zhang et al. [18] developed a random walk with a restart algorithm and PPI network for identifying the proliferative diabetic retinopathy-related genes. The random walk with a restart algorithm was applicable for a two-class classification, but not for multiclass classification. Al-Safi et al. [19] and Iswisi et al. [20] developed a Harris Hawks optimization (HHO) algorithm for an effective feature selection. Additionally, a majority voting learning method was utilized to diagnose the disease type in medical centers. Al-Safi, et al. [21] integrated the HHO algorithm and artificial neural network for heart disease diagnosis. In addition, several optimization algorithms like a black widow spider optimization algorithm [22], particle swarm optimization (PSO) [23], hybrid particle swarm optimization [24], artificial bee colony (ABC) optimization algorithm [25], polar bear optimization [26], and principal component analysis [27] were also preferred in gene selection.

3. Problem Statement and Motivation

By reviewing the existing literatures, some common concerns faced by the researchers in breast cancer detection using microarray data are listed as follows:

- Microarray data acquisition and pre-processing unit consist of a major problem of being difficult to acquire the quality medical data by a user, due to the limit of capturing technology or adverse environmental conditions.
- While experimenting with supervised machine learning methods, the semantic space is maximized between the feature values that lead to poor classification performance.

- The clustering techniques improve the gene classification accuracy, but it is time-consuming since it computes the neighborhood term in every iteration step. To address the highlighted concerns, a new ensemble-based rough set model is implemented in this manuscript to improve the performance of TNBC and non-TNBC detection by using microarray data.

4. Methodology

The ensemble-based rough set model includes four phases in microarray data classification as follows:
- Data collection: TNBC microarray expression datasets (GSE45827, GSE76275, GSE65194, GSE3744, GSE21653, and GSE7904);
- Data pre-processing: robust multi-array average technique;
- Optimal gene selection: rough set theory; and
- Gene classification: ensemble classifier.

A flowchart of the ensemble-based rough set model is illustrated in Fig. 1.

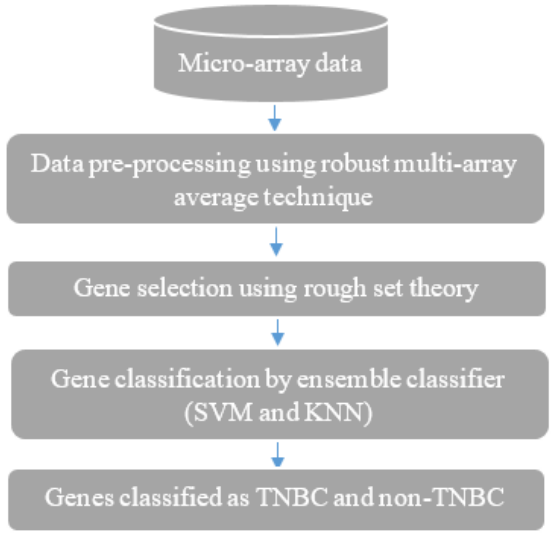

Fig. 1. Flowchart of ensemble-based rough set model.

4.1 Data Collection and Pre-processing

In this manuscript, the proposed model's effectiveness is tested on TNBC microarray expression datasets, which are collected from a gene expression omnibus (GEO) data repository using the keywords of “breast cancer, basal like breast cancer, and triple negative breast cancer.” The datasets are manually reviewed for fulfilling the criteria falling under (i) studies without drug treatments, (ii) studies involved in human sample selection, and (iii) datasets belonging to the gene expression profiles of breast cancer between non-TNBC and TNBC. The data statistic about the undertaken datasets such as accession number, organism, and number of samples are represented in Table 1. These datasets are publicly available at http://www.ncbi.nlm.nih.gov/geo/.

In this study, the collected TNBC microarray expression datasets (GSE45827, GSE7904, GSE3744, GSE21653, GSE65194, and GSE76275) consist of 405 TNBC samples, and 463 non-TNBC samples, as mentioned in Table 1. Next, the RMA technique is applied to correct, normalize, and summarize the probe level information of the Affymetrix data. In the RMA technique, the raw intensity variables are corrected, log-2 values are transformed, and then the quantiles are normalized. After normalizing the

collected microarray data, the fifty outlier samples are eliminated, and then the residual 818 samples are used for further intended operations. The eliminated fifty outlier samples are depicted in Table 2.

Table 1. Data statistic of the undertaken datasets

Accession No.	Organism	Number of samples	
		TNBC	Non-TNBC
GSE45827	Homo sapiens	41	89
GSE7904	Homo sapiens	18	25
GSE3744	Homo sapiens	18	22
GSE21653	Homo sapiens	75	162
GSE65194	Homo sapiens	55	98
GSE76275	Homo sapiens	198	67

Table 2. Eliminated 50 outlier samples

Dataset	Eliminated samples
GSE45827	GSM1116215, GSM1116087, GSM1116190, GSM1116092, GSM1116146, and GSM1116093
GSE7904	GSM194406 and GSM194408
GSE3744	GSM85484, GSM85482, and GSM85497
GSE21653	GSM540108, GSM540323, GSM540109, GSM540324, GSM540110, GSM540325, GSM540130, GSM540332, GSM540139, GSM540343, GSM540141, GSM540322, GSM540148, GSM540319, GSM540201, GSM540231, GSM540195, GSM540214, and GSM540317
GSE76275	GSM1978928, GSM1978939, GSM1978917, GSM1978900, GSM1974760, GSM1978916, GSM1974736, GSM1974750, GSM1974732, GSM1974716, GSM1974723, GSM1974584, GSM1974605, GSM1974666, and GSM1974717
GSE65194	GSM1588987, GSM1588986, GSM1589015, GSM1589012, and GSM1589116

4.2 Gene Selection

After removing the outlier samples, gene selection is carried out by using a rough set-based feature selection algorithm. A rough set is a novel intelligent mathematical tool, which is utilized to deal with data incompleteness and uncertainty. A rough set model works based on a lower and upper approximation of a set, and its major benefit of use is that it does not require any additional or preliminary information of data such as probability in statistics or assignment in Dempster-Shafer theory and membership grade in a fuzzy set theory. The attribute reduction is a major application of rough set theory, which is accomplished by relating the equivalence relations generated by attribute sets. Hence, reduced or removed attributes delivers a similar degree of the original data by utilizing the dependency degree as a measure. The systematic procedure of rough set theory is given below [28, 29].

4.2.1 Information system

In rough set theory, the knowledge representation is made utilizing information system that is represented as four tuple, as mentioned in Equation (1).

$$S =< U, A, V, f > \tag{1}$$

where, U indicates closed universe with a finite set of N objects$\{x_1, x_2 \dots, x_n\}$, A denotes finite set of attributes $\{a_1, a_2 \dots, a_n\}$ that is further subdivided into two disjoint subsets of C and D. Hence, C denotes conditional attributes, and D states decision attributes. Whereas, $f: U \times A \to V$ is stated as a total decision function and called an information function $f(x, a) \in V_a$ for each $a \in A, x \in U$. Additionally, $V = U_{a \in A} V_a$ and V_a is denoted as a domain of the attribute.

4.2.2 Indiscernibility relation

In rough set theory, the most significant characteristic is indiscernibility relation $IND\ (R)$, which is denoted in Equation (2).

$$IND\ (R) = \{(x, y) \in U \times A: a \in R, a(x) = a(y)\} \tag{2}$$

where, $a\ (x)$ represents attribute value a of object x. If $IND\ (R) \in \{(x, y)$, then x and y are denoted as indiscernible values with respect to R. Therefore, $[x]_R$ is expressed as equivalence classes of the R-indiscernibility relation.

4.2.3 Lower and upper approximations

Lower and upper approximations are considered as two basic operations in a rough set theory. For any attribute set $R \subseteq A$, and any concept $X \subseteq U$, X is approximated by lower and upper approximations that are denoted in Equations (3) and (4).

$$R(X) = \{x \in U: [x]_R \subseteq X\} \tag{3}$$

$$\bar{R}(X) = \{x \in U: [x]_R \cap X \neq \emptyset\} \tag{4}$$

Meanwhile, the R-boundary region of X is denoted in Equation (5).

$$Bnd\ (X) = R(X) - \bar{R}(X) \tag{5}$$

where, A denotes the boundary region and is non-empty.

4.2.4 Core and attribute/gene reduction

In rough set theory, a few conditional attributes does not deliver proper information about the objects in U. Therefore, the redundant attributes are used for removing any unwanted attributes without losing the necessary classification information. Hence, the core and reduct attribute sets are the main concepts in rough set theory. The reduct attribute is utilized to reduce the attributes from A that provides proper gene classification with a full set of attributes. The given D and $C \subseteq A$ is the minimum attribute set such that $IND(D)$ is equal to $IND(C)$. Where $RED(A)$ represents the reduct of A, and the intersection of all reducts of A is denoted in Equation (6).

$$CORE\ (A) = \cap RED\ (A) \tag{6}$$

4.2.5 Dependency degree

Several measures are determined for representing how much the attribute C depends on the decision attributes D. The most common measure is dependency degree $\gamma_C(D)$ that is denoted in Equation (7).

$$\gamma_C(D) = |POS_C(D)|/|U| \tag{7}$$

where, $|POS_C(D)|$ indicates the positive region and $|U|$ states the cardinality set. Therefore, the workflow of a rough set theory is illustrated in Fig. 2. In addition, the fitness comparison of different techniques such as PSO, genetic algorithm (GA), ant colony optimization (ACO), ABC, grey wolf optimization (GWO), reliefF, infinite, and rough set theory by varying the iteration number is illustrated in Fig. 3.

4.3 Gene Classification

The selected 38 genes in each gene expression omnibus ID are fed to the ensemble classifier for classifying low-risk genes (non-TNBC), and high-risk genes (TNBC). An ensemble classifier significantly

enhances the machine learning results by integrating several models, which results in better gene classification related to individual classification techniques. In this manuscript, the ensemble classifier combines a kNN and SVM classifier. The kNN classifier depends on the calculation of distance between training and testing samples A for identifying the nearest neighbors. In the kNN classifier, k value represents the number of nearest neighbors, and helps decide the nearest neighbor values that influences gene classification. In the kNN classifier, the nearest neighbors are selected based on training and testing samples A. Numerous distance measures are utilized to calculate the distance between training and testing samples A like Chebyshev distance, city-block, Euclidean, Minkowski, etc. Among these available distance measures, the Euclidean distance is utilized to calculate the distance between training and testing samples A. The formula of Euclidean distance measure is denoted in Equation (8).

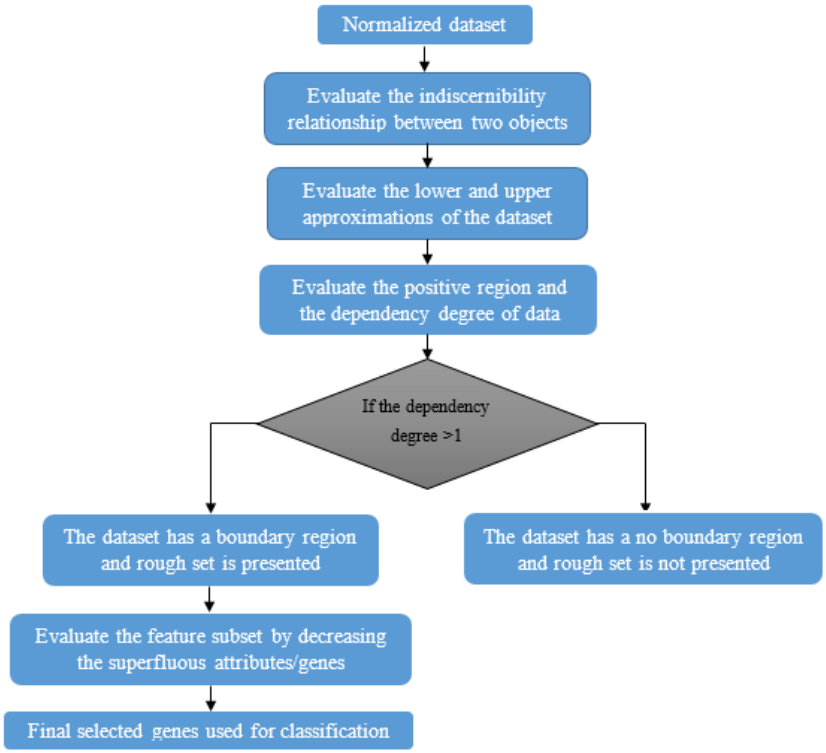

Fig. 2. Rough set theory workflow.

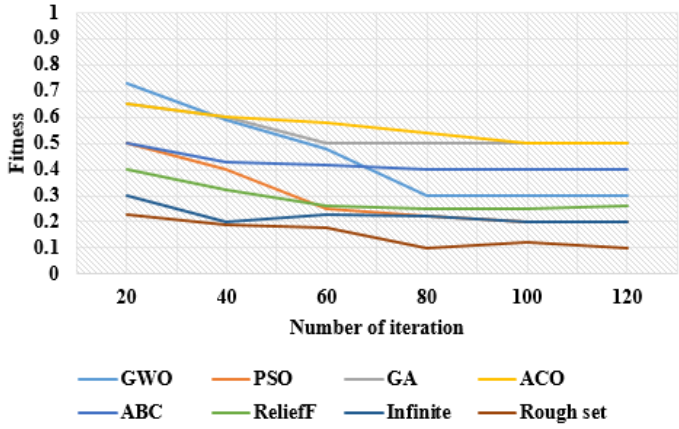

Fig. 3. Fitness comparison of different techniques by varying iteration number.

$$Euclidean\ distance = \sqrt{\sum_{i=1}^{N}(q_i - p_i)^2} \quad (8)$$

where, N indicates the number of samples A, while q_i and p_i represent testing and training samples, respectively [30]. Additionally, the SVM is a discriminative classifier, which is represented by a separate hyper-plane. Compared to other machine learning classifiers, the SVM classifier works well in solving the two-class problems by using Vapnik-Chervonenkis theories and structure principles. The mathematical formula to calculate the linear discriminant function is denoted in Equation (9).

$$w.z + b = 0 \quad (9)$$

In the SVM classifier, the hyper-plane is applied between the two classes: low-risk genes (non-TNBC) and high-risk genes (TNBC) [31] that is denoted in Equation (10).

$$pi[w.z + b] - 1 \geq 0, i = 1,2,..N \quad (10)$$

Next, $\|w\|^2$ is minimized in Equation (10) and the ideal discriminant function is denoted in Equation (11). Where, α_i indicates the Lagrange function with Lagrange multipliers.

$$f(z) = sign\{(w^*z) + b^*\} = sign\{\sum_{i=1}^{N} \alpha_i^* . pi(z_i^* - z) + b^*\} \quad (11)$$

Finally, the interior product $(z_i^* - z)$ is interchanged by a linear kernel function $k(z, z')$ in Equation (11). The linear separability of the estimated samples is improved, and the discriminant function is re-written as represented in Equation (12).

$$f(z) = sign\{\sum_{i=1}^{N} \alpha_i^* . pi. k(z, z_i) + b^*\} \quad (12)$$

The parameter settings of the kNN and SVM classifiers are listed as follows: the number of neighbors is 10; the distance metric is Euclidean; kernel function is Gaussian; kernel scale is 0.56; and box-constrained level is one. These parameters control the learning process and determine the model parameter values, where the learning algorithm ends up learning. The selection of appropriate parameters maximizes the model's predictive accuracy. The experimental investigation of the proposed ensemble-based rough set model is given in the next section.

5. Experimental Results

The ensemble-based rough set model's efficiency is validated using MATLAB (version 2020) on a system configuration with 64 GB random access memory, 4 TB hard disk, Intel Core i9 Processor, and Windows 10 operating system. The efficiency of the proposed ensemble-based rough set model is evaluated using performance measures like MCC, F-measure, precision, recall, and accuracy. The precision performance metric quantifies the number of positive class prediction, which belong to the positive class, while the recall performance measure quantifies the number of positive classes in the datasets of GSE45827, GSE76275, GSE65194, GSE3744, GSE21653, and GSE7904. Further, the F-measure includes a single score for balancing the issues of both recall and precision in a single number. The mathematical formulas of precision, recall, and f-measure are denoted in Equations (13)–(15) respectively. In addition, the MCC and accuracy are utilized for measuring the ratio between the overall samples and number of correctly classified samples. The equations of the MCC and accuracy are defined in Equations (16) and (17), respectively.

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (13)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (14)$$

$$F - measure = \frac{2TP}{FP+2TP+FN} \times 100 \quad (15)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100 \quad (16)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (17)$$

where, FP, FN, TP, and TN are represented as false positive, false negative, true positive, and true negative.

5.1 Quantitative Study

In this scenario, the ensemble-based rough set model's effectiveness is validated on six datasets related to TNBC, namely GSE45827, GSE76275, GSE65194, GSE3744, GSE21653, and GSE7904. The ensemble-based rough set model's efficiency is evaluated by utilizing a 10-fold cross-validation with 80% training and 20% testing of the microarray data. In this study, the performance valuation is carried out by using different feature selection techniques such as reliefF, infinite, and rough set theories using MCC, F-measure, precision, recall, and accuracy. By investigating Table 3, a rough set theory with an ensemble classifier obtained a mean accuracy of 97.24%, MCC of 97.28%, F-measure of 96.95%, precision of 97.01%, and recall of 96.62%, where the obtained experimental results are superior related to reliefF and infinite feature selection techniques. The comparison results of the different feature selection techniques are illustrated in Fig. 4. In this study, a rough set theory effectively reduces the number of attributes contained in the dataset and enables the discovery of data dependencies without the need for additional information.

Table 3. Experimental results of different feature selection techniques (unit: %)

Dataset	Feature selection	MCC	F-measure	Precision	Recall	Accuracy
GSE45827	ReliefF	94.09	94.52	95.67	90.30	92.90
	Infinite	96.88	96.47	95.06	95.07	95.08
	Rough set	98.30	97.80	96.90	97.34	96.86
GSE7904	ReliefF	93.02	95.02	94.06	90.89	90.03
	Infinite	96.49	96.80	94.80	94.32	95.02
	Rough set	97.59	98.74	97.98	95.38	97.80
GSE3744	ReliefF	93.88	92.34	95.90	92.44	94.20
	Infinite	95.94	94.60	96.95	92.92	94.90
	Rough set	96.74	95.56	98.95	95.93	97.90
GSE21653	ReliefF	93.94	93.06	94.30	93.56	92.70
	Infinite	95.90	93.90	93.80	95.44	94.06
	Rough set	97.48	95.94	96.67	97.80	96.38
GSE76275	ReliefF	92.65	94.04	93.03	94.30	94.09
	Infinite	94.07	95.90	94.08	95.03	95.90
	Rough set	96.67	96.84	95.54	96.90	96.77
GSE65194	ReliefF	92.88	92.92	92.03	93.92	92.30
	Infinite	94.86	95.20	95.40	94.09	95.20
	Rough set	96.93	96.83	96.02	96.39	97.74

In Table 4, the performance valuation is accomplished using different classifiers such as a random forest, kNN, naïve Bayes, and ensemble classifier with a rough set-based feature selection technique. As

seen in Table 4, the ensemble classifier achieved maximum performance in gene classification compared to the individual classifiers with respect to MCC, F-measure, precision, recall, and accuracy. The comparison results of the different classification techniques are illustrated in Fig. 5. In this manuscript, the ensemble classifier delivers superior classification results than any single contributing model. The proposed ensemble classifier decreases the dispersion or spread of gene classification.

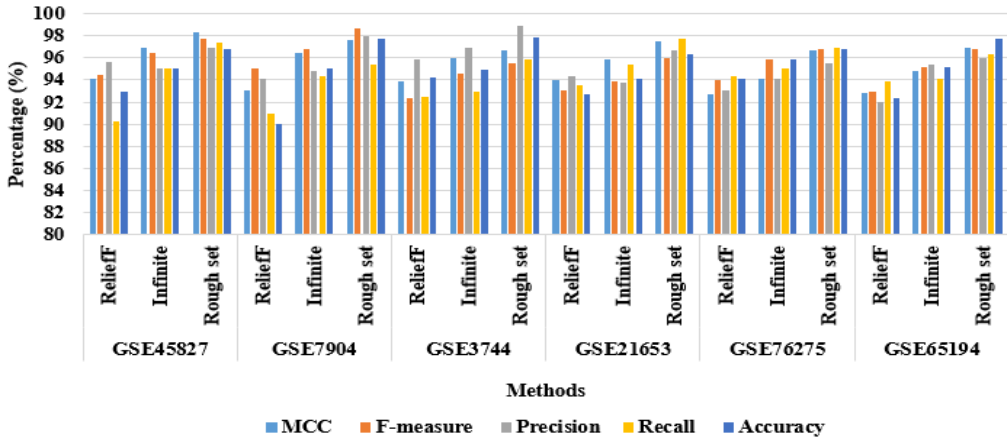

Fig. 4. Comparison results of different feature selection techniques.

Table 4. Experimental results of different classification techniques (unit: %)

Dataset	Classifiers	MCC	F-measure	Precision	Recall	Accuracy
GSE45827	Random forest	92.00	91.92	90.63	87.30	89.90
	kNN	92.89	94.44	92.03	92.03	92.03
	Naïve Bayes	96.80	95.50	94.00	95.92	94.30
	Ensemble	98.30	97.80	96.90	97.34	96.86
GSE7904	Random forest	91.02	92.02	91.06	89.89	92.03
	kNN	93.49	94.80	92.80	90.32	94.02
	Naïve Bayes	94.04	96.67	93.02	93.28	95.50
	Ensemble	97.59	98.74	97.98	95.38	97.80
GSE3744	Random forest	92.87	93.39	92.90	90.44	93.20
	kNN	91.90	93.62	94.00	91.92	91.02
	Naïve Bayes	93.82	94.09	95.80	93.28	93.10
	Ensemble	96.74	95.56	98.95	95.93	97.90
GSE21653	Random forest	90.00	94.06	92.00	90.56	89.70
	kNN	90.90	92.90	93.82	94.44	92.06
	Naïve Bayes	95.38	94.37	95.30	95.46	94.34
	Ensemble	97.48	95.94	96.67	97.80	96.38
GSE76275	Random forest	87.65	94.04	94.03	90.00	90.09
	kNN	92.07	93.90	93.03	92.03	90.90
	Naïve Bayes	93.34	94.80	93.90	93.00	93.80
	Ensemble	96.67	96.84	95.54	96.90	96.77
GSE65194	Random forest	88.88	93.92	90.00	89.92	93.30
	kNN	90.56	93.29	89.45	92.02	93.20
	Naïve Bayes	93.92	95.40	92.30	94.90	94.00
	Ensemble	96.93	96.83	96.02	96.39	97.74

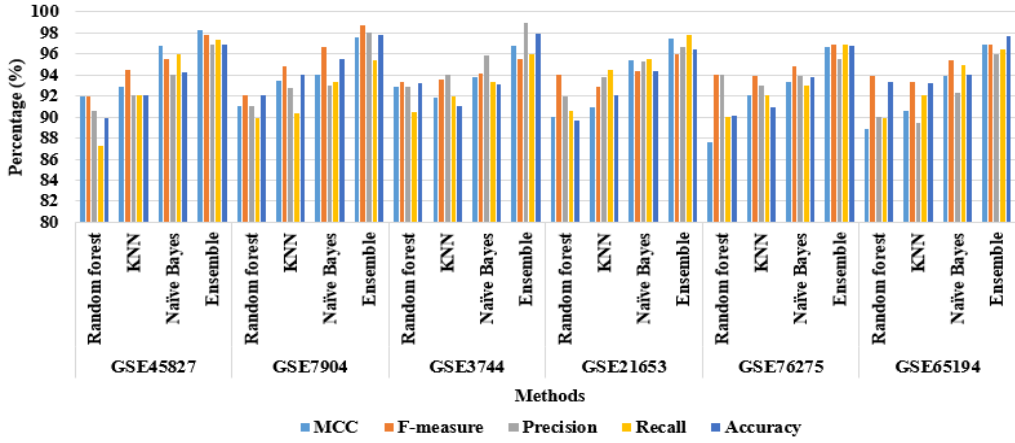

Fig. 5. Comparison results of different classification techniques.

5.2 Comparative Study

The comparative investigation between the proposed ensemble-based rough set model and existing model is represented in Table 5 and Fig. 6. Naorem et al, [16] have combined a correlation-based feature selection algorithm and naïve Bayes classifier for classifying non-TNBC and TNBC genes. The extensive experiments specified that the selected candidate genes were therapeutic targets for TNBC treatment. In the resulting section, the presented model achieved better gene classification performance on GSE3744, GSE7904, GSE45827, and GSE65194 datasets with respect to MCC, F-measure, precision, recall, and accuracy.

Table 5. Comparative results (unit: %)

Dataset	Models	MCC	F-measure	Precision	Recall	Accuracy
GSE45827	Existing [16]	80.60	90	92.10	89.70	89.72
	Proposed	98.30	97.80	96.90	97.34	96.86
GSE7904	Existing [16]	63.10	81.10	82.80	81	80.95
	Proposed	97.59	98.74	97.98	95.38	97.80
GSE3744	Existing [16]	63.10	81.20	82.30	81.10	81.08
	Proposed	96.74	95.56	98.95	95.93	97.90
GSE65194	Existing [16]	80.80	90.10	92.20	89.90	89.86
	Proposed	96.93	96.83	96.02	96.39	97.74

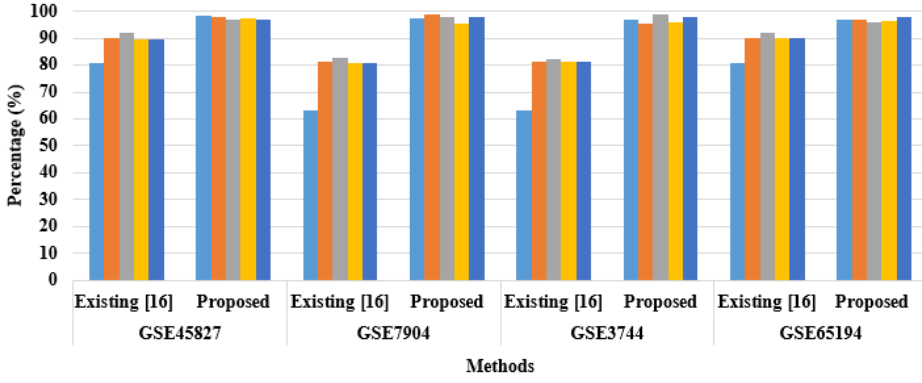

Fig. 6. Graphical comparison of ensemble-based rough set model and existing model.

5.3 Discussion

By investigating Table 5, the ensemble-based rough set model obtained superior gene classification performance related to the existing models. By utilizing a rough set theory, the optimal genes are selected from each GEO-IDs on datasets like GSE45827, GSE76275, GSE65194, GSE3744, GSE21653, and GSE7904. This process significantly decreases the computational time and complexity of the classifier, and overcomes the problems mentioned in literatures [11–13, 15]. However, the computational complexity of the ensemble-based rough set model is linear $O(N)$, where, O indicates order of magnitude and N states input size, while the proposed model consumes 34.20 seconds to train and test the data, which is limited related to comparative classifiers such as random forest, kNN, and naïve Bayes.

6. Conclusion

In this manuscript, an ensemble-based rough set model is proposed for identifying the key genes of TNBC and non-TNBC. The ensemble-based rough set model includes two key phases of gene selection and classification. After eliminating the outlier's samples, a rough set theory is applied for selecting the optimal TNBC and non-TNBC genes from the 818 samples. The selected optimal TNBC and non-TNBC genes are given as the input to the ensemble classifier (combination of both SVM and kNN classifier) to classify the low-risk genes (non-TNBC) and high-risk genes (TNBC). In the resulting section, the ensemble-based rough set model's effectiveness is validated based on MCC, F-measure, precision, recall, and accuracy. The experimental investigations showed that the ensemble-based rough set model achieved a mean accuracy of 97.24%, which is better compared to other feature selection techniques (i.e., reliefF and infinite), and individual classifiers (i.e., random forest, kNN, and naïve Bayes). The proposed model significantly reduces the computational time and complexity, which are the major issues highlighted in the literature section. As a future direction of work, a new deep learning model can be developed and analyzed on the unstructured multi-modal data to further improve gene classification on other disease for early treatment and diagnosis.

Author's Contributions

Conceptualization, SP, KRB, PBD. Funding acquisition, JF, JN. Investigation and methodology, SP, KRB, PBD. Project administration, JF. Resources, SP, SK. Supervision, PBD. Writing of the original draft, SP, KRB. Writing of the review and editing, PBD, JF. Validation, SP, KRB, PBD, JN. Formal Analysis, PBD, JF, JN. Data curation, SP, KRB, PBD. Visualization, SP, KRB, JF, SK. All the authors have proofread the final version.

Funding

This work was supported by the Ministry of Education, Youth, and Sports (Grant No. SP2022/18, SP2022/34, and SP2022/5) conducted by VSB-Technical University of Ostrava.

Competing Interests

The authors declare that they have no competing interests.

References

[1] C. H. Lee, W. H. Kuo, C. C. Lin, Y. J. Oyang, H. C. Huang, and H. F. Juan, "MicroRNA-regulated protein-protein interaction networks and their functions in breast cancer," *International Journal of Molecular Sciences*, vol. 14, no. 6, pp. 11560-11606, 2013.

[2] E. van den Akker, B. Verbruggen, B. Heijmans, M. Beekman, J. Kok, E. Slagboom, and M. Reinders, "Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis," *Journal of Integrative Bioinformatics*, vol. 8, no. 2, pp. 222-238, 2011.

[3] A. Chakraborty, S. Mitra, D. De, A. J. Pal, F. Ghaemi, A. Ahmadian, and M. Ferrara, "Determining protein–protein interaction using support vector machine: a review," *IEEE Access*, vol. 9, pp. 12473-12490, 2021.

[4] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, and Q. Ma, "Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics*, vol. 35, no. 14, pp. 2395-2402, 2019.

[5] J. P. Sarkar, I. Saha, S. Rakshit, M. Pal, M. Wlasnowolski, A. Sarkar, U. Maulik, and D. Plewczynski, D. "A new evolutionary rough fuzzy integrated machine learning technique for microRNA selection using next-generation sequencing data of breast cancer," in *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, Prague, Czech Republic, 2019, pp. 1846-1854.

[6] Y. B. Wang, Z. H. You, X. Li, T. H. Jiang, X. Chen, X. Zhou, and L. Wang, "Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network," *Molecular BioSystems*, vol. 13, no. 7, pp. 1336-1344, 2017.

[7] J. Zhang, K. Jia, J. Jia, and Y. Qian, "An improved approach to infer protein-protein interaction based on a hierarchical vector space model," *BMC Bioinformatics*, vol. 19, article no. 161, 2018. https://doi.org/10.1186/s12859-018-2152-z

[8] Z. Li, W. Xie, and T. Liu, "Efficient feature selection and classification for microarray data," *PLoS One*, vol. 13, no. 8, article no. e0202167, 2018. https://doi.org/10.1371/journal.pone.0202167

[9] H. Aydadenta and A. Adiwijaya, "A clustering approach for feature selection in microarray data classification using random forest," *Journal of Information Processing Systems*, vol. 14, no. 5, pp. 1167-1175, 2018.

[10] 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, no. 1, pp. 159-176, 2019.

[11] M. Li, Y. Guo, Y. M. Feng, and N. Zhang, "Identification of triple-negative breast cancer genes and a novel high-risk breast cancer prediction model development based on PPI data and support vector machines," *Frontiers in Genetics*, vol. 10, article no. 180, 2019. https://doi.org/10.3389/fgene.2019.00180

[12] Y. D. Cai, Q. Zhang, Y. H. Zhang, L. Chen, and T. Huang, "Identification of genes associated with breast cancer metastasis to bone on a protein–protein interaction network with a shortest path algorithm," *Journal of proteome research*, *16*(2), 1027-1038, 2017.

[13] J. P. Sarkar, I. Saha, A. Sarkar, and U. Maulik, "Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers," *Computers in Biology and Medicine*, vol. 131, article no. 104244, 2021. https://doi.org/10.1016/j.compbiomed.2021.104244

[14] S. Mahapatra, V. R. Gupta, S. S. Sahu, and G. Panda, "Deep neural network and extreme gradient boosting based hybrid classifier for improved prediction of protein-protein interaction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 155-165, 2021.

[15] J. Pan, L. P. Li, C. Q. Yu, Z. H. You, Z. H. Ren, and J. Y. Tang, "FWHT-RF: a novel computational approach to predict plant protein-protein interactions via an ensemble learning method," *Scientific Programming*, vol. 2021, article no. 1607946, 2021. https://doi.org/10.1155/2021/1607946

[16] L. D. Naorem, M. Muthaiyan, and A. Venkatesan, "Integrated network analysis and machine learning approach for the identification of key genes of triple-negative breast cancer," *Journal of Cellular Biochemistry*, vol. 120, no. 4, pp. 6154-6167, 2019.

[17] D. Wang, J. R. Li, Y. H. Zhang, L. Chen, T. Huang, and Y. D. Cai, "Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms," *Genes*, vol. 9, no. 3, article no. 155, 2018. https://doi.org/10.3390/genes9030155

[18] J. Zhang, Y. Suo, M. Liu, and X. Xu, "Identification of genes related to proliferative diabetic retinopathy through RWR algorithm based on protein–protein interaction network," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1864, no. 6, pp. 2369-2375, 2018.

[19] H. Al-Safi, J. Munilla, and J. Rahebi, "Patient privacy in smart cities by blockchain technology and feature selection with Harris Hawks Optimization (HHO) algorithm and machine learning," *Multimedia Tools and Applications*, vol. 81, no. 6, pp. 8719-8743, 2022.

[20] A. F. Iswisi, O. Karan, and J. Rahebi, "Diagnosis of multiple sclerosis disease in brain magnetic resonance imaging based on the Harris Hawks optimization algorithm," *BioMed Research International*, vol. 2021, article no. 3248834, 2021. https://doi.org/10.1155/2021/3248834

[21] H. Al-Safi, J. Munilla, and J. Rahebi, "Harris Hawks Optimization (HHO) algorithm based on artificial neural network for heart disease diagnosis," in *Proceedings of 2021 IEEE International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, Tumkur, India, 2021, pp. 1-5.

[22] A. T. H. Al-Rahlawee and J. Rahebi, "Multilevel thresholding of images with improved Otsu thresholding by black widow optimization algorithm," *Multimedia Tools and Applications*, vol. 80, no. 18, pp. 28217-28243, 2021.

[23] S. Ahmed, M. Frikha, T. D. H. Hussein, and J. Rahebi, "Optimum feature selection with particle swarm optimization to face recognition system using Gabor wavelet transform and deep learning," *BioMed Research International*, vol. 2021, article no. 6621540, 2021. https://doi.org/10.1155/2021/6621540

[25] F. A. Alsarori, H. Kaya, J. Rahebi, D. E. Popescu, and D. J. Hemanth, "Cancer cell detection through histological nuclei images applying the hybrid combination of artificial bee colony and particle swarm optimization algorithms," *International Journal of Computational Intelligence Systems*, vol. 13, no. 1, pp. 1507-1516, 2020.

[24] M. Wozniak, A. Sikora, A. Zielonka, K. Kaur, M. S. Hossain, and M. Shorfuzzaman, "Heuristic optimization of multipulse rectifier for reduced energy consumption," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5515-5526, 2021.

[26] A. Sikora, A. Zielonka, and M. Wozniak, "Heuristic optimization of 18-pulse rectifier system," in *Proceedings of 2021 IEEE Congress on Evolutionary Computation (CEC)*, Kraków, Poland, 2021, pp. 673-680.

[27] W. Dong, M. Wozniak, J. Wu, W. Li, and Z. Bai, "De-noising aggregation of graph neural networks by using principal component analysis," *IEEE Transactions on Industrial Informatics*, 2022. https://doi.org/10.1109/TII.2022.3156658

[28] X. Zhang, C. Mei, D. Chen, and Y. Yang, "A fuzzy rough set-based feature selection method using representative instances," *Knowledge-Based Systems*, vol. 151, pp. 216-229, 2018.

[29] M. S. Raza and U. Qamar, "Feature selection using rough set-based direct dependency calculation by avoiding the positive region," *International Journal of Approximate Reasoning*, vol. 92, pp. 175-197, 2018.

[30] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based k-nearest neighbor classifier," *Expert Systems with Applications*, vol. 115, pp. 356-372, 2019.

[31] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," *European Journal of Operational Research*, vol. 267, no. 2, pp. 687-699, 2018.