Journal of Industrial and Systems Engineering
Vol. 15, No. 4, pp. 31-40
Autumn (November) 2023

JISE

Investigating the impact of nutrition and lifestyle on breast cancer: A data mining approach

Hana Nazarpourfard[1], Mohammad Mehdi Sepehri[1*], Roghaye Khasha[3]
[1]MSc in Industrial Engineering, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
[2]Professor, Faculty of Industrial and Systems Engineering, Tarbiat Modares University, Tehran, Iran
[3]Research Scholar, Center of Excellence in Healthcare Systems Engineering, Tarbiat Modares University, Tehran, Iran

Abstract

Background: Breast cancer (BC) is the most common cancer and one of the main causes of death among women. This study was conducted to investigate the relationship between BC and nutrition and lifestyle, as well as compare machine learning models in predicting this disease.

Methods: We designed a questionnaire related to nutrition and lifestyle with a nutritionist's guidance and provided them to 569 patients. After data gathering, we developed some machine-learning algorithms like logistic regression (LR), K-Nearest Neighbor (KNN), Decision tree (DT), and Support vector machine (SVM) classifiers. To make more accurate models, we used an oversampling method to avoid skewing the model due to the lack of balance in the target classes, a grid search method to adjust the model's hyperparameters and finally random forest to identify each variable's importance.

Results: The results of this research showed that the accuracy of the DT model was 0.95, SVM and LR were 0.93, and KNN was 0.86. The results indicated the better performance of DT among other models.

Conclusions: Our findings show that it is possible to predict the type of cancerous tumor with relatively high accuracy without using specific information about the tumor itself. In particular, in our study, the decision tree has shown better accuracy compared to other models.

Keywords: Breast cancer, Nutrition, lifestyle, Data mining, Classification

1. literature review

22.9% of women's cancers are related to breast cancer (BC), which has the highest malignancy rate among them. In 2018, more than 2 million new cases were found (1). In the United States, 12% of women suffer from this disease. For example, in 2017, more than 250,000 new BC cases were registered (2). BC is considered a threat to women's life because this disease is the leading cause of death in women (3). Using

* Corresponding Author
ISSN: 1735-8272,

information extracted from the PubMed database, the trend of research on breast cancer has been upward from 2011 to 2020, which indicates the importance of this disease for researchers in this field.

BC is a heterogeneous disease including several distinct entities that have clinical behaviors (4). Some studies have shown the effect of nutrition and lifestyle on this disease. According to a survey conducted by Chlebowski (2013), engaging in physical activity as part of one's lifestyle is considered one of the most essential factors in reducing the risk of breast cancer and its recurrence (5). Kruk (2014) has concluded that postmenopausal women who engage in regular moderate-to-vigorous physical activity are likely to have a lower risk of breast cancer, while those who consume a high-fat diet, use combined estrogen and synthetic progestogen hormone therapy, or drink alcohol may have an increased risk. However, findings regarding the impact of smoking and psychological stress on breast cancer risk are currently contradictory (6). Ferrini et al (2015) have mentioned soy products as preventive factors against cancer and concluded that proper nutrition and diet could reduce cancer (7). Dieli-Conwright et al (2016) and Bathaee et al., (2023) have pointed to a suitable diet, food supplements and nutrition to reduce breast cancer recurrence (8). Wang (2017) emphasized the importance of early-stage breast cancer detection in reducing breast cancer death rates over the long-term. To achieve a good prognosis, it is critical to identify cancer cells at an early stage. While researchers have explored various diagnostic approaches for breast cancer, including mammography, magnetic resonance imaging, ultrasound, computerized tomography, positron emission tomography, and biopsy, these techniques have limitations. They can be expensive, time-consuming, or even inappropriate for young women. Therefore, there is an urgent need to develop highly sensitive and rapid early-stage breast cancer diagnostic methods (9). Seiler et al (2018) have reported that in both pre-and postmenopausal women, obesity raises the risk of breast cancer and can also have negative effects on disease recurrence and survival rates.

Unhealthy eating habits marked by a high intake of refined starches, sugar, and saturated and trans-saturated fats, coupled with a low intake of fibre, omega-3 fatty acids, and natural antioxidants are associated with inflammation and appear to increase the risk of breast cancer and mortality (10). Ghosn et al. (2020) found a protective relationship between a healthy lifestyle style (HLS) and breast cancer risk, particularly in postmenopausal women. This study suggested that adopting a healthy lifestyle, which includes a balanced diet, regular physical activity, and avoidance of smoking, can help reduce the incidence of breast cancer in the community (11).

The classification of cancer patients into high-risk or low-risk groups is an important step for treatment, and many research teams from different fields have been studying the application of machine learning (ML) methods to achieve this goal (12). For example, Chaurasia et al (2018) have used three algorithms - Naïve Bayes, RBF Network, and J48 - while taking into account the characteristics of breast cancer cells to predict whether they were benign or malignant. The results of the study demonstrated that the accuracy of J48, RBF Network, and Naïve Bayes were 93.41%, 96.77%, and 97.36%, respectively (13).

A review by Lee et al (2021) showed that the most common ML methods used in this field were decision trees, artificial neural networks, support vector machines (14). The study by Shanbezadeh et al (2022) and Movahed et al., (2023) was conducted on 1052 samples. In this study, the data mining (DM) process was performed using selected algorithms, including J-48 and decision tree random forest (RF), multi-layer perceptron (MLP), Naïve Bayes (NB), Ada boost (AB), and logistic regression (LR). The evaluation results of different DM algorithms showed that the J-48 DT algorithm had the best performance (AUC = 0.922), followed by the AB, MLP, LR, and RF algorithms (AUC: 0.899, 0.819, 0.716, and 0, respectively) (15).

Previous studies have highlighted the role of nutrition and lifestyle in breast cancer, as well as utilizing machine learning techniques to predict it based on tumor information. However, the relationship between

nutrition/lifestyle and benign/malignant breast cancer has received less attention. In this study, we tried to use other information such as education level, monthly income and other information listed in Table 2 in addition to nutritional information and people's mobility, and check its relationship in predicting the type of cancer cell. Our research aims to determine whether information about nutrition and lifestyle can effectively distinguish between benign and malignant breast cancer. Additionally, we seek to compare models based on an individual classifier.

2. Methods

This study includes 4 main steps, each of which is briefly explained in Figure 1.

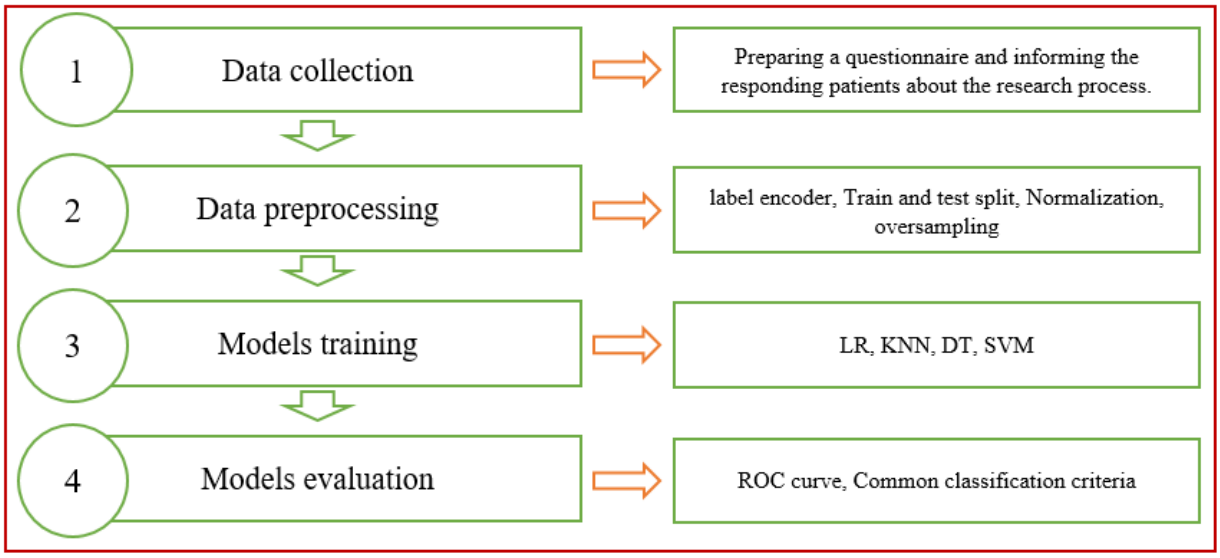

Figure 1. The main steps of research framework

2.1. Data collection

With a nutritionist's help, we designed a questionnaire related to nutrition and lifestyle and provided them to 569 patients with breast cancer. All the patients were informed about the study and answered the questions with their consent. This questionnaire started in January 2022 and lasted for 2 months. The education level and marital status of survey respondents are explored in Table 1 through statistical information.

Table 1: The statistical information about some categorical attributes in the dataset

Feature	Category	Percentage
Type of cancer	Benign	62.74%
	Malignant	37.26%
Education level	High School	27.94%
	Bachelor	36.56%
	Master's Degree and Higher	35.50%
Marital status	Single	38.31%
	Married	31.11%

	Divorced	30.58%

We classified the questionnaire questions into three categories; the first category is questions related to demographic variables, the second category is questions related to control variables, and finally, non-control variables. We did not use any condition to select the respondents and selected them randomly. Since this disease is more prevalent among women, all our participants were women. 62.74% of these people had benign cancer; the rest were malignant. The weight range of 70 to 90 kg was the most frequent among the participants. We waited long to collect the data because we asked the respondents to answer all the questions completely and accurately. Table 2 shows the research variables in detail.

Table 2. Description of features

Category	**Variables**	**Description**
Demographic	Education level	Master's degree or higher = A
		Bachelor and below = B
		High school = C
	Socio-economic	Monthly income
	Marital State	Single = A
		Married = B
		Divorce = C
	Employee status	Yes = 1
		No = 0
Control variables	Weight	Continuous
	Sport (exercise)	Continuous and regular = A
		No = B
		Once in a while = C
	Kind of meat	Red Meat = A
		Fish = B
		Chicken = C
	Kind of Vegetable	No = A
		Continuous and regular = B
		Once in a while = C
	Number of children	Integer
	Oral contraceptive use (OCU)	Yes = 1
		No = 0
	Duration of oral Contraceptive use (DOCU)	By day
	Age at last pregnancy	Integer
	Duration of breastfeeding (DOB)	By month
	Spontaneous abortions	Yes = 1
		No = 0
Non-control variables	Age at menarche	Integer
	Age at last pregnancy	Integer

	Height	Integer (cm)
	Work connected radiation with (radiation)	Yes = 1
		No = 0
Target	Diagnosis	Malignant = 1
		Benign = 0

2.2. Data preprocessing

Due to our sensitivity, we asked all the respondents to answer completely. Therefore, our data did not have duplicate or missing value. Features with outliers were removed by interquartile range (IQR) method. IQR is a measure of variability in a dataset. It is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of the data. The IQR is useful for identifying outliers because it excludes extreme values that fall outside the middle 50% of the data (16).

We considered the data training and testing with a ratio of 80 and 20, respectively. The majority of the data (68%) belonged to class 0, while a minority (32%) belonged to class 1. Therefore, we used oversampling to avoid model bias due to an imbalance in the target classes. The Synthetic Minority Oversampling Technique (SMOTE) is suitable for imbalanced data. It has significantly contributed to new supervised learning paradigms, including multilabel classification (17). We had several categorical features in our data, so we used one-hot encoding to handle them. One-hot encoding is a technique used in data preprocessing to convert categorical variables into a numerical representation that can be used for machine learning algorithms. In one-hot encoding, each category or level of a categorical feature is converted into a binary vector of 0s and 1s. The length of the binary vector is equal to the number of categories in the feature. Each category is then represented by a vector with a single 1 in the position corresponding to its index in the original list of categories, and all other positions filled with 0s (18). We rescaled data to a range of [0,1] based on equation 1 for the models that predict through feature distance.

$$W_i = \frac{W_i - min(W_i)}{(max(W_i) - min(W_i))} \quad (1)$$

Where, W_i is the ith characteristic, and $min(W_i)$ and $max(W_i)$ are the minimum and maximum values of W_i, respectively.

2.3. Creating machine learning models

A machine learning mechanism is like a human that learns from its experiences to use them to make decisions. The difference is that machine learning does this through computational methods and is expandable (19). The most common types of machine learning based on the method are supervised, unsupervised, and semi-supervised learning (20). Supervised learning involves learning the correspondence between a set of input variables X and output variable Y and using this correspondence to predict outputs from unseen data (21). Therefore, our problem is the classification of supervised learning. The decision tree (DT) classification algorithm can be done in serial or parallel steps according to the algorithm's efficiency. The serial tree is built through the training data and helps predict the value of the target variable using predictor variables. The architecture of the decision tree is such that the highest node is named as the root, other nodes are called internal nodes, and the lower nodes are designated as terminal nodes (without output links) (22). One of the criteria used to determine the optimal direction for dividing records is the entropy criterion, according to equation 2.

$$E(S) = \sum_{i=1}^{c} -\rho_i \, log_2 \, p_i \quad (2)$$

If the attribute takes on c different values, then the entropy S is related to c-wise classification. ρ_i is the ratio of S belonging to class i.

Logistic regression (LR) is a supervised machine learning algorithm. It is used to predict the probability of certain classes based on some dependent variables. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or interval type. LR is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type (23). The k-nearest neighbor (KNN) algorithm is a supervised learning classifier based on the closest training examples in the problem space. KNN is a type of instance-based learning or lazy learning where the function is only approximated locally (24). The KNN classifier fundamentally relies on a distance metric. The better that metric reflects label similarity, the better the classified will be. The most common choice is the Minkowski distance based on equation 3.

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \quad (3)$$

Minkowski distance is typically used with p being 1 or 2, which correspond to the Manhattan distance and the Euclidean distance, respectively.

The support vector machine (SVM) is a mathematical function that can be linear and non-linear and can distinguish two objects. One of the advantages of using this classifier is that it is still effective if the number of dimensions is more than the number of samples (24).

It is worth mentioning that in this study, machine learning models were implemented by Python libraries of Scikit-learn, Pandas, and Numpy. Regarding the hardware, our used CPU is an Intel i7 2.60 GHz with 8 GB installed memory.

3. Results

After adjusting the hyperparameters of each model built through the Grid search method, we compared the models. Grid search is a traditional method for setting and optimizing hyperparameters. This method performs a simple search on a specific subset of the hyperparameter space of the training algorithm (25). The results of this research showed us that the accuracy of each of the models is such the DT is 0.95, SVM and LR are 0.93, and KNN is 0.86, as well as other evaluation criteria of the models in Table 3, are located.

Table 3. Models' evaluation

Models	Accuracy	Specificity	Recall	Precision	F-score
DT	**0.954**	**0.969**	**0.939**	**0.968**	**0.953**
SVM	0.931	1.0	0.876	1.0	0.934
LR	0.931	1.0	0.876	1.0	0.934
KNN	0.863	0.916	0.819	0.921	0.867

The receiver operating characteristic (ROC) curve is a plot of the true positive rate/sensitivity (y-axis) versus the false positive rate/1 specificity (x-axis) for candidate threshold values between 0 and 1. The area under the ROC curve is known as the AUC and is a valuable separability performance metric. It shows the capability of distinguishing between classes. The higher AUC indicates a better prediction model performance (26). We evaluated the AUC criteria for the models mentioned in Figure 2.

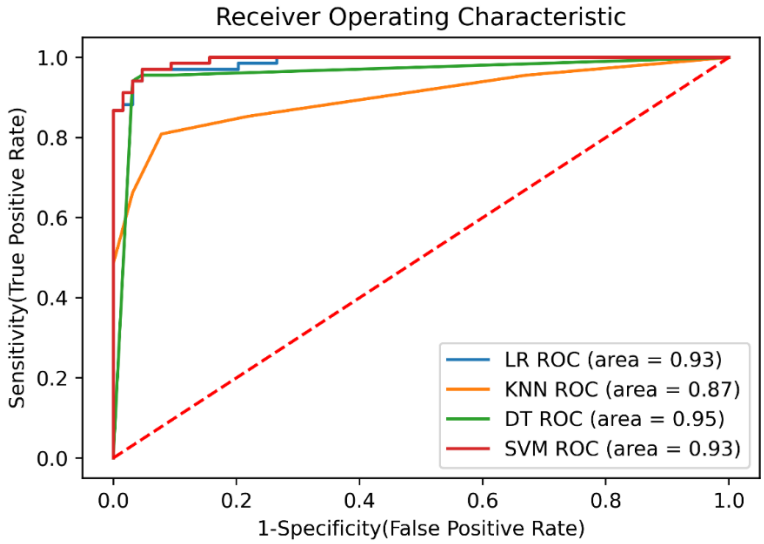

Figure 2. ROC curve.

Feature selection is a necessary stage of data analysis for applying to select a small set of relevant features. Also, the Random Forest classifier is an instrumental base for the wrapper algorithms solving all relevant problems because it provides the variable importance measure (27). We used Random Forest Feature selection to examine the set of features that have the most impact (Figure 3).

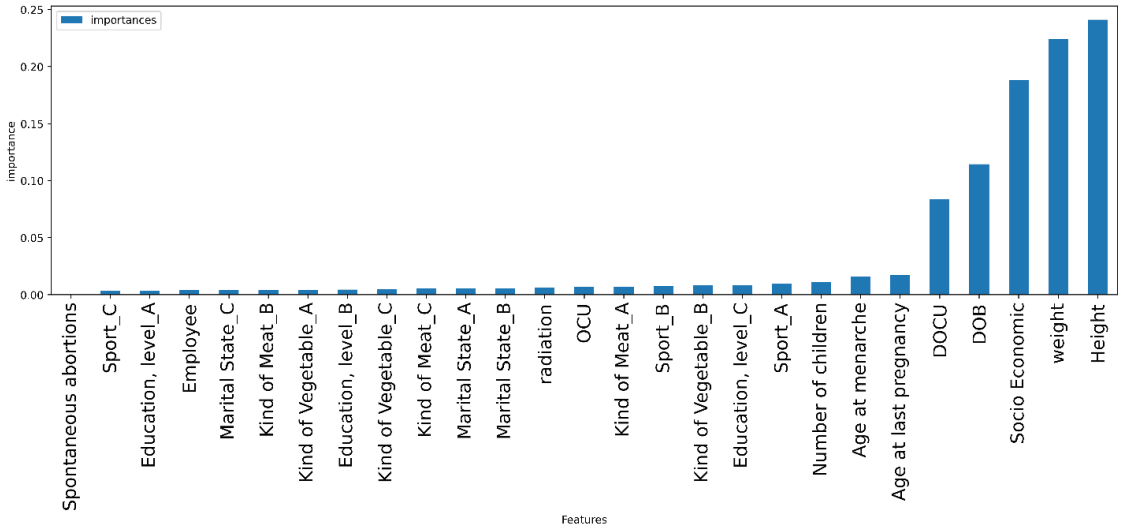

Figure 3. Impact of features

The results presented in Figure 3 indicate that economic conditions, along with factors such as weight, height, breastfeeding, and the use of oral Contraceptive have a significant impact on the type of breast

cancer (benign or malignant). In this review, comparing the typical individual classifier was the priority of our research. The results indicate the better performance of DT among other models.

4. Discussion

This study showed us that we could determine with relatively good accuracy whether a breast cancer cell is benign or malignant through the type of nutrition and lifestyle. Our research differs from previous studies in terms of the data collected. We used a questionnaire that included three categories of data for prediction, with the economy having the greatest impact. The second category consisted of control variables affecting weight, breastfeeding, and the use of oral contraceptives while the third category comprised non-control variables that had the most significant effect on height. Unlike other studies, we did not incorporate information related to cancer tumors in our predictions of malignancy and benignity.

Previous research has stated that obesity and being overweight are the leading causes of cancer. In addition to breast cancer, it is directly related to other diseases such as diabetes, high blood pressure, and stroke (28). Our study also indicates the importance of weight in cancer malignancy. A necessary experience that we gained from this study was the effect of setting hyperparameters, which had a significant impact on the results. This study has several limitations. First, the collected information is related to the climate and the country of Iran in the city of Tehran. Second, we didn't consider variables such as the weather, the amount of sleep people has 24 hours a day, bad personal habits such as drinking alcohol or smoking, and liquid consumption, including coffee, herbal teas, or tea because we did not have the data about these parameters available.

5. Conclusions

This study aimed to investigate the impact of nutrition and lifestyle on breast cancer. While previous research has recognized the significance of nutrition in cancer, it has not been utilized to predict the type of cancer. However, our findings reveal that we can predict the type of cancerous tumor with relatively high accuracy without using specific information about the tumor itself. In particular, the decision tree model was shown to provide better accuracy compared to other models. Due to the excellent performance of the decision tree, we can use ensemble models for prediction in the future in addition to considering more variables.

References

Akram M, Iqbal M, Daniyal M, Khan AU. Awareness and current knowledge of breast cancer. Biological research. 2017;50:1-23.

Alsagheer RH, Alharan AF, Al-Haboobi AS. Popular decision tree algorithms of data mining techniques: a review. International Journal of Computer Science and Mobile Computing. 2017;6(6):133-42.

Argolo DF, Hudis CA, Iyengar NM. The Impact of Obesity on Breast Cancer. Current Oncology Reports. 2018;20(6):47.

Ayodele TO. Types of machine learning algorithms. New advances in machine learning. 2010;3:19-48.

Bathaee, M., Nozari, H., & Szmelter-Jarosz, A. (2023). Designing a new location-allocation and routing model with simultaneous pick-up and delivery in a closed-loop supply chain network under uncertainty. Logistics, 7(1), 3.

Bisong E. Logistic Regression. In: Bisong E, editor. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. Berkeley, CA: Apress; 2019. p. 243-50.

Chaurasia V, Pal S, Tiwari B. Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms & Computational Technology. 2018;12(2):119-26.

Chlebowski RT. Nutrition and physical activity influence on breast cancer incidence and outcome. The Breast. 2013;22:S30-S7.

Cunningham P, Cord M, Delany SJ. Supervised Learning. In: Cord M, Cunningham P, editors. Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 21-49.

De Cicco P, Catani MV, Gasperi V, Sibilano M, Quaglietta M, Savini I. Nutrition and Breast Cancer: A Literature Review on Prevention, Treatment and Recurrence. Nutrients. 2019;11(7):1514.

Dieli-Conwright CM, Lee K, Kiwata JL. Reducing the Risk of Breast Cancer Recurrence: an Evaluation of the Effects and Mechanisms of Diet and Exercise. Current Breast Cancer Reports. 2016;8(3):139-50.

Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research. 2018;61:863-905.

Ferrini K, Ghelfi F, Mannucci R, Titta L. Lifestyle, nutrition and breast cancer: facts and presumptions for consideration. Ecancermedicalscience. 2015;9.

Ghosn B, Benisi-Kohansal S, Ebrahimpour-Koujan S, Azadbakht L, Esmaillzadeh A. Association between healthy lifestyle score and breast cancer. Nutrition Journal. 2020;19(1):4.

Gonçalves L, Subtil A, Oliveira MR, de Zea Bermudez P. ROC curve estimation: An overview. REVSTAT-Statistical journal. 2014;12(1):1–20-1–.

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal. 2015;13:8-17.

Kruk J. Lifestyle components and primary breast cancer prevention. Asian Pac J Cancer Prev. 2014;15(24):10543-55.

Kursa MB, Rudnicki WR. The all relevant feature selection using random forest. arXiv preprint arXiv:11065112. 2011.

Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, et al. Predicting breast cancer 5-year survival using machine learning: A systematic review. PloS one. 2021;16(4):e0250370.

Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. arXiv preprint arXiv:191206059. 2019.

Movahed, A. B., Aliahmadi, A., Parsanejad, M., & Nozari, H. (2023). A systematic review of collaboration in supply chain 4.0 with meta-synthesis method. Supply Chain Analytics, 100052.

Raikwal J, Saxena K. Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. International Journal of Computer Applications. 2012;50(14).

Seger C. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. 2018.

Seiler A, Chen MA, Brown RL, Fagundes CP. Obesity, Dietary Factors, Nutrition, and Breast Cancer Risk. Current Breast Cancer Reports. 2018;10(1):14-27.

Shanbehzadeh M, Nopour R, Erfannia L, Amraei M. Comparing Data Mining Algorithms for Breast Cancer Diagnosis. Shiraz E Medical Journal. 2022.

Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321(3):288-300.

Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC medical research methodology. 2014;14:1-13.

Wang L. Early Diagnosis of Breast Cancer. Sensors. 2017;17(7):1572.

Weigelt B, Geyer FC, Reis-Filho JS. Histological types of breast cancer: How special are they? Molecular Oncology. 2010;4(3):192-208.

Zhou Z-H. Machine learning: Springer Nature; 2021.