Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 4, November 2020, pp.2146-2159

Application and comparison of different classification methods based on symptom analysis with traditional classification technique for breast cancer diagnosis

F. S. Al-Juboori [1,2], N. P. Alexeyeva [1]
[1] Department of Mathematics, St. Petersburg State University, 28, University Avenue, 198504, St. Petersburg, Russia
[2] University of Information Technology and Communications, St Al-Nidal, Baghdad, Iraq

ABSTRACT

Novel approach for classification technique such as Artificial Neural Network (ANN), Linear Discriminant Analysis (LDA) and Random Forest (RF) using factor or dichotomic variables has been introduced. This study searches for the highly informative finitely linear combinations (symptoms) of variables in the finite field on the based of the Fisher's exact test and accurately predict the target class for each case in the data. There are several super symptoms have comparable $p - values$. In this case, it becomes possible to choose as a nominative representative the factor which is more accessible for interpretation. The super symptom means a linear combination of various multiplications of k dichotomous variables over a field of characteristic 2 without repeating. In algebra, such functions are called Zhegalkin polynomials or algebraic normal forms.

This process essentially yields the new variable of the identical nature or factor. The purpose of this study is to suggest a classifiers in accordance with symptoms analysis, and accurately predict the class in the dataset by using different algorithms. The proposed method for super symptoms with the most famous classification methods was compared to traditional classification methods. The performance of the classifiers has investigated based on breast cancer data set for training algorithm. Moreover, these three different algorithms have been studied very well based on symptom analysis and thus we do focus on the fact that the best results are from which algorithm.

Keywords: symptom analysis, artificial neural network, linear discriminant analysis , random forest, breast cancer, accuracy.

Corresponding Author:
F. S. Al-Juboori
Department of Mathematics, St. Petersburg State University
St. Petersburg, Russia
Email: fatema_sadik79@yahoo.com

1. Introduction

Medical data sets have high-level information content, it appears only when the data is converted to specific algorithms. Breast cancer stands for the second foremost cause of cancer deaths in women nowadays. The most in effect method for reducing breast cancer deaths is discovering it previously. Thus finding effective diagnosis method and an accurate is very important because we can discover breast cancer in early stage and the same time allow doctors to discriminate benign breast tumors from malignant ones. Neural networks were extensively employed for breast cancer identification [16][17]. Feed forward neural networks (FFNNs) are commonly used for classification. Feed forward neural networks have been trained with standard back propagation algorithm [18]. They represent supervised networks and necessitate a preferred response to be trained. The objective is to use the information it possesses in the analysis of diagnostic errors. They can be similar to the performance of optimal statistical classifiers in the challenging problems. In this area, numerous researches has been implemented. The problem of breast cancer diagnosis has involved many scholars in computational intelligence and statistical fields.

In 2008 the researcher (Maciej A.Mazurowskia and his companions) are researching training, the classification of the neural network for medical decision creating with an effect of unbalanced data sets on the classification performance, where clinical data were diagnosed for breast cancer patients. The consequences indicate that classifier performance declines with even modest class imbalance in the training data.

Additionally, it is shown that back propagation (BP) is usually desirable over particle swarm optimization (PSO) for imbalanced training data especially with minor data sample and huge amount of features [12].
In 2009 the researcher (N.Alexeyeva and her companions) used the approaches and procedures for dichotomic variable structuring in accordance with finitely geometries, and their automorphism collections turn out to be more urgent. The foremost notion of this technique is to realize new variables that can be employed in place of base ones. New variables have redistributed information structure. Accordingly, the factor variable itself is feasibly considered as the constituent of some Galois field. Consequently it is potential to investigate finitely linear combinations (symptom) of factors in this field . This notion can be employed for the dichotomic variables and field F_2 [6].
In 2013 the author (N.Alexeyeva) proposed a symptom syndrome method for analysis of multidimensional binary data [3].
In 2009 and 2010 the researcher (N.Alexeyeva) used application example in symptom analysis and symptom syndrome analysis of categorical series, and the study of the combinatorial structure of binary signs based on finite geometries, linear combinations over a finite field are used to provide a symptom-syndrome approach to solving clinical diagnostic problems [4] [5].
In 2013, (Filippo Amato) researched artificial neural networks in medical diagnostics (where discussed the problem of simplifying the diagnostic process in the daily routine and avoiding the error of diagnosis. Artificial intelligence used special methods in assisted diagnosis computer and artificial neural networks and adaptive learning algorithms can deal with different types of medical data are incorporated into their categorized outputs and there are examples of how to use artificial capabilities of neural networks in medical diagnosis. It was through this conclusion that artificial neural networks (ANN) possess the capability for processing huge quantities of data diseases and increasing diagnostic accuracy and increasing patient satisfaction [13].
In 2017, the researcher (Jasmina D. Novakovic) search solve classification problems radial basis function (RBF) and filter methods discuss the problem of avoiding diagnostic problems, by using machine learning that diagnoses tumors, heart disease, hepatitis and some medical problems used in artificial neural networks aim to exhibit and compare diverse algorithms for the constructing system that learns from experience and creates decisions and predictions and decrease the probable amount or errors percentage. It was through this conclusion that techniques should be used to solve dimension reduction problems data such as clustring methods, extraction of features, analysis and comparison of these techniques can also improving the performance of algorithms and classifying learning [14].
In (2012) compared diverse classification performances for evaluating the accuracy between three dissimilar breast cancer datasets in which confusion matrix using 10-fold cross validation method has considered [23].
In (2012) machine learning algorithms are effective because their process of searching for a model function can explain and differentiate the class and concept data, which the model is determined based on the data training analysis that is class object data whose label class is already known. The kinds of learning algorithm are Linear Discriminant Analysis, Super Vector Machine ,Logistic Regression, Naive Bayes, Neural Network, Decision Tree and K-Nearest Neighbor [24].
In (2020) the researcher (N.Alexeyeva) used application example in symptom analysis of multidimensional categorical data with applications [26].
As a result of the tremendous developments witnessed by the world in various fields and activities, especially in information technology, it has become logical and even necessary to use the approaches of information systems in the medical fields. The use of ANN in medical fields was the best evidence of the introduction of information technology in health services. These networks can after training find the pathological mark by entering it and then you will be able to find the complete stored sample which represent the best diagnosis of the pathological case[15].

2 The symptom analysis

The typical approach to handle factor variables with dual levels is to consider them as dichotomic to investigate their linear combinations in the real field.
For instance, suppose variable X_1 be equivalent to 1 in the case of object hight of interest is large and 0 if not; variable X_2 be equivalent to 1 if wight is big and 0 if not. In that case, dichotomic variable $X_{12} = X_1 + X_2(mod2)$ characterizes the inadequacy betwen weight and hight. Variable X_{12} derived in this method can be in higher informativeness for a statistical investigation[6]. Thus in order to see how super symptom can be used in breast cancer data with NN, first of all it is need to defined symptom and super symptom.

2.1. Symptom and super symptom

Consider random vector $X = (X_1, \dots, X_m)^{\mathrm{T}}$ with components taking values 0 and 1. Usually 0 and 1 mean lack and presence of factors respectively. The new variable $X_i + X_j(\mathrm{mod}2))$ [1] means presence of any one in the absence of another factor. We consider also more than two variables.

Definition 1 Let be $\tau = (t_1, \dots, t_k) \subseteq (1,2, \dots, m)$. Then $X_\tau = \sum_{i=1}^{k} X_{t_i}(\mathrm{mod}2)$ is called the symptom X_τ of the rank k. For example $\tau = (1,3,4)$, then $X_\tau = X_1 + X_3 + X_4(\mathrm{mod}2)$. The components of the vector X are trivial symptoms of rank 1. We define symptom of rank zero degenerate, it takes a value of 0 with a probability of 1.

Definition 2 Let be $\tau = (t_1, \dots, t_k) \subseteq (1,2, \dots, m), k \leq m$. Denote the result of multiplication of several dichotomous variables $X_{t_1}, \dots, X_{t_k}$ by $X^\tau = X_{t_1} \cdot \dots \cdot X_{t_k}$. If consider some $\tau_1, \dots, \tau_L \subseteq (1,2, \dots, m)$, where $\tau_i \neq \tau_j \; \forall i \neq j$, then $\sum_{i=1}^{L} X^{\tau_i}(\mathrm{mod}2)$ is called the super symptom.

For example, $S_1 = X_1 + X_2 + X_1X_2(\mathrm{mod}2)$ and $S_2 = X_1X_2 + X_1X_3 + X_2X_3(\mathrm{mod}2)$ are super symptoms. The first means presence X_1 or X_2 or both together when S = 1, which corresponds to the logical sum. When $S_1 = 0$ then $X_1 = 0$ and $X_2 = 0$ at the same time., the second means presence not less two out three factors X_1, X_2, X_3 at $S_2 = 1$. When $S_2 = 0$ then all X_1, X_2, X_3 are equal 0 or $X_i = 1$ separately, i.e the vector $X = (X_1, X_2, X_3)$ takes values $(0,0,0), (1,0,0), (0,1,0), (0,0,1)$.

Alternatively the symptoms and the super symptoms can be defined through a special parameterization of dichotomous vectors $(\alpha_1, \dots, \alpha_m)$ and $(\beta_1, \dots, \beta_M)$, where $M = 2^m - 1$, $\alpha_k, \beta_i \in \{0,1\}$, all components of vector are non-zero at the same time: $\sum_{k=1}^{m} \alpha_k \neq 0$, $\sum_{i=1}^{M} \beta_i \neq 0$. We introduce a special parameterization $a = \sum_{k=1}^{m} \alpha_k 2^{k-1}$ and $b = \sum_{i=1}^{M} \beta_i 2^{i-1}$. Then the symptoms and the super symptoms are defined respectively as linear combinations and polynomials kind of

$$G_a(X_1, \dots, X_m) = \sum_{k=1}^{m} \alpha_k X_k(\mathrm{mod}2), \; a = 1, \dots, M = 2^m - 1,$$
$$F_b(X_1, \dots, X_m) = \sum_{a=1}^{M} \beta_a \prod_{k=1}^{m} X_k^{\alpha_k}(\mathrm{mod}2), \; where \;\; b = 1,2, \dots, 2^M - 1 \,. \qquad (1)$$

3. Classification

Classification represents the process of classifying to which set of groupings for a new observation fits in. The goal of classification is to envisage the outcome y using collection symptom $(X_0, \dots, X_k)$ in the data fitting problem. In the classification problem, the dependent variable y takes on only a finite number of values, but in statistic dependent variable y called a categorical and sometimes called label. The algorithm that applies classification is identified as a classifier. The classifier term denotes the mathematical function that the classification procedure carries out.

3.1. Classifier

A classifier predicts categorical label (classes). Boolean classifiers have been used widely in many application fields. For that reason, we will start by looking at the Boolean classification problem, for example (Disease Detection), when building a model, the data used comes from hospital records or medical studies and the result that we obtain is the diagnosis accompanying the existence or nonexistence of the disease and confirm the doctor. Therefore, the examples are compatible with patients in general, especially cancer patients, with a result of $y = 1$ this means that the patient suffers from cancer or any specific disease, and $y = -1$ means no. Furthermore, the vector x contains relevant medical features(variables) associated with the patient, including for instance age, sex, results of tests, and specific symptoms. We denote by $X = (X_1, \dots, X_p)$ the predictors, by $y = -1,1$, of a type of class. For the particular classifiers consider a random vector $\tilde{X}(\tau) = (X_{\tau_1}, \dots, X_{\tau_k})$, where $\tau = (\tau_1, \dots, \tau_k) \subset (1,2, \dots, p)$.

As in real-valued data fitting, we suppose that an approximate relationship of the form $y \approx f(x)$ holds, where $f: R^n \longrightarrow \{-1,1\}$. This notation means that the function f takes super symptoms argument, and gives a resulting value that is either 1 or -1. The model will have the form $\hat{y} = \hat{f}(x)$, where $\hat{f}: R^n \longrightarrow \{-1,1\}$. The model $\hat{f}$ is also called a classifier, since it classifies super symptoms into those for which $\hat{f}(x) = 1$ and those for which$\hat{f}(x) = -1$. As in real-valued data fitting, we choose or construct the classifier $\hat{f}$ by using some observed data[25].

[1] The expression $a(\mathrm{mod}2)$ means the remainder of a division of the number a by 2. This corresponds to that all operations are performed over the field $\mathbb{F}_2$.

3.2. Prediction errors

For a given data of super symptoms with predicted outcome $\hat{y} = \hat{f}(x)$, there are only four possibilities:

- True positive. $y = +1$ and $\hat{y} = +1$.
- True negative. $y = -1$ and $\hat{y} = -1$.
- False positive. $y = -1$ and $\hat{y} = +1$.
- False negative. $y = +1$ and $\hat{y} = -1$.

In some applications we interest equally about making the two types of errors; in others we may interest more about making one type of error than another. There are only four cases of probability. In the first two cases the predicted categorical is correct, and in the last two cases, the predicted categorical is an error. In addition to that, we indicate to the third case as a false positive or type I error, and we indicate to the fourth case as a false negative or type II error.

3.3. Least squares classifier

To construct a Boolean model or classifier from a set of data, several complex methods have been developed. So we will discuss a very simple method based on the least squares that can work well, but not in the most complex methods. Initially, we implement ordinary real-valued least squares fitting of the outcome, after that we choose basis functions $f_1, \ldots, f_p$, and then choose the parameters $\theta_1, \ldots, \theta_p$ in order to minimize the sum squared error.

$$(y^{(1)} - (\tilde{f}(x^{(1)}))^2 + \ldots + (y^{(N)} - \tilde{f}(x^{(N)}))^2,$$

where $\tilde{f}(x) = \theta_1 f_1(x) + \ldots + \theta_p f_p(x)$. We use the notation $\tilde{f}$, since this function is not our final model $\hat{f}$. The function $\hat{f}$ is the least squares fit over our data set, and $\tilde{f}(x)$, for a general vector x, is a number. Our final classifier is then taken to be

$$\hat{f}(x) = sign(\tilde{f}(x)), \quad (2)$$

where $sign(a) = +1$ for $a \geqslant 0$ and -1 for $a < 0$. We call this classifier the least squares classifier. Therefore, When the value of $\tilde{f}(x)$ is close 1 we have confidence in our guess $\hat{y} = 1$; when it is small and minus (say, $\tilde{f}(x) = -0.03$), we guess $\hat{y} = -1$, but the confidence in the guess will be low.

By our interpretation of the Least Squares Classifier, we can easily explain the coefficients in the following models, such as Artificial Neural Network, Linear Discriminant Analysis and Random Forest. But each model differs in terms of the formula Least Squares Classifier and how to calculate Prediction errors and compare models through Experimentation.

4. Methods

Before starting to apply classification methods, in statistical analysis variable selection or variable subset selection methods is an active research area. This methods could be used to reduce the amount of features in the study dataset. The reasons to use variable selection are: Its enables the algorithm to train quicker and enhances the model accuracy if a right subset is selected. Therefore, the objective standing behind variable subset selection is to choose the subset of variables by ignoring features with less important information. Three methods of classifiers are compared in this research: Artificial Neural Network, Linear Discriminant Analysis and Random Forest.

4.1. Artificial neural network

Artificial Neural Networks (ANN) stands for highly significant fields of artificial intelligence that have a close connection and an active role in many of the various artificial intelligence applications. It processes information in a way that simulates the human mind. The human brain has billions of neural cells that deal with information. Every neural cell represents a tiny processing system. The unified web of neurons is considered as organic neural network that transfers information by electrical signals [1]. Warren McCulloch and Walter Pitts, In 1943, initiated the foremost mathematical model of a neuron. In their reorted study "A logical calculus of the ideas immanent in nervous activity", they explained the simple mathematical model for a neuron that characterizes a particular cell of the neural system that has inputs, processes those inputs, and yields an output. This model is termed as the McCulloch-Pitts neural model [2].

The neural network consists of a complex group of several elements processing the node (neuron) is based on mathematical models for data processing of three layers, these are: Input, Hidden, and output each have

different weights. The neural network is trained to produce the best results through the approved weight control process. (Figure 1) illustrates artificial neuron components.

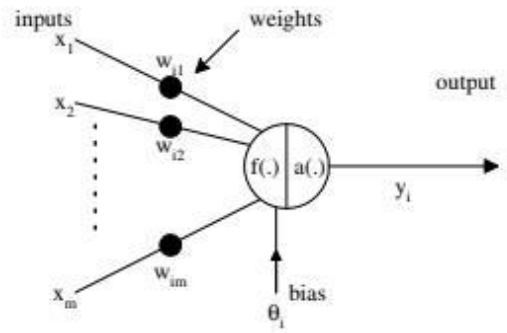

Figure 1. Components of the artificial neural network

In this article, we can use the symptom-syndrome models [3] in which the predicate is expressed in terms of independent factors as linear combinations over field F_2, which form the finite projective space [4].
We used a network based on 11 different covariates and super symptom . In multidimensional data analysis, when individual factors are insignificant, it is possible to detect a risk group with a special combination of factors. Experience has indicated there has been no definite way to forecast appropriately for each case, but that in each case, the way for a private predict along with the requirement to find and use. Nevertheless, taking more than one way can cause raising the future accuracy of estimates.
The present study aims to shed light on the neural network method based on the different variables and compare it with the neural network which depends on the method of Super Symptom Analysis [5][6].
The quality of future predictions of a particular phenomenon that can be obtained from the neural network depends on the efficiency of neural network training which is based on several factors: Learning rate, momentum factor, the number of vectors in the neural network, the number of hidden nodes, plus the number of hidden levels.

4.1.1. Models of artificial neural networks and methods of education and programming:

Back Propagation: The back propagation model stands for highly common emloyed models in pattern recognition (such as speech recognition, medical diagnosis, etc.) and consists of multiple artificial neural network. Layers are directed and iterative until the network performance reaches the required level[9].

4.1.2. Neural network algorithm related back propagation

One of the most commonly used training algorithms is used in neural network training fully correlated with basic feed, multi-layered and non-linear. This is a generalization of the algorithm method of training error correction pattern.

This algorithm is performed in two main stages:
Stage 1: Forward Propagation.
Stage II: Back Propagation.

• Forward Propagation. This stage begins display form the entrance to the network, where each element of the input layer is allocated to one of the components of the ray representing the input, the values of the input vector components cause the input layer units to be activated, ie the network works with the front feeder system. No correlation weights are adjusted during this stage.

• Back Propagation. It is a stage of adjusting weights where the network outputs during training compare to a set of correct forms that are fed from the outside and calculate the difference between the two and use the inverse to adjust the weights. The signal allows the propagation from the output layer to the input layer in reverse during the stage of updating or adjusting the weights, this process is repetitive till the output of the network corresponds to the correct shapes given. The back error propagation network algorithm is one of the most important algorithms of supervised training networks, its name is derived from the fact that the resulting error is inversely caused by the network from one layer to another.

The back error propagation network algorithm depends on the selection of an appropriate error function whose values are determined by the actual results (actual) and the values to be obtained (desired). It also depends on the parameters of the network, such as weights and thresholds θ_j. The actual output of the treatment unit in the hidden layer is determined by Eq.(2). A

$$Y_j(t) = sigmoid[\sum_{i=1}^{n} X_i(t)W_{ij}(t) - \theta_j] \quad (3)$$

where $X = (X_1, X_2, \ldots . X_m)$ represent the minput applied to the neouron, w_{ij} represent the weights between input and hidden layer, θ_j represent the bias of value, n represented the number of inputs for neuron j In the hidden layer, t represent the number of repetition. Sigmoid is an exponential activation function, the Sigmoid function is given by the Eq(3).

$$f(X) = \frac{1}{(1+\exp^{-x})} \quad (4)$$

Thus, this function has been appeared in the output layers of the deep learning DL architectures, and they are employed for forecasting probability based output and has been employed efficaciously in binary classification problems.

The actual outputs of the processing units in the output layer are determined by Eq.(4).

$$Y_k(t) = sigmoid[\sum_{i=1}^{m} X_i(t)W_{jk}(t) - \theta_k] \quad (5)$$

Where m represent the number of inputs for neuron k in the output layer, w_{jk} represent the weight between hidden node$j$ and the input node k.

To achieve the selected error criterion, the frequency t is increased by one correct and return to the second step, one of the most commonly used criteria is the miniaturization of the sum of error square, which states that when the sum of error square becomes a full path during all training categories or the period is sufficiently small, the network is considered convergent (converged) [11].

4.2. Discriminate analysis

Linear discriminate analysis is a technique developed by Roland Fisher. It can also be called Fisher Discriminate Analysis. This technique used for supervised classification problems. (LDA) is a simple and effective method for classification. The main goal of this method is to separate samples of distinct groups. The basic idea of (LDA) is to construct decisional contours for separating the objective of the classes by means of optimization of the error criterion. Thus, it transforms data to a different space which optimally distinguish classes which can be referred to between class and with in classs. (LDA) has been as well associated with factor analysis and principal component analysis (PCA) in that they both look for linear combinations of variables that clarify the data in the best manner [19].

4.3. Random forest

Random forests reresent a grouping of tree predictors in which every tree relies on the magnitudes of a random vector sampled self-reliantly and with the identical distribution for all trees in the forest. Random forest (RF) stand for an ensemble learning technique for classification that works by constructing a multitude of decision trees at training time and generate the class based on the method of classes (classification) of the individual trees [20]. The introduction of proper random forests was firstly reported in [21]. This study refers to a technique of building a forest of uncorrelated trees by means of a classification and regression tree(CART) in term of procedure, combined with randomized node optimization and bagging. (RF) offers an improvement over bagged trees through small tweak that decorates the trees. Moreover, the generalization error of a forest of tree classifiers is dependent on the individual trees strength in the forest along with the correlation among them.

5. Application to real data

In this section, a real scenario is presented where the methodology is applied to the prediction of target class in abreast cancer dataset. Again, we compare the performance of classifier techniques based on symptoms analysis method with traditional classification techniques. For simplicity, we focus on super symptom described in definition (2).

5.1. Data description

The research data set was proposed to be collected from Cancer Oncology Hospital in Medicine City in Baghdad for a set of patients who have been organized for biopsy, mammograms interpreted by radiologists. Laboratory and Clinical Investigations, Ultrasound of mammary glands and provided elastography data on

mammographic results as a part of the standard mammographic workup. The number of patient was 101 in 2017. For this analysis case with remove all missing value. The data set has randomly divided into training and testing subsets of 70% and 30% patients, respectively. Encoding of 11 selected variables is presented in Table 1.

Table 1. Encoding signs

code	name	indicat
A	Age	1- age less than 59 (73%)
		0 - age greater than 59 (27%)
D	The Oncotype DX test	1 - ILC nodal severe type tumore (13%)
		0 - IDC ducts type tumor
		in Pipe lactiferous (87%)
G	Grades of best cancer	1 - poorly differentiated tumor (67%)
		0 - moderately differentiated tumor (33%)
E	Estrogen receptor positive	1 - yes(73%), 0 - no(37%)
P	Progesterone receptor positive	1 - yes(75%), 0 - no(25%)
H	The human epidermal growth factor receptor	1 - HER2-positive breast cancer(73%)
		0 - there is no antigen in tissue(27%)
K	Proliferative activity of cells	1 - greater than 15(60%)
		0 - ki67 less than 15 (60%)
S	Operation removal of the tumor or breast	1 - mastectomy or Lumpectomy (56%)
		0 - excisional biobsy (44%)
T	Advanced type of tumor	1 - the size of the main tumor
		more than 3 cm(68%), 0 - otherwise (32%)
L	Lymph nodes	1 - tumor spreading
		to the lymph node(82%) , 0 - no (18%)
M	Metastasis	1 - distant metastasis(92%)
		0 - No distant metastases (8%)

5.2. Experimentation

For measuring performance, the way to examine which model gives the best predictions is cross validation [10]. The following expression are used [22].

- True Positive (TP): This Patient indicates malignant samples that were classify as malignant.
- True Negative (TN): This Patient indicates benign samples that were classify as benign.
- False Positive (FP):This Patient indicates malignant samples that were classify as benign
- False Negative (FN): This Patient indicates benign samples that were classify as malignant.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (7)$$

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

A different measure of success has been the value of area under a receiver operating characteristic (ROC) curve. This value is drawn based on true positive rate (TPR) sensitivity in Equation 7. and the false positive rate (FPR) 1-Specificity (Eq. 8).

5.2.1. Artificial neural Network model

The algorithm of Resilient Back Propagation is used in this study. Multi-layer networks characteristically employ sigmoid transfer functions in the hidden layers identified in (2),(3) respectively. These functions are frequently termed "squashing" functions, since they compress an infinite input range into a finite output range [8]. The actual outputs of the processing in the output layer identified in (4). In addition to that the method of supervised training also used, and the purpose of training is to regulate the weights till the measured error between the desired output and the actual output is decreased. The neuron is built by applying a program (R

Language) to obtain prediction and accuracy values for a series of 65 patients after remove missing value. The first step in the program is to determine the input of the neural network, where the input is (11) of different covariates: A, D, G, E , P, H , K , S , T, L, M . Patients baseline characteristics are shown in Table 1. The second step we can identify missing value and remove them from original data (n=101) with the intention of determining the number of hidden nodes, which are determined by training, and in training includes many computer experiments, and then choose the best number of repetition that makes the accuracy value as high as possible. The network was trained using a training set (70%) of patients from the data set drawn randomly, n = 43). The model was evaluated using a validation set from the remaining patients (n = 22).

After creating a network based on back propagation algorithm with the input layer, a hidden layer and output layer. There are two group was created for the classification task. The first group include input layer consisted of 8 neurons that symbolized $W_1 = (A, D, G, E, P, H, K, S)$. The second group include input layer consisted of 5 neuroses that denoted $W_2 = (A, D, G, H, K)$ with Super Symptom $R = P(1 + ES)(mod2)$ identified in (1) means practically that progesterone receptor was positive and as a surgical procedure the excisional biopsy was performed.

The hidden layer has dissimilar neuroses (the optimum amount of hidden neuron for a network is difficult to predict). The number of hidden neuron was varied between 1 and 20.

The output layer consisted of a single neuron representing diagnostic outcome ML means the presence of distant metastases and the tumor spreading to the lymph nodes. And then allocate random value to the input weight and the bias, the algorithm is used to train and test network and compare the result by using accuracy.

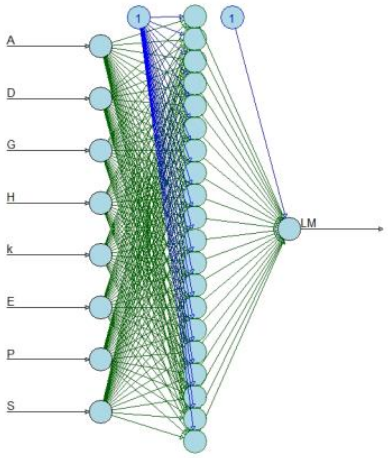

Figure 2. Artificial Neural Network Architecture

Figure 2 illustrates three layer back propagation based architecture of neural network has made for the classification task. The highly exact network training performance was gotten with 20 hidden nodes for the first group. A series of trails had to be made to achieve optimization of result.

All consequences are by using independent variable sample for testing data set based on the breast cancer data set. The test was performed using resilient back propagation network based on all variables and super symptom analysis, we can assess the accuracy of the prediction model with the greatest strength of the independent variables that it predicts in the breast cancer data set [7]. And Table 2 shows the accuracy, misclassification error (training) and (testing) of the network output values .

Table 2. Performance of the training and testing algorithm

Artificial neural	Number of hidden layers	Misclassification Error		
network (ANN)		training	testing	accuracy of diagnosis
W_1 and class ML	1	0.720	0.772	0.227
	2	0.860	0.863	0.136
	3	0.604	0.727	0.272
	4	0.441	0.727	0.272
	8	0.488	0.681	0.318
	10	0.465	0.636	0.363

Artificial neural	Number of hidden layers	Misclassification Error		
network (ANN)		training	testing	accuracy of diagnosis
	20	0.325	0.545	*0.454
$W_2 + R$ and class ML	1	0.302	0.272	*0.727
	2	0.348	0.409	0.590
	3	0.395	0.409	0.590
	4	0.348	0.409	0.590
	8	0.372	0.454	0.545
	10	0.372	0.318	0.681
	20	0.348	0.318	0.681

Table 2 illustrates the influence of training algorithm on the accuracy of diagnosis with back propagation algorithm. From this table, it is obvious that the feed forward neural network by means of back propagation algorithm based on $W_2 + R$ super symptom with class ML produced the best results for diagnosis.
Moreover the network was trained using several hidden nodes in order to select the best amount of nodes, and Table 2 shows the amount of accuracy and thus considered that the best number of nodes for the hidden layers is 20 for first group, and one hidden layer for second group, because the error value in this case is the least.

5.2.2 Discriminate analysis model

The same 65 patients after remove missing value.51 patients for training and 14 for testing, The four general steps of Linear Discriminate Analysis (LDA) classifier is

• Compute the d-dimensional mean vectors for various classes $m_i(i = 0,1)$ of the 2 dissimilar classes from the dataset and the scatter matrices (both within-class and between-class).

The within-class scatter matrix S_W is computed by the following equation:

$$S_W = \sum_{i=1}^{c} S_i \quad (9)$$

where $S_i = \sum_{x \in D_i}^{c} (x - m_i)(x - m_i)^T$ and m_i is the mean vector $m_i = \frac{1}{n_i} \sum_{x \in D_i}^{c} x_k$

The between-class scatter matrix $S_B$is calculated by:

$$S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T \quad (10)$$

where m stands for the total mean, and mi and N_i are the sample mean and sizes of the respective classes.
Then , we compute the generalized eigenvalue problem for the matrix $S_W^{-1} S_B$ to obtain the linear discriminants.

• Calculate the Eigen vectors $(e_1, e_2, \ldots, e_d)$ and respective Eigen values $(\lambda_1, \lambda_2, \ldots, \lambda_d)$ for the scatter matrix.

• Now sort the Eigen vectors by decreasing order of Eigen values and choose k Eigen vectors with maximum value of Eigen values to produce a $(d \times k)$ dimensional matrix$W$ (where every column represents an eigenvector). The eigenvectors are imperative as they will create the new axes of our new feature subspace. Furthermore, the related eigenvalues are of specific interest as they will explain to us how "informative" the new "axes" are.

• Use this above produced $(d \times k)$ Eigen vector matrix W to transform the samples onto the different subspace. And it can be represented shortly by matrix multiplication as $Y = X \times W$, where X is a $(n \times d)$ dimensional matrix representing the n samples, and y are the transformed $(n \times k)$ dimensional samples in the new subspace.
The comparison of performance analysis of the classifiers are prepared between W_1 based on classical method and $W_2 + R$ based on symptom analysis by Uusing 65 patients of the breast cancer dataset.

Table 3. Confusion Matrix of LDA for training and testing set

Classifier	Number of variables	TP	FN	TN	FP
LDA based on training set	W_1 and class ML	36	13	1	1
	$W_2 + R$	37	14	0	0

Classifier	Number of variables	TP	FN	TN	FP
	and class ML				
LDA based on testing set	W_1 and class ML	8	5	0	1
	$W_2 + R$ and class ML	9	5	0	0

Linear discriminant analysis algorithm can be evaluated by confusion matrix which is shown in Table 3 . The performance of the algorithm can be represented by a confusion matrix that provides detailed layout for this performance. Moreover this matrix is used to show the correct and incorrect instances, each of column of the matrix signifies the actual class instances while the row of this matrix signifies the predicted class instances as shown above.

5.2.3. Random forest model

The dataset has been divided into training and testing sets, one third of the dataset are allocated to testing phase. In the first run, there are 46 observations for training and 19 for testing, number of trees to grow is set to 300. Therefore, After training the model, it is possible to try to ignoring each of the predictors one by one in addition to that see which of them negatively affected the accuracy and GINI Index by being removing. After ensured that every observation is predicted at least a few times. Therefore, We have two group to create random forest for variable importance, the first group is W_1 and the second group W_2 + R Super Symptom R in figure (3)(4) respectively.

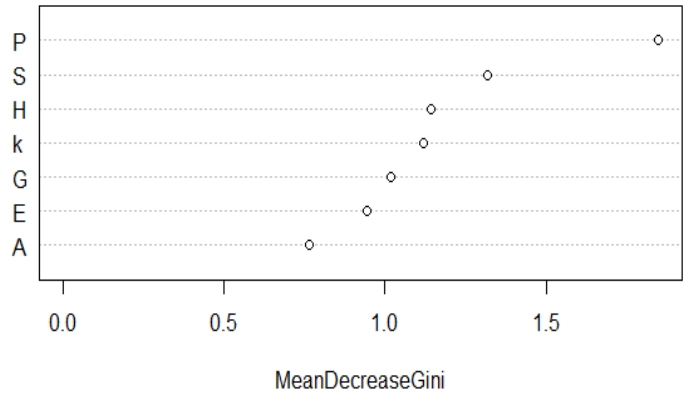

Figure 3. Variable Importance for First Group W_1 (Top 7 -Variable importance)

Figure 3 illustrate impotent variable for average degrease gini measures and how pure the nodes are at the end of the tree without each variable in group W_1. It is obvious that Pprogesteron receptor positive and Ooperation removal of the tumor or breast are influential variables. Age variable added little insight to the model. So this variable is not that important for prediction, the other predictors are of moderate influence.

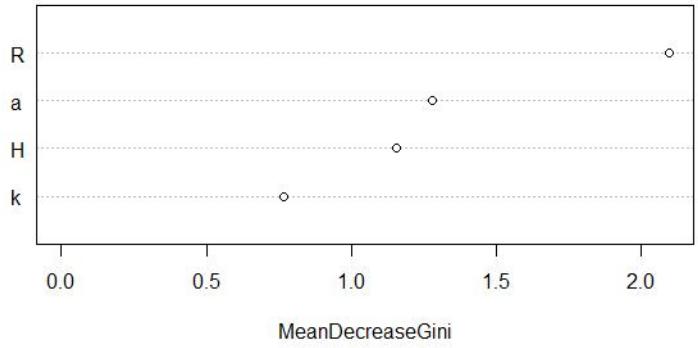

Figure 4: Variable Importance for Second GroupW_2 + Super Symptom R (Top 4 -Variable importance)

Figure 4 above consists of second group W_2 + R Super Symptom R , Iit is obvious that Super Symptom R has highest contribution towards parameter of Mean Degrease Gini. So it is an influential variable.

6. Experiment and results

The experimental results of the breast cancer disease for prediction using Artifisial Neural Network (ANN), linear discriminant analysis (LDA) and random forest (RF) are analysed in this section. The related data to breast cancer diseases are collected from 65 patients who are provided by Cancer Oncology Hospital in Medicine City in Baghdad .

6.1. Comparing classifier model based on symptom analysis with other classifier learning systems (full training mode)

For testing the abilities of our approach in a more noteworthy way, we continued as follows. Initially, we designed a full training data experiment for using 65 patients of the breast cancer dataset, we used 0.70 (training) versus0.30 (testing) split and applied the average value of the performance metrics accuracy of all results. Furthermore, the statistical tests intend to prove if there are significant differences, in this case the super symptom analysis $R = P(1 + ES)(mod2)$ with class ML have $p\_value = 0.000003$ of Fisher exact test.

Table 4. Results for diagnosing of breast cancer

Classification methods	Number of variables	Misclassification Error		
		training	testing	accuracy of diagnosis
Artificial neural network (ANN)	W_1 and class ML	0.325	0.545	0.454
	$W_2 + R$ and class ML	0.302	0.272	*0.727
Linear discriminant analysis (LDA)	W_1 and class ML	0.274	0.428	0.571
	$W_2 + R$ and class ML	0.27407	0.357	*0.642
Random forest (RF)	W_1 and class ML	0.153	0.369	0.631
	$W_2 + R$ and class ML	0.261	0.316	*0.684

The properly classified data for breast cancer diagnosis has been examined and its accuracy is calculated for the 3 classifier are shown in Table 4 above. After completing the training of the three classifier model. From Table 4 above, the testing results shows that classifier model based on super symptom $W_2 + R$ in terms of accuracy performs better than the first group W_1 .

Table 5. Results for classifier model based on symptom analysis $W_2 + R$ and class ML

Techniques	ANN	LDA	RF
TN	0	0	0
TP	16	9	31
FP	0	1	1
FN	6	5	5
TN+TP+FP+FN	22	14	19
TP+FN	22	14	18
TN+FP	0	0	1
Accuracy	0.727	0.642	0.684
Sensitivity	0.727	0.642	0.722
Specificity	0	0	0

Table 6. Results for classifier model based on W_1 and class ML

Techniques	ANN	LDA	RF
TN	0	0	0
TP	11	8	12
FP	5	1	2
FN	6	5	5
TN+TP+FP+FN	22	14	19
TP+FN	17	13	17
TN+FP	11	1	2
Accuracy	0.454	0.571	0.631
Sensitivity	0.647	0.615	0.705
Specificity	0	0	1

From these results, we can deduce that, comparing our method symptom analysis with other classifier learning systems. To achieve a preliminary comparison between the obtained performance of classifiers with the proposed method super symptom analysis and the ones obtained with classical methods (ANN, LDA, RF), we have executed three models in 0.30 testing mode. The comparison is shown in Tables 5 and 6, which show, for each classifier, the following information: true positive rate, false positive rate, true negative rate, false negative rate, accuracy, sensitivity, specificity.
As for the classifier learning based on symptom alalysis (observe that the finest consequence arisen from the suggested method).

7. Iterative algorithm for constructing super symptoms

In addition to the supersymptom $R_1 = R + 1 = 1 + EPS + P(\mathrm{mod}2)$, for which the uncertainty coefficient with the variable ML equals 28.85%, we can consider the supersymptoms $R_2 = GK + GKS + S(\mathrm{mod}2)$ and $R_3 = 1 + HPS + HS + P + PS(\mathrm{mod}2)$ with uncertainty coefficients 21.53% and 23.94%, respectively. These supersymptoms R_1, R_2, R_3, determine groups that are more favorable from the point of view of the emergence of ML, which cannot be described only by the unidirectional action of factors: according to R_1, a more prosperous group is formed by both those with $P = 0$ and as well as those with $P = 1$, but only in combination with $ES = 1$; according to R_2, a more prosperous group is formed by those patients with $S = 1$ and as well as those with $S = 0$, but only in combination with $GK = 1$; according to R_3, the risk group for ML is formed by those patients with $P(1 + \mathrm{S})(\mathrm{mod}\ 2) = 1$ and as well as those with $S(1 + \mathrm{P})(\mathrm{mod}\ 2) = 1$, but only in combination with $H = 1$.
If we apply the procedure for identifying the most informative supersymptom on these three new variables, then we get the non-occurrence factor ML in the form of the product ($R_1\ R_2\ R_3$) taking into account in general six variables E, P, S, G, K, H, with the uncertainty coefficient 28% for the variable ML. In the presence of all factors at the same time, ML was observed in 52% cases, and in the absence of at least one in 97% cases. Thus, applying one additional iteration, we obtain a significant supersymptom for six variables at once with a significant reduction in the enumeration of combinations. This result is important for determining further actions in terms of the development of this scientific approach, in particular, it is supposed to study the accuracy and reasons for the rapid convergence of the iterative procedure of symptom-syndrome analysis.

8. Conclusion

In this study, we have suggested a new procedure for categorical classifier learning. The proposed algorithm is tested on a real life problem such as diagnostic problems of cancer (medical diagnosis). Our proposal is based on defining symptom analysis for learning a classifier as (ANN)(LDA) and (RF). The study shows that the precision (accuracy) for the prediction analysis of breast cancer data is acceptable and can help physicians in decision making for early diagnosis. The accent here is on the role of classifier learning based on symptom analysis in cancer management and prognosis. Additional work as future trend is required to raise the accuracy of classifying breast cancer diagnosis.

Acknowledgement
The work was carried out with the support of the RFBR, under Grant No. 20-01-00096-a.

References

[1] Neural Networks. https://djinit-ai.github.io/2018/02/08/neural-networks.html. [accessed 12 oct 2019].

[2] K. Kawaguchi, Artificial Neural Networks.The McCulloch-Pitts Model of Neuron. http://wwwold.ece.utep.edu/research/webfuzzy/docs/kk-thesis/kk-thesis-html/node12.html.[accessed 12 oct 2019].

[3] N. Alexeyeva, Analysis of biomedical systems. Reciprocity. Ergodicity. Synonymy. Publishing of the Saint-Petersburg State University, Saint-Petersburg, 2013.

[4] N.Alexeyeva, P.Gracheva, E.Podkhalyuzina, K. Usevich, Symptom and syndrome analysis of categorial series, logical principles and forms of logic. In Proceedings, 3rd International Conference on BioMedical Engineering and Informatics BMEI, pp. 2603–2606, 2010.

[5] N.Alexeyeva, A.Alexeyev, E.Verbitskaya, E.Krupitsky, Final geometry and a logic principle of projectivity in the statistical analysis of the medical data. In Bulletin of the international statistical institute LXII-2007, pages 3021–3024. Lisbon, Portugal, 2009.

[6] N.Alexeyeva, P. Gracheva, B. Martynov, I. Smirnov, The finitely geometric symptom analysis in the glioma survival study", 2nd International Conference on Biomedical Engineering and Informatics, pp. 1-4, 2009.

[7] H. B. Burke, D. B. Rosen, and P. H. Goodman. "Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival." In *Advances in neural information processing systems*, pp. 1063-1067. 1995.

[8] F. Paulin, and A. Santhakumaran. "Classification of breast cancer by comparing back propagation training algorithms." *International Journal on Computer Science and Engineering* 3, no. 1, pp.327-332, 2011.

[9] N. Ampazis, "Introduction to Neural Networks",1998. *www. iiitnrcps-t. gr/neural/index*.

[10] B. Abraham, J. Ledo, "Statistical Methods For Forcasting " John Willy, Sons, New York, 1983.

[11] D. B. Sandy, "Statistical Aspects of Neural Networks", Quantitative Economics and Statistics, 1225 Connecticut Avenue, NW, Washington, USA, 2003.

[12] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural networks, vol.21, no.2-3, pp.427-436, 2008.

[13] F. Amato, et al. "Artificial neural networks in medical diagnosis", 2013.

[14] J. Đ. Novakovic, and A. Veljovic, "Solving medical classification problems with RBF neural network and filter methods", *International Journal of Reasoning-based Intelligent Systems*, vol.9, no.2, pp.80-89, 2017.

[15] A. Z. Issa, Neural networks architecture algorithms and applications, 2000 (Tanslated in Arabic).

[16] T. Kiyan And T. Yildirim, "Breast Cancer Diagnosis Using Statistical Neural Networks", *Journal Of Electrical And Electronics Engineering*, , vol. 4, Number 2, pp.1149-1153, 2004.

[17] S. D. Swarkar, A. Ghatol, A. P. Pande, "Neural Network Aided Breast Cancer Detection and Diagnosis Using Support Vector Machine" Proceedings of the International conference on Neural Networks, Cavtat, Croatia, pp. 158-163, 2006.

[18] R. D. Leone, R. Capparuccia and E. Marelli, "A Successive Overrelaxation Backpropagation Algorithm for NeuralNetwork Training" IEEE Transactions on Neural Networks, vol. 9, No. 3, pp. 381-388, 1998.

[19] A. M. Martinez, A. C. Kak, "PCA versus LDA", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.23, no.2, pp. 228–233, 2001.

[20] T.K. Ho, "Random Decision Forests", Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal. pp. 278–282, 1995.

[21] L. Breiman, "Random Forests". Machine Learning, vol.45, no.1, pp. 5–32, 2001.

[22] G. Forman, "An extensive empirical study of feature selection metrics for text classification", *Journal of machine learning research*, vol.3, pp.1289-1305, 2003.

[23] G. I. Salama, M. Abdelhalim, and, M. A. Zeid, "Breast cancer diagnosis on three different datasets using multiclassifiers", *Breast Cancer (WDBC)*, vol.32, no.569, p.2, 2012.

[24] J. Han, M. Kamber , Data mining: concepts and techniques. 3rd edition. USA: Elsevier, pp. 5–6, 2012.

[25] S. Boyd and L. Vandenberghe, *Introduction to applied linear algebra: vectors, matrices, and least squares*. Cambridge university press, 2018.

[26] N. P. Alexeyeva, F. S. Al-Juboori, and E. P. Skurat, "Symptom analysis of multidimensional categorical data with applications", *Periodicals of Engineering and Natural Sciences (PEN), vol.8*, no.3, pp.1517-1524, 2020.