Article

JOURNAL OF UNIVERSITY OF BABYLON
For Pure and Applied Sciences (JUBPAS)

Vol.31; No.3.| 2023

Attribute Selection Based on a Hybrid Approach for Improving Classification of Breast Cancer Recurrence

Ameer K. AL-Mashanji [1*]

[1]Presidency of University, University of Babylon, amir.mashanji@uobabylon.edu.iq, Babylon, Iraq.

***Corresponding author email:** ameeruobabylon@gmail.com: 07704319686

اختيار السمات على أساس نهج هجين لتحسين تصنيف تكرار سرطان الثدي

امير علي كاظم*[1]

1 رئاسة الجامعة، جامعة بابل، amir.mashanji@uobabylon.edu.iq، بابل، العراق

Accepted: 25/8/2023 **Published:** 30/9/2023

ABSTRACT

Background:

A commonly occurring disease among women worldwide is breast cancer, the second deadliest form of cancer. However, death chances are remarkably reduced when the cancer is detected and prevented at an early stage.

Materials and Methods:

The main contribution of the current study is to propose a hybrid approach to attribute selection by combining the information gain method with the correlation method and to exploit the strengths of these methods for improving classification accuracy. The dataset has been obtained from the publicly open UCI machine learning repository. The dataset is used to classify the target class into breast cancer recurrence and non-recurrence. Classification algorithms Naïve Bayes, J48 Decision Tree, and Multi-Layer Perceptron were adopted for performing the accuracy of prediction.

Results:

The proposed hybrid approach has been combined with each classification model, improving the performance of each model through the reduction of lower-ranked attributes, due to their insignificant contribution and the possibility of misguiding the classifying algorithm. After selecting a set of upper-ranked attributes carefully, it has been found that the accuracy rate, RMSE, and computational costs have improved for all three algorithms. The J48 Decision Tree achieved a significant performance, and it obtained a relatively higher accuracy (75.87 %).

Conclusions:

It can be concluded that (Inv nodes, deg-malig, node-caps, tumor size, irradiat, and breast) are strong attributes in a dataset and (Age, breast-quad, and menopause) are weak attributes. As noted, the implementation of the hybrid approach improved the accuracy of all classifiers.

Keywords:

Attribute Selection methods, Breast cancer disease, Classification methods, and Performance measures.

ISSN: 2312-8135 | Print ISSN: 1992-0652

info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

1. INTRODUCTION

Cancer is found to be among the largest health issues faced by humanity, breast cancer in particular is found to have the highest death rate among women worldwide, more than any other form of cancer [1]. The only way to prevent the occurrence of this deadly disease is by detecting and predicting it at an early stage. In particular, the prediction of recurring cancer has become a major issue globally [2]. The recurrence of breast cancer involves cancer that later returns to the same or opposite breast or the chest wall, and it cannot always be detected [3]. An essential process in classification is an attribute selection as part of the data preprocessing. Not all attributes are always found to be of relevance during the classification. Some attributes might be redundant and have a negative effect on the classifying algorithm efficiency in terms of time consumption and costs [4]. The hybrid attribute selection approach has been applied as one of the evaluating criteria to diagnose breast cancer. In this study, two types of attribute-selecting method have been used which depend on the ranks. The purpose is the reduction of attributes found in the dataset which are 10 attributes. To that end, it has been proposed a hybrid approach that exploits the strengths of these two methods. This process resulted in the selection of (6) attributes which are input directly into the Naïve Bayes (NB), Multi-Layer Perceptron(MLP), and J48 Decision Tree (DT) classifiers, which produced a better accuracy. The performance measures used for evaluation include: recall, accuracy, and the F1-measure, and its values which have been computed and presented to compare classification results with the original attribute set.

The content of this paper can be outlined in the following way. Section (2) presents a summary of the previous works conducted on diagnosing breast cancer. Section (3) involves breast cancer disease overview. Section (4) presents a description of materials and methods. Section (5) clarifies the methodology. The experimental results analysis is presented in Section (6), and the conclusions are explained in Section (7).

2. RELATED WORK

Bhukya and Sadanandam [5] proposed an approach for breast Cancer Recurrence classification using a series of Machine Learning (ML) classifiers namely, NB, DT, Adaboost, K-Nearest Neighbor (KNN), Logistic Regression (LR),Support Vector Machine (SVM), and Random Forest with a rough set for attribute selection desirable. They used the Wisconsin Breast Cancer Diagnostic (WBCD) dataset to analyze their experiment. Finally, they concluded that the Random Forest gained better results than other models with the highest accuracy of (95.23%). The attribute selection algorithm used significantly improved the accuracy of the classifiers. Kalpna Guleria et al. [6] used four types of ML models which are the KNN, NB, LR, and DT. They used the Breast Cancer dataset which has been obtained through the UCI (University of California Irvine) ML repository for predicting benign or malignant breast cancers. The results showed that the NB provides the highest accuracy among all techniques. Puja Gupta et al. [7] utilized five types of ML techniques, including KNN, Random Forest, SVM, Artificial Neural

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Network (ANN), and DT. They were adopted for predicting the occurrence of breast cancer tumors. The results indicated that the ANN achieved the highest accuracy rates among the other algorithms. The dataset adopted in the research was the WBCD. Ghani et al. [8] applied a multi-classification technique, including the DT, ANN, KNN, and NB algorithms. The results indicated how the ANN provided the best classification accuracy among the other techniques. The authors used the Coimbra breast cancer dataset in their study. On the other hand, Al Batainehin in [9] utilized five ML algorithms, which are the MLP, KNN, CART DT, NB, and SVM. They adopted the WBCD dataset for predicting benign or malignant breast cancers. The accuracy of the MLP on the data was found to be better than the other four algorithms.

3. BREAST-CANCER DISEASE

One of the commonly occurring diseases found among women is breast cancer. It can be described as the event whereby cells in the breast tend to grow abnormally. Diagnosing this medical condition properly in advance is necessary to prescribe suitable medications [10]. With cancer, the growth of organ cells tends to take place uncontrollably. This irrational cell growth forms tumors, which can be either cancerous or non-cancerous. The former type of tumor is life-threatening when spread through the body, whereas non-cancerous tumors are not so life-threatening [11].

4. MATERIALS AND METHODS

4.1 About Breast-Cancer Dataset

The data adopted throughout this research is downloaded freely from the (UCI) ML repository for examining the proposed methodology [12]. The benchmark breast cancer dataset contains (286) instances with (10) attributes of patients who have undergone breast cancer surgeries. Each instance in the dataset has (10) attributes with a class label. The dataset uses to classify the classes breast cancer, non-recurrence, and recurrence. The existing attributes, namely (1)deg-malig, (2)age, (3)node-caps, (4)breast, (5)irradiat, (6)menopause, (7) tumor-size, (8)inv-nodes, (9)breast-quad, and (10)Class Label. Table (1) presents a summary of the dataset information. Table (2) depicts the statistics of classes in the dataset while the distribution of classes is shown in Figure (1). Table (3) shows the description of attributes.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 1:The dataset information

Name of dataset	Breast-Cancer
Number of instances	286
Number of attributes	10
Class variable	no- recurrence, recurrence

Table 2: The statistics of classes in the dataset

Class	Instances	Distribution
No-recurrence	201	70.29 %
Recurrence	85	29.71 %
Total	286	100 %

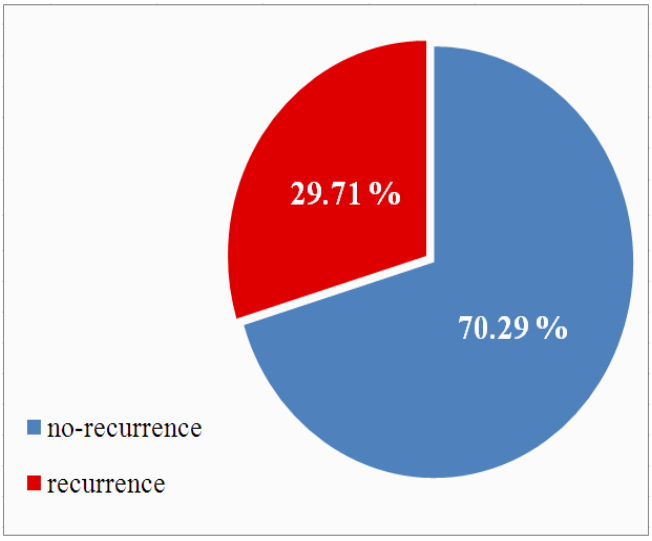


Figure. 1 The classes distribution

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 3: The attributes description

Attribute Name	Description	Type
#1 (deg-malig)	Malignant degree within the patient's body	Numeric
#2 (age)	Patient Age in years	Categorical
#3 (node-caps)	Nodes Presence or Nodes absence around the tumor	Categorical
#4 (breast)	Tumor location in the left or right breast	Categorical
#5 (irradiat)	Tumor non-spread or spread within the patient's body	Categorical
#6 (menopause)	Patient reaches menopause	Categorical
#7 (tumor-size)	Tumor size within the patient's body	Categorical
#8 (inv-nodes)	Node size	Categorical
#9 (breast-quad)	Tumor location in the breast-quad	Categorical
#10 Class Label	recurrence. no- recurrence	Categorical

4.2 Data Preprocessing

The data preprocessing stage is necessary for increasing the data quality in such a way that it will result in high-quality mining [13].

4.2.1 Handling Missing Value

In medical and health-related datasets, it commonly occurs that certain values are missing. Such aspects need to be treated in advance so that they do not fail the classification process or predict diseases incorrectly. Two main techniques are commonly adopted in the treatment of missing values in data sets. These are deleting and imputing values [14]. The first technique is used to deal with missing values which undergo no processing. In the medical context, this process is considered unethical as it may lead to the loss of useful information. Therefore, imputation is a more suitable solution, as it replaces missing values with estimated ones. In this study, the second approach is used. Specifically, the distribution frequency method has been applied whereby a case is imputed using values from the most similar cases [14].

4.2.2 Label Encoder

It is a common coding method for handling categorical attributes. In this method, a unique integer value is assigned for each label based on alphabetical order in the attribute. Then

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

the data is passed to machine learning algorithms because machine learning algorithms do not deal directly with categorical variables [14].

4.3 Ranker Attribute Methods

Selecting attribute is a technique used for optimization through the reduction of data dimension in ML. It involves selecting the best subset of input variables through the removal of attributes that have no predicting information [15]. The attributes in such kinds of methods are selected according to the performance measures with no regard to the predicting algorithms. Thus, they are to be used before the prediction model [16]. In this study, two ranker attribute methods have been applied to evaluate and rank attributes in the breast cancer dataset, which are the correlation and information gain methods.

4.3.1 Correlation Method: This method measures the correlation between all attributes and the target class. The attribute weight ranges between (1) and (-1), so the attribute is considered very weak if its weight is close to zero, meaning that the attribute is not related to the target class, while it is considered very robust if its weight is close to ± 1, meaning that the attribute is highly related to the target class [17]. The following equation computes the correlation value between each attribute with the target class.

$$\mathrm{cor}(x,y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(Xi - \bar{X})^2}\sqrt{\sum(Yi - \bar{Y})^2}} \quad \ldots\ldots\ldots (1)$$

where

Xi: is referring the attribute, **Yi** is referring the target class, $\bar{Y}$: is referring the average of the target class, and $\bar{X}$: is referring to the average of the attribute.

4.3.2 Information Gain: It is one of the important methods that are used to choose attribute and reduce dimensions for effective classification. The attribute is determined as an important attribute if it exceeds a specific threshold. The following equation can be used to calculate IG between any attribute with the target class [18].

$$E(T) = -\sum_{a \in A} P(a)\, log_2\, p(a) \quad \ldots\ldots\ldots (2)$$

$$IG(T.t) = E(T) - E(T\backslash t) \quad \ldots\ldots\ldots (3)$$

where

E(T): is the random variable entropy of T (target class), and **E(T\t):** is the conditional entropy of T given the value of an attribute (t).

4.4 Decision Trees Technique

DT is classifying and predicting method that has a form of a tree flowchart, whereby nodes and internodes are used in representing data. Instances node differing attributes are

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

separated using test cases. Internal nodes result from attribute test cases, and leaf nodes indicate the class variable (target class) [19],[20].The data can be classified using different DT algorithms, such as ID3, C4.5, and J48. All DT nodes are obtained via the calculation of the highest information gain for the attributes. In case a certain attribute leads to unambiguous end products (explicit classification of the class attribute), its branch will be terminated to be assigned the target class value [21]. As for the work presented in this paper, the J48 DT algorithm is used for the model establishment.

4.5 Naïve Bayes Technique

NB is a classifying method used for finding a probabilistic relationship between classes and attributes. It depends on the Bayesian theorem whereby the probability of the target is calculated using a given predictor or attribute value. It has provided satisfactory results in a wide range of applications, especially whenever there is a high input dimensionality [22].

4.6 Multilayer Perceptron Technique

MLP is a supervised learning classifier for feed-forward back-propagation networks. It is the most frequently used technique in classification tasks. It consists of input, output, and more hidden (assign modifiable weighting coefficients to input layers components) layers [8]. MLPs construct a multidimensional space (by the hidden nodes activation), and separate the two classes (no-recurrence, recurrence) as much as possible. Firstly, it passes weights assigned to different layers, determines the output, and compares it with the target output [8].

4.7 Performance measures

Four common performance measures have been adopted in the evaluation of the classification accuracy of algorithms. Figure (2) presents the confusion metrics results of the classifying process (both correct and incorrect results), which are used for measuring the classification quality [23].

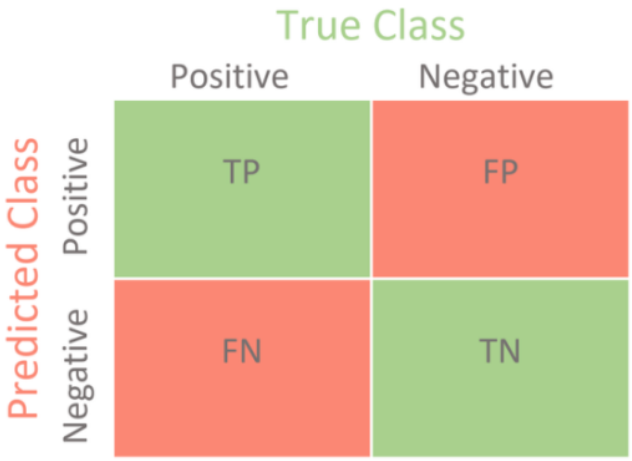


Figure. 2 The confusion metrics

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Whereby

- TP (True Positive): These are the cases that have been assigned a correct positive label by the classifier.
- TN (True Negative): These are the cases that have been assigned a correct negative label by the classifier
- FP (False Positive): These are the cases that have been assigned an incorrect positive label by the classifier.
- FN (False Positive): These are the cases that have been assigned an incorrect negative label by the classifier.

1)Accuracy: Accuracy is the ratio of the number of all correct predictions to the total number of the dataset often calculated using the following equation:

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \quad \ldots\ldots\ldots (4)$$

2) Recall: It is also known as (sensitivity). It represents the rate of predictions that have been identified as positive effectively.

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad \ldots\ldots\ldots (5)$$

3) Precision: It is also known as Confidence. It is represented by the rate of both TP and TN cases that have been identified as being positive. It is an indicator of how efficiently the classifier works.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \quad \ldots\ldots\ldots (6)$$

4) F-Measure: This measure represents the mean of precision and recall, and it considers both false positive and negative results.

$$\text{F} - \text{Measure} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad \ldots\ldots\ldots (7)$$

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

5. METHODOLOGY OF THE STUDY

The architecture of the proposed methodology involves four main stages for achieving the aims of this study, as follows: initially, data preprocessing, a hybrid attribute selection, classification models and finally evaluating classification models based on various measures. Figure (3) shows the block diagram of the methodology stages introduced in this study.

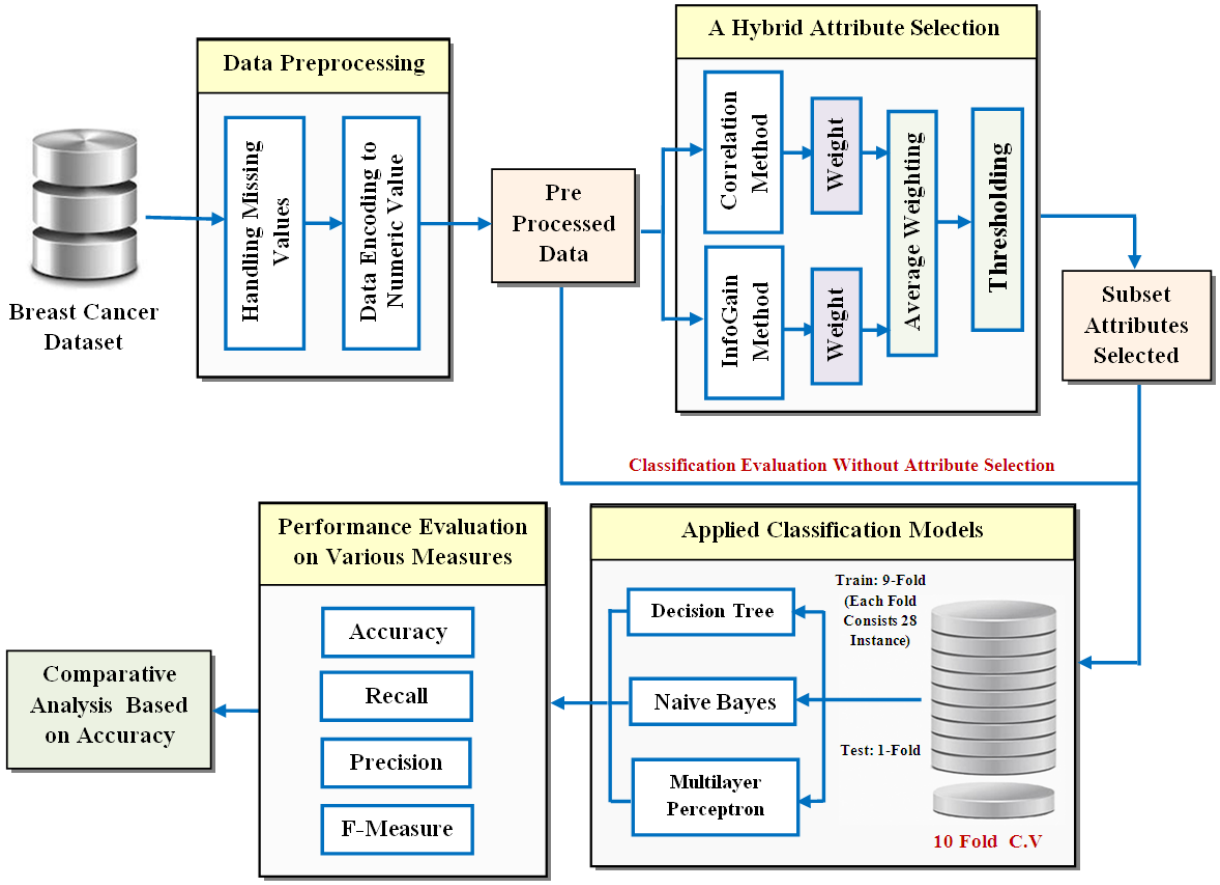
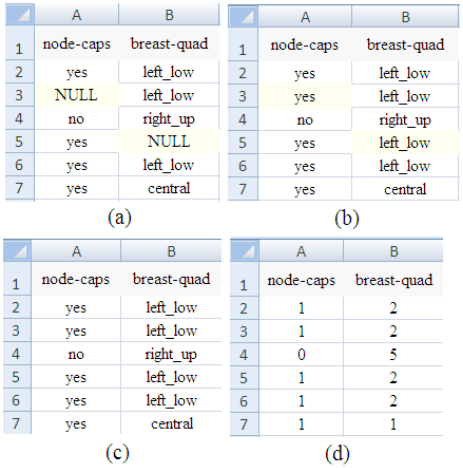


Figure.3 The Architecture of the proposed methodology

Stage 1 Data Preprocessing: Data Preprocessing is intended to prepare the dataset in an appropriate form for the machine learning technique of classification. This stage has been used in this study because the adopted breast cancer dataset contains some missing values. It has been found some missing values in node-caps and breast-quad attributes. These missing values are estimated by replacing them with the most frequent value in the column. After that, Label Encoder is applied to the categorical attributes to map all attributes data as numerical values (see attribute node-caps and breast-quad in Figure (4)).

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

For more details about data preprocessing steps look at the Algorithm (1). Finally, the dataset is ready for further analysis.



	A	B
1	node-caps	breast-quad
2	yes	left_low
3	NULL	left_low
4	no	right_up
5	yes	NULL
6	yes	left_low
7	yes	central

(a)

	A	B
1	node-caps	breast-quad
2	yes	left_low
3	yes	left_low
4	no	right_up
5	yes	left_low
6	yes	left_low
7	yes	central

(b)

	A	B
1	node-caps	breast-quad
2	yes	left_low
3	yes	left_low
4	no	right_up
5	yes	left_low
6	yes	left_low
7	yes	central

(c)

	A	B
1	node-caps	breast-quad
2	1	2
3	1	2
4	0	5
5	1	2
6	1	2
7	1	1

(d)

Figure.4 Data Preprocessing Steps. (a) Before handling missing (b); After handling missing;(c) Before Label Encoder (d) After Label Encoder.

Algorithm (1): Data Preprocessing

Input: Array ($DS_{(n,m)}$) where n is the number of instances, and m is the number of attributes.

Output: Two-dimensional array ($DS_{(n,m)}$) after preprocessing

// Handling missing values

Begin

1. **for each** attribute (T) in the (DS)
2. **if** value (v) in attribute (T) is missing then
3. v = most frequency value in (T)
4. **end if**
5. **end for**

// Label Encoder

6. **for each** attribute (T) in the (DS)
7. **if** an attribute (T) is categorical **then**
8. Apply Label Encoder to attribute (T)
9. **end if**
10. **end for**
11. **return** Updated (DS)

End

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Stage 2 Attribute Selection: Attribute selection methods are applied to identify the most important attributes which directly affect the target class (no-recurrence, recurrence). Thus, a subset of the most important attributes is selected among the original ones. The combining correlation method and information gain method is applied in this study for attribute selection. These methods evaluate the attributes and give a different weight value for each one of them and then the average weight is taken. All weak attributes are excluded separately, this is achieved through a predefined threshold value. Algorithm (2) summarizes the all aforementioned details.

Algorithm (2): A Hybrid Attribute Selection

Input: Array DS(n,m) where n: number of instances, m: number of attributes, a predefined threshold Θ.
// Output of Algorithm (1)

Output: ST[] Significant attributes array

// A Hybrid Correlation and Information Gain Methods

Begin

1. Let COR[] holding attributes weights of Correlation
2. Let IG[] holding attributes weights of Information Gain
3. Let ST[] holding the Significant attributes
3. **for** T = 1 to m
4. Compute the weight (w) of the attribute (T) and target class based on Equation (1) and add value in in COR[].
5. Compute the weight (w) of the attribute (T) and target class based on Equations (2,3) and add value in IG[].
6. W_T=(COR[T]+IG[T]) /2 **// Contribution of Study**
7. **if** (W_T <Θ) **then** **// Θ < 0.3**
8. Ignore this attribute T
9. **else**
10. Add attributes T to ST[]
11. **end if**
12. **end for**
13. **return** ST[]

End

Stage 3 Prediction Stage: This stage represents the most important step in the proposed methodology. Three different classification models have been used for validating the accuracy of the attribute selection, and to ensure that the selected attributes are indeed most likely to influence the target class (no recurrence, recurrence). In the first technique, the method aims to

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

provide a representation of the data in the form of a group of trees through J48 DT. The second technique is Naive Bayes for predicting the relationship between attribute and target class and the third model is MLP.

Stage 4 Evaluation of Prediction Model: In this stage, Accuracy, Recall, Precision, and F1- Measure performance measures are utilized for measuring the performance of classification models.

6. EXPERIMENTAL RESULTS AND DISCUSSION

The method suggested in this study depends on the idea of classification, as the main task is to classify the class labels into recurrence and non-recurrence cases for the breast cancer dataset. Initially, data preprocessing is applied to prepare the row data for future analysis. After that, a hybrid of the correlation method and information gain method is implemented, and taking the average weight for each attribute to determine the optimal attributes subset for improving the accuracy of models.

The correlation and information gain give weight to each attribute based on the relationship between the target class and these attributes. Table (4) shows the average weight in descending order yielded from the combination of the two methods. Figure (5) presents the bar chart of the average weight for each attribute.

Table 4: The average weight for each attribute

Attribute	Correlation	InfoGain	Avg. Weight	Selected
1#deg-malig	0.212	0.077	0.289	✔
2#Irradiat	0.193	0.025	0.219	✔
3#inv-nodes	0.260	0.069	0.164	✔
4#node-caps	0.276	0.051	0.164	✔
5#tumor-size	0.070	0.057	0.063	✔
6#Breast	0.058	0.002	0.030	✔
7#breast-quad	0.050	0.008	0.029	✘
8#Menopause	0.050	0.002	0.026	✘
9#Age	0.0342	0.010	0.022	✘

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

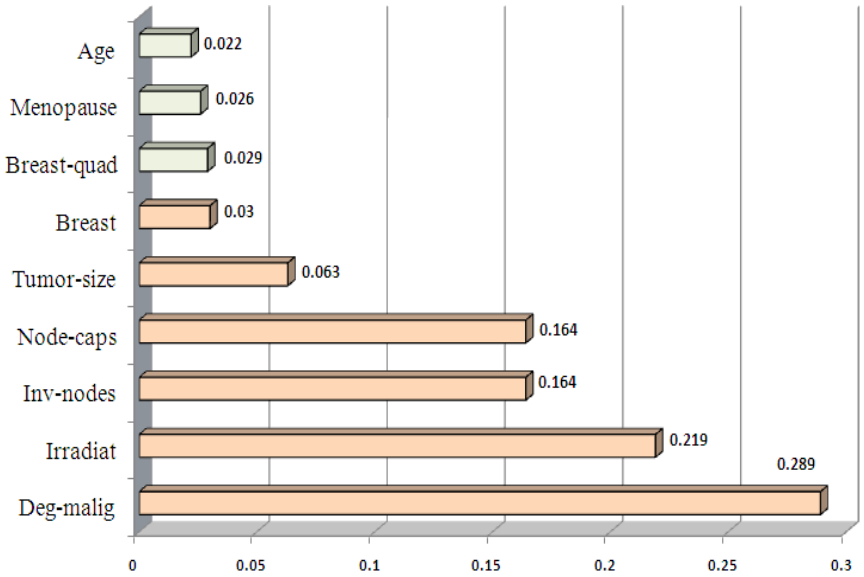


Figure.5 Average weights for all attributes

Any attribute with an average weight less than the threshold value (0.3) has been discarded (see attributes like them menopause, breast quad, and Age in Table 4). After this process, only (6) attributes remain, which represent the most significant attributes that affect the target class. These attributes are phased for building the classification model. In the current study, a K-fold cross-validation has been opted (shown in Figure (6)) that depends on dividing the training dataset into k subsets of equal sizes. In each iteration, one portion is reserved for the validation dataset and the rest of the (k-1) splits are retained as training data.

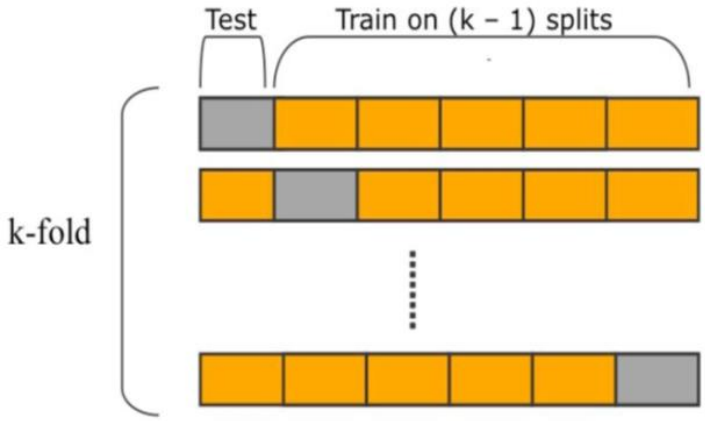


Figure.6 K-fold cross-validation

Several results are obtained by applying classification algorithms, such as NB, MLP, and J48 DT based on a breast cancer dataset with 10 fold cross-validation method. Table (5) presents the accuracy rates of the three classifiers, while Figure (7) presents the bar chart of accuracy values for these classifiers.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 5: Accuracy values

Classifier	All Attributes	With a Hybrid Attribute Selection
NB	71.67 %	74.12 %
MLP	64.68%	69.23 %
J48 DT	74.94 %	75.87 %

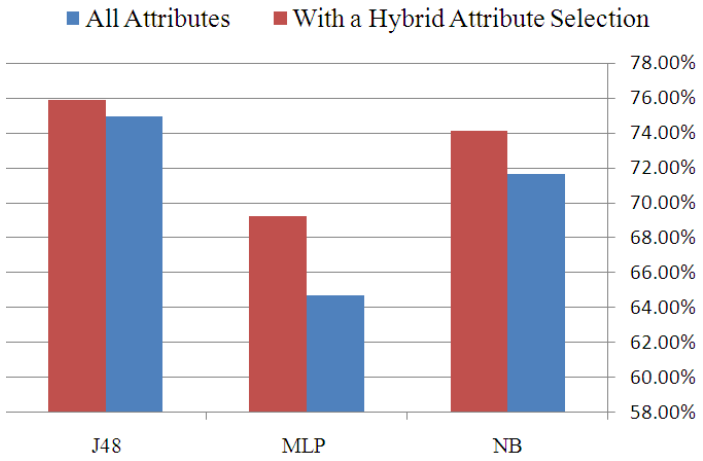


Figure 7. Comparison of classifier's accuracy

Figure (7) indicates that the J48 DT has the highest accuracy with (75.87%) followed by NB with (74.12%), and the last technique is the MLP which has the lowest accuracy (69.23%). The recall values and the bar chart for classification techniques used are shown in Table (6) and Figure (8) respectively.

Table 6: Recall values

Classifier	All Attributes	With a Hybrid Attribute Selection
NB	0.72	0.74
MLP	0.65	0.69
J48 DT	0.76	0.76

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

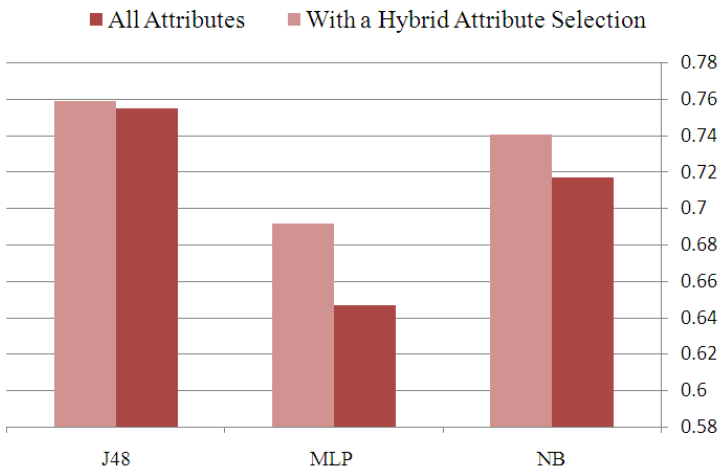


Figure.8 Comparison of classifiers recall

The precision values and the bar chart of all classifiers are presented in Table (7) and Figure (9) respectively.

Table 7: Precision values

Classifier	All Attributes	With a Hybrid Attribute Selection
NB	0.704	0.731
MLP	0.648	0.677
J48 DT	0.752	0.760

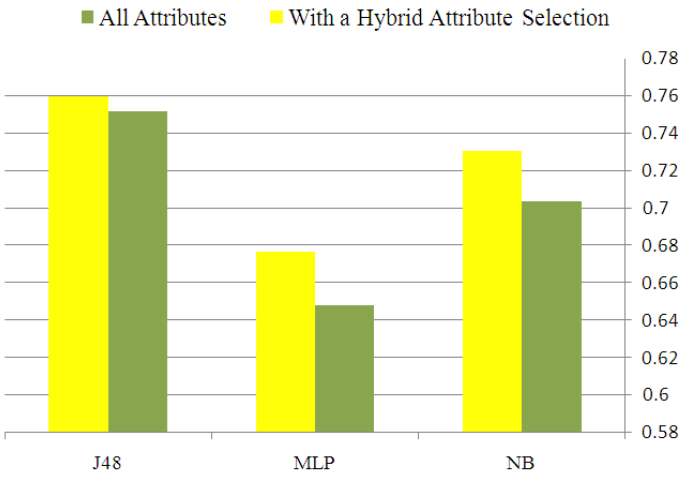


Figure.9 Comparison of classifiers precision

Table (8) shows the f-measure values to the all classifiers, while Figure (10) presents the bar chart of f-measure values.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

Table 8: F-Measure values

Classifier	All Attributes	With a Hybrid Attribute Selection
NB	0.708	0.734
MLP	0.647	0.682
J48	0.713	0.716

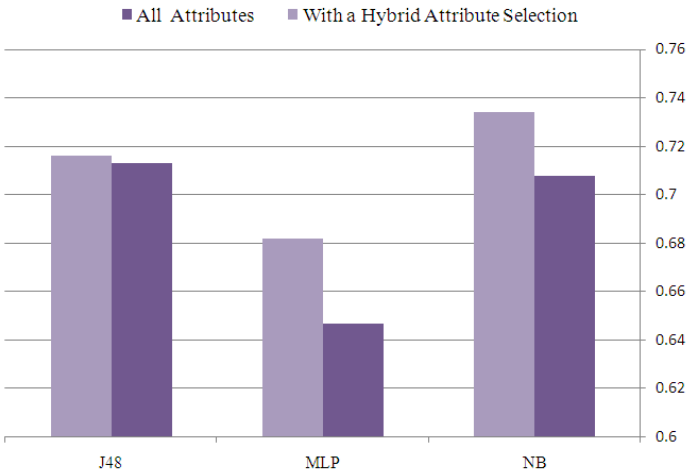


Figure.10 Comparison of classifiers f-measure

The Root Mean Square Error (RMSE) is commonly applied as a parameter to verify experimental results. It refers to the standard deviation of differences between the real (observed) and predicted values. Practically, Figure (11) illustrates that J48 DT has the RMSE with a corresponding value of (0.4324), followed by NB with (0.4534) and MLP with (0.5423) without an attribute selection method.

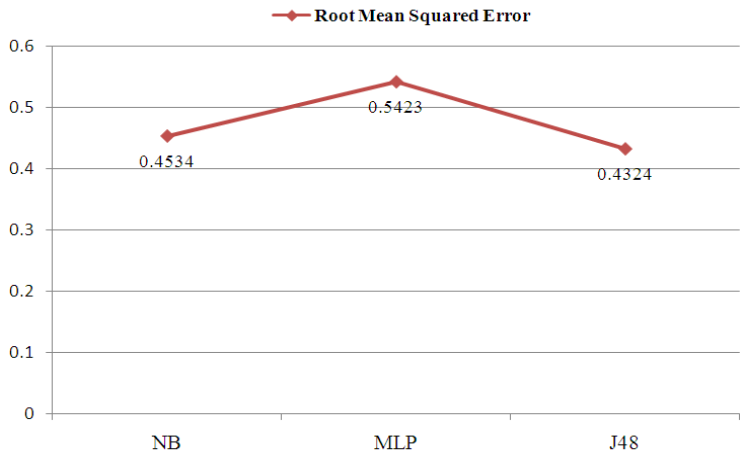


Figure.11 RMSE all attribute selection

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

While, Figure (12) shows that the J48 DT has a very low RMSE value of (0.4314), followed by NB with (0.4487) and MLP with (0.5008) with a hybrid attribute selection.

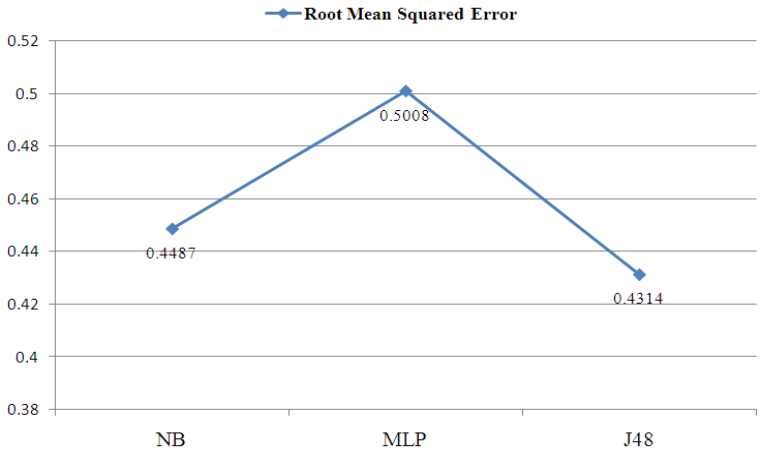


Figure.12 RMSE with a hybrid attribute selection

Figure (13) presents the time required in second to build the models for both with a hybrid attribute selection and with all attributes. It clarifies that J48 DT and NB are faster classifiers than MLP. It is worth mentioning that the time of the MLP is reduced to (1.25) seconds when using the relevant attributes which have been identified using a hybrid attribute selection.

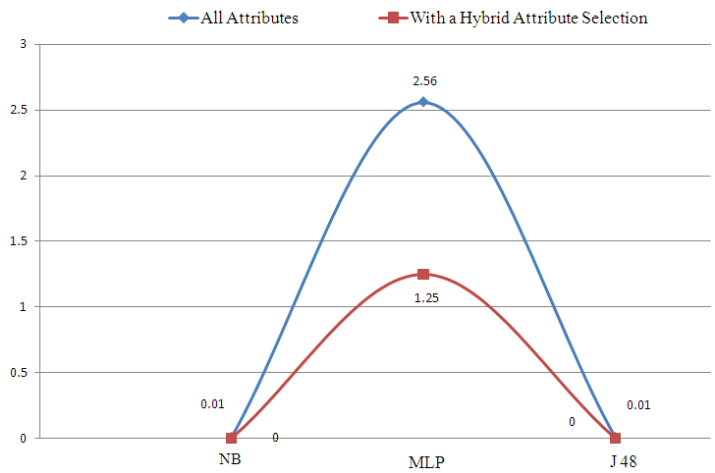


Figure.13 Speed time for all model

It has been found that the proposed hybrid approach contributed to reducing the error and execution time of all classification algorithms, as well as improving classification accuracy. Another additional comparison was done with research results conducted by Alaoui et al. [21] analyzed the same issue. In their work, the authors used a series of six models for classifying breast cancer recurrence namely: Logistic Model Tree, Simple Logistic, NB, Stochastic Gradient

Descent, IBK, and Sequential Minimal Optimization on the same dataset (Breast Cancer). They used all attributes and concluded that the Simple Logistic achieved the best classification accuracy of (74,94%).

As mentioned above, the current study identified only (6) attributes that are highly relevant to the target class with a classification accuracy of (75.87%). The practical experiments for this study were conducted using JAVA programming language version (8.2) in a Net Beans environment with a Windows-64 Operating System.

7. CONCLUSIONS

Attribute selection is a technique used to optimize a certain procedure. It aims to reduce the data dimension in ML processes. Attribute selection involves selecting the best subset of attributes by excluding the ones that have (almost) no prediction information. The present study made use of combining the correlation method and information gain method to select the significant attributes whereby three classifiers are adopted (MLP, J48 DT, and NB). Consistent with the classification results, it has been observed that the J48 DT classifier yields the highest classification accuracy with (75.87%) with (6) predictive attributes. It can be concluded that (inv-nodes , tumor-size , Irradiat, deg-malig, node-caps, and Breast) selected as the best attributes in the dataset. As it has been noted, the implementation of the hybrid approach on the breast cancer dataset led to several beneficial aspects include the reduction of processing time, reducing the number of tests for a patient, and the improvement of prediction quality regarding the (re)occurrence of breast cancer in patients. As for the future direction of research, it is suggested that more attribute selection methods are explored to yield better outcomes. The present work could contribute to increase the efficiency and reliability when predicting diseases. This will in turn help in developing advanced healthcare methodologies.

Conflict of interests.

There are non-conflicts of interest.

References

[1] F. F. Ting, Y. J. Tan, and K. S. Sim, "Convolutional neural network improvement for breast cancer classification", *Expert Systems with Applications*, vol.120, no.5, pp. 103-115,2019. DOI: doi.org/10.1016/j.eswa.2018.11.008

[2] R. Vijayarajeswari, P. Parthasarathy, Vivekanandan, S., and A. A. Basha, "Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform", *Measurement*, vol.146, pp. 800-805, 2019. DOI: doi.org/10.1016/j.measurement.2019.05.083

[3] W. T. Tran,, K., Lu Jerzak, F. I. Klein, J. S. Tabbarah, A. Lagree, and A. Sadeghi-Naini, " Personalized breast cancer treatments using artificial intelligence in radiomics and pathomics", *Journal of Medical Imaging and Radiation Sciences*, vol.50, no.4, pp. S32-S41,2019. DOI: doi.org/10.1016/j.jmir.2019.07.010

[4] A. Chinnaswamy, R. Srinivasan, and S. M. Poolakkaparambil, "Rough set based variable tolerance

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

attribute selection on high-dimensional microarray imbalanced data", *Data-Enabled Discovery and Applications*, vol.2 ,no.1, pp.1-16, 2018. DOI: doi.org/10.1007/s41688-018-0019-0

[5] H. Bhukya, and M. Sadanandam, "RoughSet based Feature Selection for Prediction of Breast Cancer", *Wireless Personal Communications*, vol.130, no.3, pp.2197-2214, 2023. DOI: doi.org/10.21203/rs.3.rs-1542645/v1

[6] K. Guleria, A. Sharma, U. K. Lilhore, and D. Prasad, "Breast Cancer Prediction and Classification Using Supervised Learning Techniques", *Journal of Computational and Theoretical Nanoscience*, vol.17, no. 6, pp.2519-2522, 2020. DOI: 10.1166/jctn.2020.8924 2

[7] P. Gupta, and S. Garg, " Breast cancer prediction using varying parameters of machine learning models", *Procedia Computer Science*, vol.171, pp. 593-601, 2020. DOI: doi.org/10.1016/j.procs.2020.04.064

[8] Ghani, M. U, T. M. Alam, and F. H. Jaskani, "Comparison of classification models for early prediction of breast cancer", *In 2019 IEEE International Conference on Innovative Computing (ICIC) ,*2019, pp. 1-6. DOI: 10.1109/ICIC48496.2019.8966691

[9] A. Al Bataineh , "A comparative analysis of nonlinear machine learning algorithms for breast cancer detection", *International Journal of Machine Learning and Computing*, vol.9, no. 3, pp.248-254, 2019. DOI: 10.18178/ijmlc.2019.9.3.794

[10] M. Gupta, and B. Gupta , "A comparative study of breast cancer diagnosis using supervised machine learning techniques ", *In 2018 IEEE second international conference on computing methodologies and communication (ICCMC)* , 2018 , pp. 997-1002. DOI: 10.1109/ICCMC.2018.8487537

[11]M. Ezzat, and A. Idri, "Reviewing Data Analytics Techniques in Breast Cancer Treatment", *In World Conference on Information Systems and Technologies* , 2020, pp. 65-75 ,Springer, Cham.

[12]UCI ,Web source:http://archive.ics.uci.edu/ml/datasets.html,last accessed on Jan 2021.

[13]S. Tang, S. Yuan,, and Y. Zhu, " Data preprocessing techniques in convolutional neural network based on fault diagnosis towards rotating machinery ", *IEEE Access*, vol.8, pp.149487-149496, 2020. DOI: 10.1109/ACCESS.2020.3012182

[14]M. Kakkar, S. Jain, A. Bansal, and P. S. Grover, " Combining data preprocessing methods with imputation techniques for software defect prediction ", *In Research Anthology on Recent Trends, Tools, and Implications of Computer Programming*, pp. 1792-1811. IGI Global, 2021. DOI: 10.4018/IJOSSP.2018010101

[15]A. Jović, K. Brkić, and N. Bogunović, " A review of attribute selection methods with applications", *In 2015 Ieee 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, 2015, pp. 1200-1205. DOI: 10.1109/MIPRO.2015.7160458

[16]S. H. Ebenuwa, M. S. Sharif, M. Alazab, and A. Al-Nemrat , " Variance ranking attributes selection techniques for binary classification problem in imbalance data", *IEEE Access*, vol.7, pp.24649-24666, 2019. DOI: 10.1109/ACCESS.2019.2899578

[17]S. Singla, P. Ghosh, and U. Kumari, " Breast cancer detection using genetic algorithm with correlation based feature selection: experiment on different datasets", *International Journal of Computer Sciences and Engineering*, vol.7, no.4, pp. 406-410, 2019. .DOI: doi.org/10.26438/ijcse/v7i4.406410

[18]M. W. Huang, C. H. Chiu, C. F. Tsai, and W. C. Lin, " On combining feature selection and over-

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

sampling techniques for breast cancer prediction", *Applied Sciences*, vol.11, no.14, pp.6574, 2021. DOI: doi.org/10.3390/app11146574

[19]M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad, and C. T. Lin, "Machine learning techniques for the diagnosis of Alzheimer's disease: A review", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),* vol.16, pp.1-35,2020. DOI: doi.org/10.1145/3344998

[20]N. Nahar, and F. Ara,"Liver disease prediction by using different decision tree techniques", *International Journal of Data Mining & Knowledge Management Process*, vol.8, no.2, pp.01-09, 2018. DOI: 10.5121/ijdkp.2018.8201

[21]S. S. Alaoui, Y. Labsiv, and B. Aksasse,"Classification algorithms in data mining", *Int. J. Tomogr. Simul*, vol. 31, pp.34-44, 2018.

[22]M. U. Ghani, T. M. Alam, and F. H. Jaskani,"Comparison of classification models for early prediction of breast cancer", I*n 2019 IEEE International Conference on Innovative Computing (ICIC)*, 2019, pp. 1-6. DOI: 10.1109/ICIC48496.2019.8966691

[23]G. Huang, Y. Li, Q. Wang, J. Ren, Y. Cheng, and X. Zhao," Automatic classification method for software vulnerability based on deep neural network", *IEEE Access*, 7, 28291-28298, 2019. DOI: 10.1109/ACCESS.2019.2900462

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com

الخلاصة

مقدمة:

يعد سرطان الثدي من الأمراض الشائعة الحدوث بين النساء في جميع أنحاء العالم، وهو ثاني أكثر أنواع السرطان فتكًا. ومع ذلك، تقل فرص الوفاة بشكل ملحوظ عند اكتشاف السرطان والوقاية منه في مرحلة مبكرة.

طرق العمل:

تتمثل المساهمة الرئيسية للعمل الحالي في اقتراح نهج هجين لاختيار السمة من خلال الجمع بين طريقة اكتساب المعلومات وطريقة الارتباط ، واستغلال نقاط القوة في هذه الطرق لتحسين دقة التصنيف. تم الحصول على مجموعة البيانات من مستودع التعلم الآلي UCI المفتوح للجمهور. تُستخدم مجموعة البيانات لتصنيف الفئة المستهدفة إلى تكرار الإصابة بسرطان الثدي وعدم تكرارها. تم اعتماد خوارزميات التصنيف Naïve Bayes و J48 Decision Tree و Multi-Layer Perceptron لحساب دقة التنبؤ.

الاستنتاجات:

يمكن استنتاج أن (deg-malig, Inv nodes, node-caps, irradiat, tumor size و breast) هي سمات قوية في مجموعة البيانات و (breast-quad, Age وbreast-quad) هي سمات ضعيفة. كما لوحظ ، أدى تنفيذ النهج الهجين إلى تحسين دقة جميع المصنفات.

الكلمات المفتاحية:

طرق اختيار السمات، مرض سرطان الثدي، طرق التصنيف، ومقاييس الأداء.

ISSN: 2312-8135 | Print ISSN: 1992-0652
info@journalofbabylon.com | jub@itnet.uobabylon.edu.iq | www.journalofbabylon.com