Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq
DOI: https://doi.org/10.33095/jeas.v29i137.2760

Machine Learning Techniques for Analyzing Survival Data of Breast Cancer Patients in Baghdad

Noor Ayad Mohammed
University of Baghdad, College of Administration and Economics, Department of statistics.
Baghdad, Iraq
noor.mohammed1101@coadec.uobaghdad.edu.iq

Entsar Arebe Fadam
University of Baghdad, College of Administration and Economics, Department of statistics.
Baghdad, Iraq
entsar_arebe@coadec.uobaghdad.edu.iq

Received: 23/5/2023 **Accepted: 16/7/2023** **Published: 1/9/ 2023**

Abstract

The machine learning methods, which are one of the most important branches of promising artificial intelligence, have a great importance in all sciences such as engineering, medical, and also recently involved widely in statistical sciences and its various branches, including analysis of survival, as it can be considered a new branch used to estimate the survival and was parallel with parametric, nonparametric and semi- parametric methods that are widely used to estimate survival in statistical research. In this paper, the estimate of survival based on medical images of patients with breast cancer who receive their treatment in Iraqi hospitals were discussed. Three algorithms for features extraction were explained: The first principal component algorithm, The second kernel principal component algorithm and The last is faster ICA algorithm. Then the important features extracted in the three algorithms for features extraction will be entered to machine learning algorithms: The first K nearest neighbor algorithm, The second survival tree algorithm (or regression tree), and the last random survival forests algorithm.

Two criteria for comparing the best models to estimate survival were relied on the MSE, and the C-Index. The best model for estimating and predicting survival is the use of the fastest ICA algorithm with the random survival forest algorithm that gave the lowest amount to MSE and the highest value to C-Index. Accordingly, we recommend doctors and medical professionals in Iraq to adopt this model to estimate survival for patients with breast cancer.

Keywords: Survival function, Feature extraction, Machine learning, Mammogram medical images, Breast cancer.

1. Introduction

Statistical analysis includes many branches, including survival models, Regression models and time series, but at the present time the survival analysis is of great importance, especially after the multiplicity of methods of analysis and the entry of artificial intelligence algorithms. The process of building an predictive model for survival times for patients with cancer diseases by using machine learning algorithms for real data delete it represented by medical scanning images is one of the great challenges, and this is due to many reasons, the first of which is the difficulty of obtaining medical data for a big sample size of patients during a specific time to study and this is mainly due to the society nature or behaviour, (e.g. Iraq), where the research data were obtained from the capital, Baghdad hospitals,

secondly to determine which algorithms for machine learning are the best that give good results for this type of data that need a process of experimenting with a number of algorithms to test their ability to predict, third: Dealing with medical images needs a lot of questions and inquiries by specialists in oncology and radiology to clarify How to make a logical and accurate interpretation to deal with such a kind of images, fourth: The scarcity of sources used for this type of data in Iraq is due to the nature of the data on mammogram images, as this device cannot store images of infected breasts for breast cancer patients for more than a month, due to the large number of daily examinations that are conducted with it for normal cases and cases with benign and malignant tumors because imaging mammogram device is considered one of the first and most important steps in diagnosing breast cancer patients, in addition to the frequent malfunctions that occur with the device, noting that the data set for the research was obtained from more than one centre specializing in oncology in Baghdad, and therefore cannot be compared to any other research that used a dataset available on the Internet and is not representative of real country data ,and therefore we cannot provide any benefit to the country because we did not study its real data. Through the above points, the importance of the research is centred, which aims to build a predictive model for the patients suffer from breast cancer survival in Iraq by introducing medical images to the steps of making the machine learning techniques as an attempt to help doctors in decision making and those interested in the medical aspect to take the necessary precautions and provide appropriate treatments to deal with a state of prediction to patients status.

The use of machine learning methods to analyse survival began within the limits of the year 2000, in this year as Zupan and et.al (2000) research dealt with an estimate of survival of the disease returning to prostate cancer patients using machine learning, due to the advantages of this modern method that overcame the normal statistical methods in dealing with the nonlinear and censored data, and dealing with such problems without the need to know the statistical distribution of the data, the research also dealt with the use of the decision-making tree algorithm and the Bayes Classifier showed much better results compared to the cox model for survival.

Fadnavis (2019) Applying the machine learning methods to analyse survival, where a survey of traditional statistical methods and different types of machine learning methods, including survival trees, naive Bayes, neural networks and deep neural networks for survival analysis and explain how to integrate classical methods with machine learning techniques to produce algorithms that have the ability to solve high dimensions problems for data of survival analysis.

Nemati and et al (2020) used machine learning methods and the nonparametric methods and semi-parametric methods to analyse survival and predicting times from the hospital to patients with (Covid-19) depending on the data of age and sex, as the results have proven the superiority of the machine learning algorithm on the rest of the other methods to predict the survival times of patient in the hospital.

2. Material and Methods

2.1 The Data Set of the Research Topic

Breast cancer is the most common type of cancer, Breast cancer occurs in every country in the world among women of all ages after puberty, but at increasing rates later in life.

Thus, the most important stages of breast cancer diagnosis is the CT scan, which is done with a special device for the breast only, through which the condition is diagnosed if it is normal or infected with a tumor, from which a biopsy is taken and analyzed to confirm whether it is benign or malignant, and then the treatment protocol is started according to age, stage and age, cancer cell, etc.

Survival analysis is concerned with the analysis of survival data, which means the time between the beginning of the event until the occurrence of the end event represented by death or any other end event, which is subject to a set of strict conditions, including that it has a specific statistical distribution, or some cases do not reach the end event called censored data, or The survival times are dependent on a large number of variables, so the machine learning techniques has the ability to deal with these problems (Yaqoob & Ali ,2021). One of the most important stages of the machine learning techniques is the availability of data on the topic of research or study, The data on the subject of the research represent the medical images resulting from the mammogram tomography of the patients diagnosed with breast cancer, and then converting the image of each patient into a digital matrix and extracting the important features from it and including them within the machine learning algorithms to estimate survival. The research sample included mammogram images of 70 diagnosed cases of breast cancer in Iraq, two images of the infected breast were extracted for each patient and the images were extracted after using the medical program for showing the medical images (Dicom Viewer), which through the options for this program can be obtained the image free of the patient's name, the date of the examination, the file number and the name of the hospital that the patient retreats, as this information is accompanied by all images of a mammogram examination for each patient.

The research sample was collected from three educational centres specializing in oncology in Baghdad, namely: Oncology Teaching Hospital in the Medical City, Al-Yarmouk Teaching Hospital, Al-Andalus Private Hospital, And for the period from 1-6-2021 to 1-11-2022

Noting that the resulting images are gray images which are one of the types of digital images whose elements take gray gradients (Jabir & Mohammed ,2020).

2.2 Featuere (Feature) Extraction algorithms (Extract important features)

After converting the medical image of each patient diagnosed with breast cancer into a digital matrix, and that this matrix has high dimensions, so it is necessary to resort to using methods to extract the important features from this matrix and include them within the method of machine learning, therefore it is necessary to address an algorithm for extraction the features is called for this big or enormous data, which can be defined as choosing a sub -set of features, features or dimensions that produce similar results for the full feature, or in other words, the removal of features (variables) not important or repeated, i.e. reducing the data entered into the reasonable size of processing and analysis, and therefore it means reducing the time, effort and cost used in the analysis of the huge data (Jabir & Essa ,2019).

The process of extracting features can also be defined as the process of reducing the representative components of the image to a less than the components that carry a sufficient amount of information, as extracting features helps in obtaining a summary of the image. The conversion of the input data to the set of feature is called the extraction of features (Abdulwahhb & Abd Alrazak , 2020).

This algorithms is of great importance, due to the noticeable increased use of huge data at the present time, and accordingly, the primary goal of choosing features is to improve prediction performance and provide accuracy and speed in results (Jabir & Essa 2019).

2.2.1 The Principal Component Analysis Algorithm

The principal component analysis algorithm is one of the most common and used algorithms in addressing the problem of multilinearity, when the number of explanatory variables is greater than the sample size for the studied status (Jabir & Essa 2019). In the case of images transferred to a digital matrix, the goal will be to reduce the linear dimensions that lead to reducing the dimensions as much as possible for the images that concern the state of the study without affecting the accuracy of the information obtained from the image, and therefore it leads to facilitating the process of image processing (Jabir& Mohammed, 2020) .

The mathematical procedure of the principal components requires first finding the Variance-Covariance matrix for the available images, which are calculated according to the following form (Jabir & Mohammed, 2019) :

Step1: Read the image of the studied case and convert it into a digital matrix.
Step2: Finding a matrix of Variance-Covariance for images of the studied status, which can be clarified in the following form

$$\boldsymbol{\Sigma_{ij} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{i=1}^{N}(X_i(\kappa,\iota)-\mu_i)\left(X_j(\kappa,\iota)-\mu_j\right)} \quad (1)$$

M, N: The number of rows and columns in the image.
X_i (κ, ι) The component is at the point (κ, ι) in the image.
X_J (κ, ι) The component is represented by the point (κ, ι) in the image.
[M_J, μ] I represents the arithmetic meam of the component and is given as follows

$$\boldsymbol{\mu_i = \frac{1}{N}\sum_{j=1}^{N} X_j(\kappa,\iota)} \qquad \boldsymbol{\mu_j = \frac{1}{M}\sum_{i=1}^{M} X_i(\kappa,\iota)} \quad (2)$$

Step3: finding a matrix of variance – covariance as follows:

$$\Sigma = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix}$$

Where:
Σ: The matrix represents variance – covariance, which is a symmetric square matrix and its determinant is not equal to zero.
C_{ij}: which is the covariance between the vectors (X_i , X_J)
Step4: Finding the Eigen value according to the following formula:

$$\boldsymbol{|\Sigma - \lambda_j I| = 0} \quad (3)$$

Where:
$\sum_{j=1}^{k}\lambda_j = k$
λ_j: represent the eigen value j
K: The sum of the eigenvalues is equal to one, but when taking the highest k of the eigen values, their sum is equal to k.
Step5: Find the Eigen vectors according to the following formula:

$$\boldsymbol{(\Sigma - \lambda_j I)a_j = 0} \quad (4)$$

a_j: represent the Eigen vector j
Step6: Finding the principal component according to the following formula:

$$\boldsymbol{F_{1j} = a_{1j}X_1 + a_{2j}X_2 + \cdots + a_{kj}X_k} \quad (5)$$

Where F_{1j} is a principal component

2.2.2 The Kernel Principal Component Analysis Algorithm

The principal component analysis algorithm is used as a preliminary step that precedes statistical treatment (classification or prediction) to reduce the high dimensions of the digital image variables of the situation under study (i.e. extracting important and related variables and excluding unimportant, but in the event that high dimensions are not linear, principal component analysis algorithm is resorted to the kernel principal component analysis algorithm (Scholkopf and et.al 1996). The basic idea of it depends on the beginning that the kernel functions allow the work on some of the spaces of the variables implicitly, while in other ways the work on the space of the total variables, and therefore the change is a conversion to the original dimensional space R^d to a new sacred space F, as the conversion is The following form:

$\emptyset: R^d \rightarrow F$

Therefore, $\emptyset$ represents the transformation from the original space to the contracted space (Abdulwahhb & Abd Alrazak 2020).

In practice, the dotted multiplication in the space of variables is expressed by the term of the kernel functions in the input space, and the main sequence of this is that any algorithm that uses only constant multiplication can be transformed into its non-linear state using the kernel methods (Abdulwahhb & Abd Alrazak 2020).

Therefore, the steps for feature extraction depending on the Kernel principal component algorithm can be summarized as follows (Scholkopf and et.al 1996):

Step1 : Read the image of the studied case and convert it into a digital matrix.

Step2 : calculate the variance and covariance matrix using the same mechanism that was calculated in the PCA algorithm, according to the following formula:

$$C_X = \frac{1}{N}\sum_{i=1}^{N} X_i X^T{}_i \qquad (6)$$

Where:

N: number of data points in the set.

Step3 : calculating the Eigen values and the Eigen vectors of the matrix of variance and covariance according to the following formulas:

$$\lambda v = C_x v \qquad (7)$$

Where:

λ: It represents the Eigen values

$v \in R^n$: It represents the Eigen vectors that are given in the following formula

$$X_{pc}{}^{k} = v^k . X \qquad (8)$$

Step4 :The original space of the variables is converted into another space as follows

$\emptyset: R^n \rightarrow H$

$X \rightarrow \emptyset (X)$

Consequently, the matrix of variance and covariance will be rewritten as follows:

$$C_{\emptyset(X)} = \frac{1}{N}\sum_{i=1}^{N} \emptyset(X_i)\, \emptyset(X_i)^T \qquad (9)$$

Step5 : Compensation for the formula of the matrix of variance and covariance in the above step in the form (7) to obtain the eigen values .

$$\lambda v_\emptyset = C_{\emptyset(X)}\, v_\emptyset$$

$$= \frac{1}{N}\sum_{i=1}^{N} (\emptyset(X_i) . v_\emptyset)\ \emptyset(X_i) \qquad (10)$$

Step6 : Depending on the above formula, each eigen vector can be written as follows:

$$v_\emptyset = \sum_{i=1}^{N} \alpha_i \emptyset(X_i) \qquad (11)$$

Where:

α_i: represent coefficients of $\emptyset\ (X_i)$
By multiplying the formula (10) above with $\emptyset\ (X_K)$ from the left and putting the formula (11) we get

$$\gamma \sum_{i=1}^{N} \alpha_i \left(\emptyset\ (X_k).\emptyset\ (X_i)\right) = \frac{1}{N} \sum_{i=1}^{N} \alpha_i \left(\emptyset\ (X_k). \sum_{j=1}^{N} \left(\emptyset\ (X_j)\ .\ \emptyset\ (X_i) \right) \right) \emptyset\ (X_j) \qquad (12)$$

Where:
k ∈ [1,N]
Step7 :Define a square matrix that dimensional called N×N called Kernel so that
$K_{ij} = (\ \emptyset\ (X_i)\ .\emptyset\ (X_j))$
Accordingly, the above formula will be into

$$\lambda\, K\alpha = \frac{1}{N}\, K^2\, \alpha \qquad (13)$$

Where:
$\alpha = (\ \alpha_1, \alpha_2, \dots, \alpha_N)^T$
The solution to the formula (13) can be reached by solving the eigen values, by multiplying the two ends of the formula (13) with K^{-1} We get

$$N\, \gamma\, \alpha = K\, \alpha \qquad (14)$$

Step8 : To solve the eigen values equation of $C_\emptyset$ (x), which is equivalent to eigen values to K and since α the values of the eigen value of the matrix K, so it is produced $\lambda_k \alpha^k \alpha^k = 1$
so
$K\alpha - \tilde{\lambda}\alpha = 0 \Longrightarrow (K - \tilde{\lambda})\alpha = 0$
The condition necessary to find the above equation is to be:
$|K - \tilde{\lambda}| = 0$
Therefore, the formula of the kernel principal component is written as follows:

$$F_{4j} = \sum_{i=1}^{N} \alpha_{ij} K\ (x_i, x_j) \qquad (15)$$

2.2.3 The Fast independent component algorithm (Fast ICA)

The Fast ICA algorithm is known as one of the most important and most used ICA algorithms and before starting to explain it must be explained and clarified the ICA algorithm or what is called the algorithm of the analysis of independent component, which is one of the statistical methods and its idea depends on making the variables involved in the study studied independent of each other, i.e. converting a complex group And as large of correlated data to partial statistically independent component, ICA has multiple and wide applications in several fields and in particular in digital signal processing, extracting important features, analysis of financial time chains, digital image processing and other applications (Benlin & et.al ,2008).
The model of the ICA algorithm can be expressed as follows:

$$X = AS + n \qquad (16)$$

where
X : Represent a vector of observation.
A :That represents the mixing matrix.
S: Independent component vector.
n: Random errors vector.

The ICA algorithm includes many algorithms to achieve this goal and the Fast ICA algorithm is one of the most common and used algorithms for the estimation of independent component, as the idea of this algorithm depends on the entropy function, which works to measure the non- Gaussian, which is a measure or standard for independence, this algorithm is characterized by speed the high implementation of the typical function is a high stability (Benlin & et.al ,2008).

2.3The Machine Learning Algorithms Used to Estimate The Survival

After the feature extraction algorithm were discussed, which are one of the steps accompanying the algorithms of machine learning here, it is necessary to clarify the machine learning algorithms, which can be defined in a simple way as a set of instructions addressed to the computer to implement a specific procedure and the machine learning programs are often written using many languages Programming like Java, Python, Matlab or R, as each of these languages includes libraries for machine learning programming that supports a variety of machine learning algorithms (Zupan & et.al , 2000).

The machine learning algorithms differ from other artificial intelligence algorithms In terms of its dependence on the availability of real data in the studied phenomenon, in addition to its ability to deal with problems that accompany the data, including high dimensions or correlation between the studied variables, the presence of missing values in the data set, and thus machine learning algorithms outperform other artificial intelligence algorithms that require the above conditions to be met in the dataset, as the machine learning algorithms depend heavily on data and a model of data is made of data and the higher number of data used in the algorithm, the greater complexity process in the algorithm and the greater to the accuracy of the results and vice versa (Nemati & et.al , 2020).

2.3.1 K Nearest Neighbors Algorithm

It is one of the nonparametric methods used with data that does not have a specific probability distribution, presented for the first time with a medical research by (Fix and Hodges, 1951) and this method has been known since then as the closest neighbor method, and this algorithm was used in the event of classification and Regression. It is characterized by its simplicity and efficiency with the big data, but it needs a large implementation time compared to the sizes of small data, as well as its ability to deal with data that suffers from extremist values (Imandoust & Bolandraftar ,2013).

The nearest neighbor algorithm is one of supervise machine learning algorithms, as it needs a training group to determine the prediction of the new case.

The principle of the work of this algorithm in the case of Regression, which includes a general case and a special case the so –called the nearest neighbor when k=1, The dependent variable of a new case is calculated as that the nearest neighbor is calculated to this new case using one of the measures of the distance, and therefore the value of the dependent variable to be found is the same as the value of the dependent variable of the closest neighbor, while the concept of the general situation is when more than one neighborhood (K>1) is determined to the new value In this case, the value of Y for the new case will be average y values for the Nearest neighbors (Imandoust & Bolandraftar ,2013).

It is possible to clarify the steps of the K Nearest Neighbor Algorithm as follows:

Step1: Entering the data after selecting the significant variables according to one of the three feature extraction algorithms described in the paragraphs 2.2.1, 2.2.2, 2.2.3 with the entry of the survival times (the dependent variable for the training cases).

Step2 : Determining K, which represents the number of neighbors for the new situation required to predict the time to stay for it by relying on the training group entered in the previous step, noting that the process of determining a specific value to K is not in the simple process as well as it is not possible to determine fixed values for all applications (In this paper, k is chosen equal to 5).

Step3 : Calculate the distance between the new value and the values of K from the group of values close to the new or required value, so the appropriate measure must be determined to calculate the distance, which usually uses the Euclidean distance scale, which is given according to the following formula:

$$\mathbf{D(x,p) = \sqrt{(x-p)^2}} \qquad (\mathbf{17})$$

Where:

$D(x,p)$:The Euclidean distance between the value required is to know the value of its dependent variable and the value of the nearest neighbor.

X :The new value is required to know the value of its dependent variable.
P :The value of the nearest neighbor to the new value.
Step4 : It is possible to predict the survival time for the new value by relying on the value of K that was found in the second step above, as the prediction value represents the average survival times to K from the neighbors that are given in the following formula

$$\mathbf{y = \frac{1}{k}\sum_{i=1}^{k} y_i} \qquad (\mathbf{18})$$

Step5 : The prediction of the new value by relying on what has been clarified with the above step is inaccurate because all points K nearest to the new point will have the same importance in finding the survival time for the new situation and thus to make the closest point of greater importance to be used for the points nearest to the point the required and the weight for each point is given according to the following formula:

$$\mathbf{w(x, p_i) = \frac{exp(-D(x,p_i))}{\sum_{i=1}^{k} exp(-D(x,p_i))}} \qquad (\mathbf{19})$$

Where:
$w(x, p_i)$ *:* The weight of the Pi point neighbor to the new point.
$D(x, p_i)$: The distance between the Pi point and the new point
$\sum_{i=1}^{k} \exp(-D(x, p_i))$ *:* The total distances for all points are on the new point.
Noting that the total weights of all points are equal to the one.

$$\sum_{i=1}^{k} w(x, p_i) = 1$$

Consequently, the predictive value of the survival time for the new situation will be given according to the following formula

$$\mathbf{y = \sum_{i=1}^{k} w(x, p_i) y_i} \qquad (\mathbf{20})$$

2.3.2 Decision Tree algorithm

The algorithm of the survival tree is one of the non-parametric supervised learning algorithms that can be used for classification and regression, as it is used to predict when the data is based on the regression tree and also called the survival tree when the value required to be predicted is the time of survival, this type of algorithms depends on the type of machine learning subject to supervision which includes the final results of a set of data entered (Sishi & Telukdarie , 2021).
The concept of making this algorithm in the case of regression (to predict the survival time for the new situation), which can be discrete or continuous values and is done by structuring decisions and results in the form and tree structure, which includes three main parts: (Sishi & Telukdarie, 2021)
1- The root nodes, which usually represent all data for the studied situations.
2- The peripheral nodes, the aim of which includes the results of the condition that precedes it, but without specifying these results, and the peripheral contract usually ends with the leaf nodes.
3- The leaf nodes, which represent the prediction of the required value, and it is the end of the decision tree, that is, no other party is branched from it, such as the peripheral contract.

The steps for conducting the regression tree algorithm for the estimation of survival are based on the following steps:

Step1: Entering the data set for the studied status with the introduction of the final results of survival times that the regression trees fall within supervised learning.

Step2: Forming the decision of the decision tree for the data set and setting the appropriate conditions for the formation of the peripheral and leaves nodes.

Step3: The standard deviation values are used to build the peripheral contract of the tree and the CV coefficient to determine at any number the branch is stopped for the peripheral nodes.
Whereas, the formulas for the arithmetic mean, standard deviation, and coefficient of variation are shown as follows:

$$\bar{Y} = \sum_{i=1}^{n} Y_i \quad (21)$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n}} \quad (22)$$

$$CV = \frac{SD}{\bar{Y}} \times 100\ \% \quad (23)$$

Step4: The third step is to do the so -called standard deviation reduction, which includes several steps:

- It begins by calculating the standard deviation of all the data entered, and then the data set is divided by relying on the conditions of certain division into several parts.

- Calculate the standard deviation of each part of the divided parts.
- Submit the standard deviation of each part of the total standard deviation of the data, which symbolizes SDR.
- Choose the group that has the highest SDR and implement the steps above again on this group until all the data is completed and access to the leaves nodes in the regression tree, which represents the prediction.
- The predictive value is reached that is extracted by finding the computational mean of the values included in the leaves nodes of the verified condition.

2.3.3 Random Survival Forests Algorithm (RSF)

It is one of the machine learning algorithms that are used for classification and regression. It is considered an expansion of the principle of survival trees. Instead of making a single survival tree according to the steps that were presented in the regression tree in the previous paragraph, several trees will be made, as the idea of random forests depends on finding more than one sample For the entered data group by using the Bootstrapping sampling and then making the steps of the regression tree for each group, and therefore the predictive value will be the computational mean of the predictive values resulting from each tree, the goal of this step is to refuse the random state and reduce the variance without increasing the bias in the data group and thus increasing Accuracy to predict survival times (Isltwaran & et.al, 2008).
The steps of this algorithm can be displayed as follows:

Step1: Enter the data set for the study of survival times.
Step2: Use Bootstrap sampling to draw samples with a number of B from the collection of data entered according to the above step, as each drawn sample includes n observations, noting that the samples are withdrawn with the return randomly.
Step3: Making the regression tree for each sample that was found by using a bootstrap sampling according to the steps of the regression tree shown in the previous paragraph, and since each tree will give prediction to the survival time, therefore we will get B from predictions.

Step4: Calculation The mean of the predictions obtained according to the third step to be the predictive value of the new singular or new observation.

3. Discussion of Results

3.1 Stages of Estimating Survival Time Using Machine Learning Algorithms

The most used programs with machine learning algorithms are Python, Matlab, R However, the Python program is considered one of the most fast and accurate programs, and it includes many codes for statistical and mathematical models, and it is considered one of the most used programs for artificial intelligence models, especially machine learning algorithms. Accordingly, the algorithms for the subject of the research were implemented and their results obtained through the use of the Python program, as the stages for estimating survival are explained as follows:

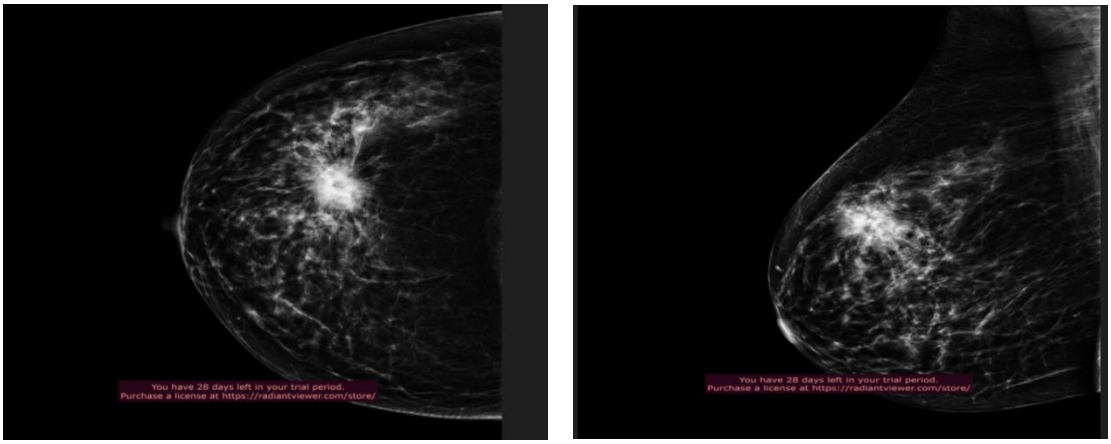
First: Reading Excle Sheet to extract Y, which represents the estimate of the survival times estimated by the tumors specialists, depending on the status of each patient who is included in the study, as well as the serial for patients and pictures of each patient, and that the main reason for making this step because the machine learning method used for estimating the survival times by supervising approach, we will have Two pictures for each patient and Y vector which represent the ages estimated by the specialist for 70 patients as follow.

Table (1): Shows the survival times of 70 patients estimated by specialist doctors measured by months

	Survival time		Survival time		Survival time
1	4	26	7	51	5
2	9	27	5	52	9
3	6	28	4	53	15
4	5	29	8	54	5
5	4	30	3	55	8
6	5	31	4	56	9
7	5	32	3	57	8
8	5	33	2	58	2
9	5	34	3	59	20
10	8	35	7	60	6
11	4	36	3	61	4
12	4	37	4	62	5
13	6	38	10	63	6
14	4	39	3	64	3
15	4	40	11	65	1
16	6	41	4	66	4
17	5	42	10	67	10
18	1	43	5	68	6
19	6	44	6	69	5
20	5	45	10	70	6
21	3	46	8		
22	8	47	3		
23	2	48	6		
24	4	49	4		
25	2	50	1		

Second: Read the file for the images or the so -called Path, any place to store the mammogram images in the JPG in the computer, which is being found by using the code called TNCER Flow, which is located within the Python program code.

And since the raw images that are taken from the mammogram device of the breast affected by tumors are as in the figure below, which is an anterior and a lateral image, and usually two images for each breast, meaning that the total images for the examination contain four images (anterior and lateral image of the right breast and a frontal and lateral image of the left breast)



(a) (b)

Figure (1): The Figure shows the raw image of a breast cancer patient, as the image (a) represents the front image of the affected breast and the image (b) represents the side image of the affected breast.

Third: Make a step to change the image size (Resize) for each image to be 1024 x 1024 for all algorithms within the work steps and the purpose of this step is to remove all the signs in the image and the unnecessary that are represented in the image above the writing at the bottom of the image.

Fourth: Change the size of the image (RIZE) again for the necessary dimensions for each algorithm, as each algorithm of the features extraction (variables extraction) method works on the size of a different image from the other, and therefore this step is a kind of pre preparation for each image before inserting them into the algorithms to reduce dimensions or choose features.

Fifth: Performing data Agmantation for each image, which means rotating the image at an angle of 90, 180, 270, so that for each image we have four images, and therefore each patient will have 8 images (because we dealt with two images for each patient, a front image and a side image of the tumor-infected breast).

Sixth: Performing the Falttn for the image, which means converting the image from a matrix to a vector consisting of n values, because the algorithms for dimensions reduction or features extraction deal with data in the form of a vector, not a matrix.

Seventh: Reshaping the two images (Reshape), the aim of which is to combine the values of the first vector resulting from the above step for the first image with the values of the second vector resulting from the above step for the second image, as the resulting vector has a dimension of 1×280 for each patient, but when it is approved for all patients It will be 280×70 in dimension.

Eighth: Introducing vectors and matrices resulting from the above step into the dimension reduction algorithms (features extraction), as three algorithms were relied upon to extract features, namely:

- PCA
- KPCA
- Fast ICA

Ninth: Store the data extracted from the above step in the drive files, and then enter the extracted variables into machine learning algorithms, in which three algorithms were used:

- KNN
- Decision Tree
- Random forest

Tenth: After implementing the above steps, 70 predictive values for survival times will be obtained for each patient within the study, based on 9 different combinations of dimension reduction algorithms with machine learning algorithms, noting that the comparison between the different results was made based on two criteria for comparison: MSE, C-Index.

Note that the comparison of results for estimating survival times according to the two criteria is shown in the two tables below.

Table (2): Results of applying machine learning algorithms to real data about survival times based on MSE (a), C-Index (b)

	PCA	KPCA	Fast ICA
KNN	0.0257	0.02087	0.016623
Decision Tree	0.0179	0.01794	0.012387
Random forest	0.0109	0.01224	0.006375

(a)

	PCA	KPCA	Fast ICA
KNN	0.5628	0.61010	0.751181
Decision Tree	0.7156	0.67045	0.78759
Random forest	0.7823	0.69510	0.85506

(b)

Eleven: Applying all of the above and including it within the non-parametric Kaplan-Meier model for estimating survival. Drawing the survival function for patients based on the methods and steps described above and as shown in the results below.

Table (3): Results of applying machine learning algorithms to data using Kaplan-Meier survival method based on MSE (a), C-Index (b).

	PCA	KPCA	Fast ICA
KNN	0.136143	0.11909	0.06485
Decision Tree	0.0665	0.09618	0.0434
Random forest	0.0562	0.083607	0.0236

(a)

	PCA	KPCA	Fast ICA
KNN	0.5628	0.61025	0.75119
Decision Tree	0.7439	0.6688	0.8235
Random forest	0.8079	0.70893	0.8928

(b)

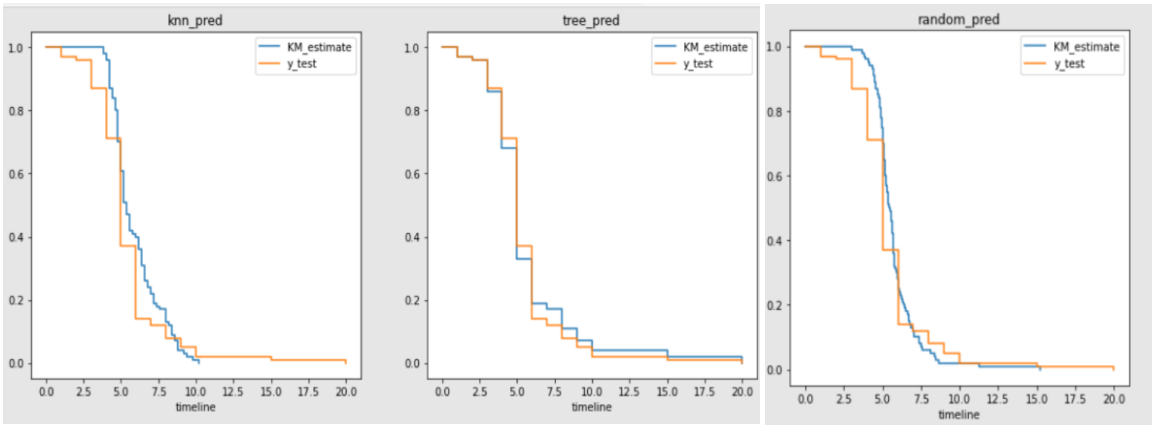


Figure (2): Principal component algorithm with all machine learning algorithms

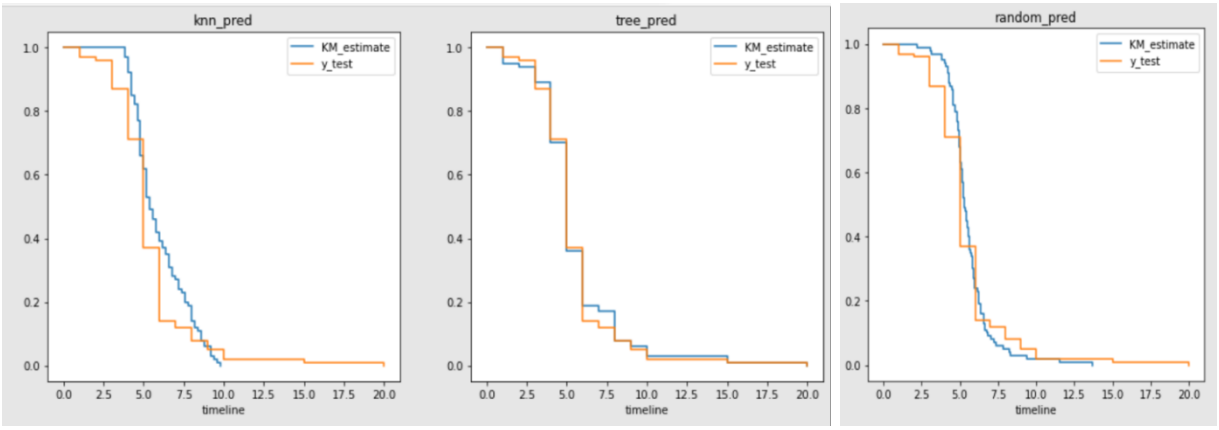


Figure (3): Kernel principal component algorithm with all machine learning algorithms

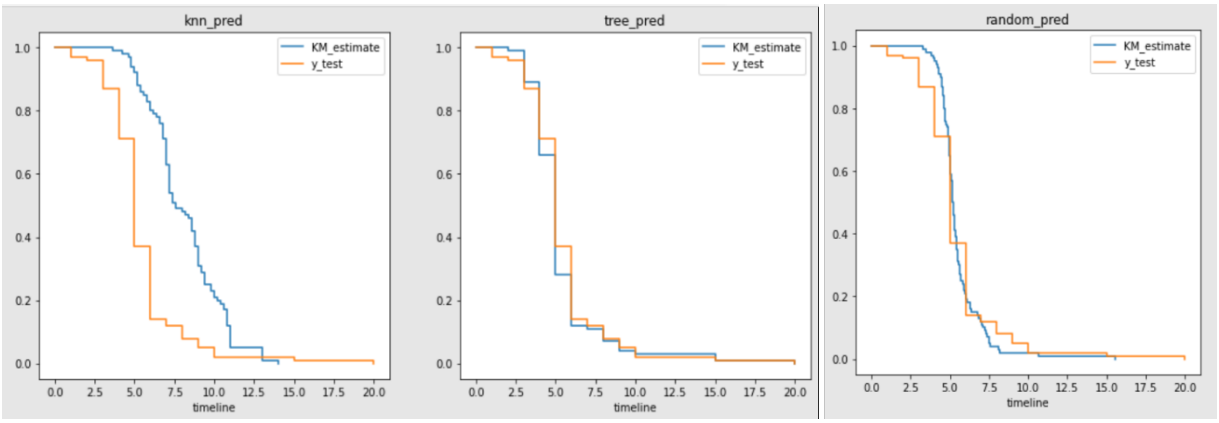


Figure (4): Fast ICA algorithm with all machine learning algorithms

4. Conclusion

According to the study sample, the results of survival estimation by using three feature extraction algorithms (selection of important features) which are the principal component algorithm, the kernel principal component algorithm and the fast ICA algorithm with three machine learning algorithms which are the K nearest neighbor algorithm, the decision tree and the random survival forests. The result showed that the best model for predicting survival for patients with Breast cancer is the model resulting from the application of a fast ICA algorithm to select important variables from medical images with the random survival forests algorithm, which gave the lowest value to the MSE criterion and the largest value to the C-Index criterion as well. This is illustrated by Figures 1,2,3 above. Accordingly, it is possible for the health institutions in Baghdad to adopt the research model to diagnose the most serious cases, and accordingly, based on the entered image of the breast cancer patient, it is possible to make a preliminary diagnosis of the level of severity of the disease and thus predict the survival of the patient. Thus, to increase the accuracy of the estimation results, it is possible to adopt the same algorithms that were used in this research paper, but with larger sample sizes, as it is known that the greater the sample size, the greater the accuracy of prediction and estimation.
Future studies can also adopt the same algorithms, but for other types of diseases.

References

1. Abdulwahhb, O.A., Abd Alrazak, M.S. (2020) "Using nonlinear dimensionality reduction techniques in big data analysis", Periodicals of Engineering and Natural Sciences ,Vol. 8, No. 1, pp.142-155.

2. Al-Rawi, A.G., Mohammed, A.M.,(2019) "Split and Merge Regions of Satellite Images using the Non-Hierarchical Algorithm of Cluster Analysis" Journal of Economics and Administrative Sciences 2019; Vol. 25, No.111 Pages: 466- 484.

3. Benlin, X., Fangfang, L., Xingliang, M., Huazhong, J.(2008) "STUDY ON INDEPENDENT COMPONENT ANALYSIS' APPLICATION IN CLASSIFICATION AND CHANGE DETECTION OF MULTISPECTRAL IMAGES", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B7.

4. Scholkopf B., Smola, A., & Muller, K. R.,(1996)," Nonlinear Component Analysis as a Kernel Eigenvalue Problem ", DBLP, Vol. 10, No. 44, PP1299 – 1319.

5. Fadnavis, R.A.(2019),"Application of Machine Learning For Survival Analysis- A Review" , IOSR Journal of Engineering (IOSRJEN), Vol.09, Issue5, pp 56-60.

6. Imandoust, S.B., Bolandraftar, M.(2013) "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background", Journal of Engineering Research and Applications, Vol. 3, Issue 5, pp.605-610.

7. Isltwaran, H., Bogalur, U.B., Blackstone, E.H., Lauer, M.S.(2008) "RANDOM SURVIVAL FORESTS" , The Annals of Applied Statistics, Vol.2, No.3, pp 841-860.

8. Jabir, A.G., Eesa, A. M. (2019) " Use Principal Component Analysis Technique to Dimensionality Reduction to Multi Source" , Journal of Economics and Administrative Sciences 2019; Vol. 25, No.115 Pages: 464- 473.

9. Jabir, A.G., Mohammed, M.A.(2020) "Using multidimensional scaling technique in image dimension reduction for satellite image" Periodicals of Engineering and Natural Sciences ISSN 2303-4521 Vol. 8, No. 1, March 2020, pp.447-454.

10. Nemati, M., Ansary, J., Nemati , N. (2020) , "Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data" , Patterns 1, 100074, August 14, 2020 (http://creativecommons.org/licenses/by-nc-nd/4.0/).

11. Sishi, M., Telukdarie, A. (2021) " The Application of Decision Tree Regression to Optimize Business Processes" Proceedings of the International Conference on Industrial Engineering and Operations Management Sao Paulo, Brazil, pp: 48-57.

12. Yaqoob, A.A., Ali, O.A.(2021) "DYNAMIC MODELING FOR DISCRETE SURVIVAL DATA BY USING ARTIFICIAL NEURAL NETWORKS AND ITERATIVELY WEIGHTED KALMAN FILTER SMOOTHING WITH COMPARISON" , Int. J. Agricult. Stat. Sci. Vol. 17, Supplement 1, pp. 1265-1274,.

13. Yaqoob, A.A., Ali, O.A.(2021), "DYNAMIC MODELING OF TIME-VARYING ESTIMATION FOR DISCRETE SURVIVAL ANALYSIS FOR DIALYSIS PATIENTS IN BASRAH, IRAQ" , Int. J. Agricult. Stat. Sci. Vol. 17, Supplement 1, pp. 1323-1332,.

14. Zupan, B., DemšAr, J., Kattan, M. W., Beck, J. R., & Bratko, I. (2000). "Machine learning for survival analysis: a case study on recurrence of prostate cancer". Artificial intelligence in medicine, 20(1), 59-75.

تقنيات التعلم الالي لتحليل بيانات البقاء لمرضى سرطان الثدي في بغداد

نور اياد محمد
جامعة بغداد/ كلية الإدارة والاقتصاد/ قسم الاحصاء
بغداد، العراق
noor.mohammed1101@coadec.uobaghdad.edu.iq

انتصار عريبي فدعم
جامعة بغداد/ كلية الإدارة والاقتصاد/ قسم الاحصاء
بغداد، العراق
entsar_arebe@coadec.uobaghdad.edu.iq

Received: 23/5/2023 **Accepted: 16/7/2023** **Published: 1/9/ 2023**

مستخلص البحث:
ان لطرائق التعلم الالي والتي هي واحدة من اهم فروع الذكاء الاصطناعي الواعدة اهمية كبيرة في جميع العلوم كالهندسية والطبية وايضا في الاونة الاخيرة دخلت بشكل واسع في العلوم الاحصائية وفروعها المختلفة والتي منها تحليل او تقدير البقاء، اذ بالامكان اعتباره فرع جديد استعمل لتقدير البقاء وكان موازيا مع الطرائق المعلمية واللامعلمية وشبه المعلمية التي تستعمل بشكل واسع لتقدير البقاء في البحوث الاحصائية.

في هذا البحث تم التطرق الى تقدير البقاء بالاعتماد على الصور الطبية الخاصة بالمرضى المصابين بسرطان الثدي والذين يتلقون علاجهم في المستشفيات العراقية. تم الاعتماد على ثلاث طرائق لاخيار الميزات المهمة من الصور وهي طريقة المركبات الرئيسية الاعتيادية، طريقة المركبات الرئيسية اللبية وطريقة اسرع ICA كما تم ادخال الميزات المهمة المستخرجة في ثلاث خوارزميات للتعلم الالي وهي خوارزمية الجار الاقرب ، شجرة البقاء، وغابات البقاء العشوائية تم الاعتماد على معيارين لمقارنة الطرائق الافضل لتقدير البقاء هي معيار MSE, معيار C- Index وكان افضل انموذج لتقدير والتنبؤ بالبقاء هو استعمال خوارزمية اسرع ICA مع خوارزمية الغابات البقاء العشوائية والتي اعطت اقل مقدار الى MSE واعلى قيمة الى C-index.

وعليه فأن أفضل انموذج لتقدير وتوقع البقاء هو استخدام خوارزمية ICA الأسرع مع خوارزمية غابة البقاء العشوائية التي أعطت أقل قدر من MSE وأعلى قيمة c- index. وبناءً على ذلك، نوصي الأطباء والأخصائيين الطبيين في العراق باعتماد هذا الانموذج لتقدير بقاء مرضى سرطان الثدي.

نوع البحث: ورقة بحثية.

المصطلحات الرئيسة للبحث: طرائق اختيار الميزات، تقنيات التعلم الالي، الصور الطبية لجهاز الماموكرام، سرطان الثدي

*البحث مستل من اطروحة دكتوراه