Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/140513

Machine Learning for Breast Cancer Classification Using K-Star Algorithm

Mohamed Sakr[1], *Abeer Saber*[2*], *Osama M. Abo-Seida*[2] *and Arabi Keshk*[1]

[1]Department of Computer Science, Faculty of Computers and Information, Menoufia University, Menoufia 32511, Egypt
[2]Department of Computer Science, Faculty of Computers and Information, Kafr El-Sheikh University, Kafr El-Sheikh 33511, Egypt

Received: 12 Mar.. 2020, Revised: 22 May 2020, Accepted: 18 Jun. 2020
Published online: 1 Sep. 2020

Abstract: Early disease detection and prevention play a very significant role in reducing deaths as well as the cost of healthcare. It was found that 8% of women were diagnosed with Breast Cancer (BC) throughout their life. BC is characterized by gene mutation, constant pain, as well as changes in size, color (redness), and breast skin texture. Machine Learning (ML) technologies play an important role in diagnosing and predicting the prognosis of BC. Also, it helps in recognizing people with BC, distinguish benign from malignant tumors using classification techniques. In the current study, we apply four various classifier algorithms: K-star, Naïve Bayes (NB), Clonal Selection Algorithm (CLONALG), and Artificial Immune Recognition System (AIRS) for BC classification model. The two algorithms were evaluated through a series of experiments over real datasets. We chose five metrics to evaluate performance of the applied algorithms, i.e. accuracy, precision, sensitivity, specificity, and Area Under the ROC Curve (AUC). The results showed that the K-star algorithm has better results than the old ones. Also, experiments indicated that the K-star algorithm provides the highest accuracy, sensitivity, specificity, precision, and AUC with 97.142%, 100.00%, 95.24%, 93.3%, and 0.998, respectively.

Keywords: Breast Cancer - Data Mining – Feature Extraction – Machine Learning

1 Introduction

Several people die annually because of preventable death. Approximately, 56 million people died worldwide in 2012, and two-thirds died of non-communicable diseases, such as cancer and diabetes, as well as cardiovascular insufficiency [1]. However, this figure can reduce at least to half through access to affordable interventions. One of the major applications of Machine Learning (ML) is health systems improvement. ML uses the historical data collected from old patients and analyze them to identify the connection between diseases and their symptoms and treatments. BC is one of the crucial reasons for death among women in urbanized nations. Every year, at least "220.000" females in The USA are diagnosed with BC and more than "40.000" expecting to die [2]. BC is cancer that develops from breast tissue.Signs of BC may include a lump in the breast, a change in breast shape, dimpling of the skin, fluid coming from the nipple, a newly-inverted nipple, or a red or scaly patch of skin.

It has been established that a particular diagnosis in early detection reduces death rate from BC. Medical experts can make mistakes when recognizing a disease. Using technologies, such as ML, accuracy of the diagnosis reaches (91.1%), while it will be only (79.9%) if it is performed by a skilled doctor [3]. Using ML approaches improves our data analysis and cancer progression understanding. An appropriate validation level in clinical practice is required for such techniques [4, 5].

Classification is one of the ML categories which is a strategy of grouping data according to their features. Classification is one of the supervised ML algorithms that mean that it needs training data to build the model which will help in classifying the new upcoming data. First, the data is split into 2 parts: training data and test data. Training data is to develop a learning model, and testing data is to examine the learning model.

In this current study, a new model for BC classification based on different ML algorithms, such as

* Corresponding author e-mail: Abeer_Saber@fci.kfs.edu.eg

K-star, NB, CLONALG, and AIRS algorithms, is proposed. The best model for BC data classification is selected based on accuracy, precision, sensitivity, specificity, as well as AUC metrics evaluated from the confusion matrices.

This paper is organized address the previous pieces of literature as follows: sections 2 and 3 address the previous pieces of literature and a brief description of the problem of the BC classification and how to solve it using ML techniques are presented. Section 4 covers experimental results over real data. Conclusion is presented in Section 5.

2 Related Work

Lee et al. [6] implemented a molecular ML algorithm for sentence classification. Their algorithm could generalize from example sentences and do sentence classifications with 100% success. Chaurasia and Pal [7] compared three techniques for BC classification and declared that Sequential Minimal Optimization (SMO) has better classification accuracy than IBK and BF Tree.
PATRÍCIO et al. [3] used logistic regression (LR), random forests (RF), and support vector machines (SVM) algorithms for predicting the presence or absence of BC based on resistin, glucose, age, and BMI. The SVM algorithm achieved the best predicting results of specificity ranging from 85% to 90% and with sensitivity ranging from 82% to 88%. Agrawal et al. [8] described and implemented the Directed Bee Colony (DBC) algorithm for the classification of diabetes, cancer and heart disease. The classification results were compared with other bioinspired algorithms(algorithms inspired from biology [9]), and showed that DBC accuracy proved to be the second-best among the algorithms applied. Jain et al. [10] integrated CFS: Correlation-based Feature Selection techniques with an improved version of Binary Particle Swarm Optimization called iBPSO for cancer classification purpose. They compared their results with 7 other common methods and showed that their model provides the best results.
Lötsch et al. [11] introduced a symbolic rule-based classifier tool with an accuracy of 95%. It included 21 single or aggregated attributes, involving psychological features, demographic and pain-related arguments. Hoadley et al. [12] presented a new PanCancer Atlas integrative analysis using iCluster that identifies 28 different molecular subtypes that arise from the 33 different types of tumors analyzed on at least four different TCGA platforms.
Mahmood et al. [13] used the K-Star classifier algorithm for the Intrusion Detection System and that have achieved high accuracy of "99.47%" for traffics classification with NSL-KDD dataset. To define the presence of liver disease or not, Thangaraju et al. [14] developed Practical Swarm Optimization (PSO) with the K-Star algorithm and accuracy was 100%.

3 Material and Methods

3.1 Data Pre-processing

Data pre-processing is a necessary step for the classification process because most of the datasets are susceptive to noisy and inconsistent data due to their different sources. As shown in Fig. 1.
The first step in the pre-processing phase is to ensure that no missing data exists in the BC dataset. Then, we use the partition membership filter, which applies a Partition Generator for generating partition membership values. After applying this filter, the numerical values become nominal and are distributed in a selected number of containers equally. Finally, data get appropriate for classification.

3.2 Classification models

Data is divided into two parts; "70%" for the training the model and "30%", for testing the model. The k-star, NB, CLONALG, and AIRS algorithms are used in two ways within the dataset in the classification process. First, the classification model is implemented after applying the partition membership filter for all parameters. Then, two different feature variables are used to define the most factors affecting BC.

3.2.1 K-Star algorithm

K-Star or K * algorithm is an instance-based classifier that uses K Nearest Neighbour (KNN) method. It aims to divide n data points into k clusters. K* uses an entropic distance measure depending on the probability of transforming one instance into another. Using entropy is very important for an instance distanced and information theory helps calculate the distance among instances. Thus, entropic distance is used for retrieving the most similar instances from the data set [15].
New data points, n, are attached to the most expected class, yi, where i = 1…k. The K* equation is computed, as follows:

$$K^*(yi,n) = -lnP^*(yi,n), \tag{1}$$

The probability of x reaching to y through a random path is represented by P*. Advantages: It presents a consistent strategy for handling the real-valued attributes, symbolic attributes, and missing values. Limitations: It has a long training time.

3.2.2 NB algorithm

To predict the values of the features for other members, the data item is classified. Based on similarity, similar data items are grouped in the same classes. Bayesian classification can successfully predict other features values if the algorithm can detect the class. The Naive Bayes classifier is a supervised learning algorithm model that applies a simple theorem of probability called Bayes with the Naïve assumption of pairwise independence between the whole features. If y is a class variable and $x_1, x_2, ..., x_n$ are dependant feature vectors, then according to Bayes' Theorem, the below relationship follows:

$$P(y|x_1,x_2,...,x_n) = \frac{P(y)P(x_1,x_2,...,x_n|y)}{P(x_1,x_2,...,x_n)} \quad (2)$$

$$P(y|x_1,x_2,...,x_n) = \frac{P(y)\Pi_{i=1}^{n}P(x_i|y)}{P(x_1,x_2,...,x_n)} \quad (3)$$

The following rule of classification can be applied since $P(x_1,x_2,...,x_n)$ is constant.

$$P(y|x_1,x_2,...,x_n) \propto P(y)\Pi_{i=1}^{n}P(x_i|y) \quad (4)$$

Which leads to

$$\hat{y} = argmax_y P(y)\Pi_{i=1}^{n}P(x_i|y) \quad (5)$$

3.2.3 CLONALG algorithm

The Clonal Selection algorithm was proposed as a simulation of the CLONALG selection hypothesis of obtained immunity. It also represents the characteristics and behaviours of antibodies in the immune system. Its hypothesis suggests that when selecting B and T-cells (antigens for lymphocytes) and connecting them to a specific antigenic, each cell divides to make duplicates of itself and differentiates to form other cell types, such as memory cells or plasma [16]. Fig. 2. shows the flowchart of the CLONALG algorithm.

3.2.4 AIRS algorithm

AIRS is a supervised learning algorithm inspired by immune system metaphors using resource competition, clonal selection, affinity maturation, and affinity recognition balls (ARBs). It comprises five phases: Initialization, antigen training, competition for limited resources, memory cell selection, and classification [17]. The AIRS algorithm flowchart is shown in Fig. 3.

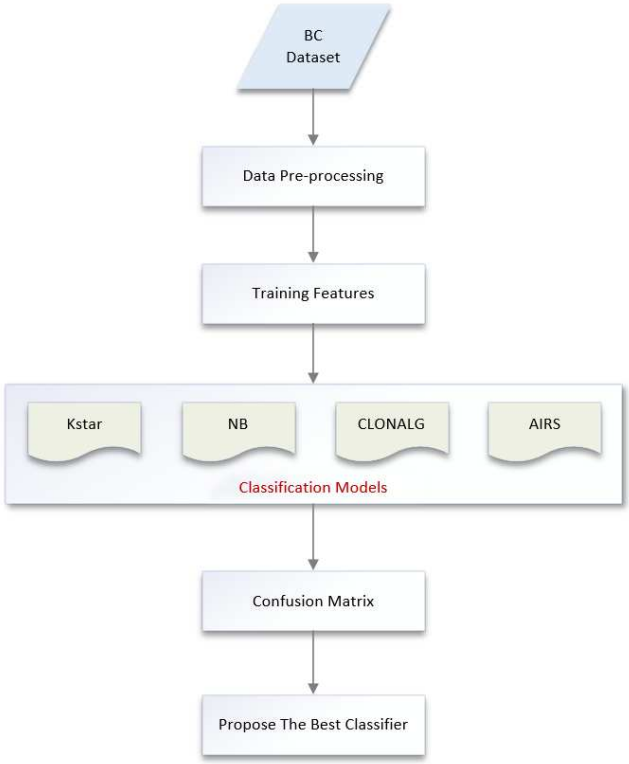

Fig. 1: The proposed model for breast cancer classification

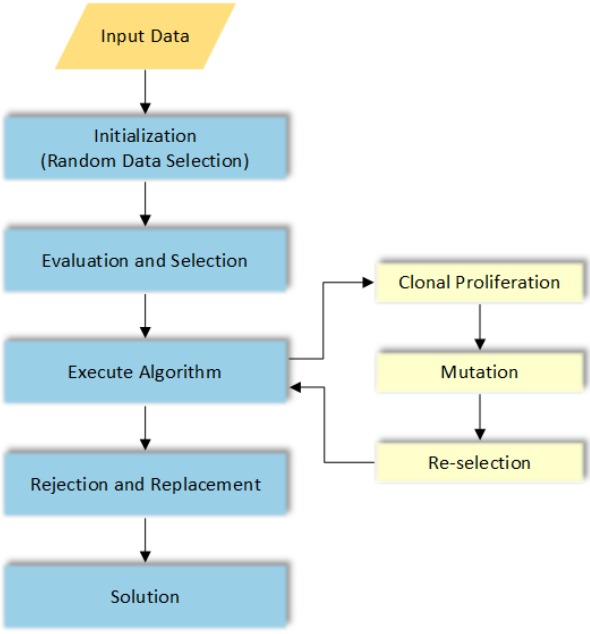

Fig. 2: Flow chart of Clonal Selection Algorithm

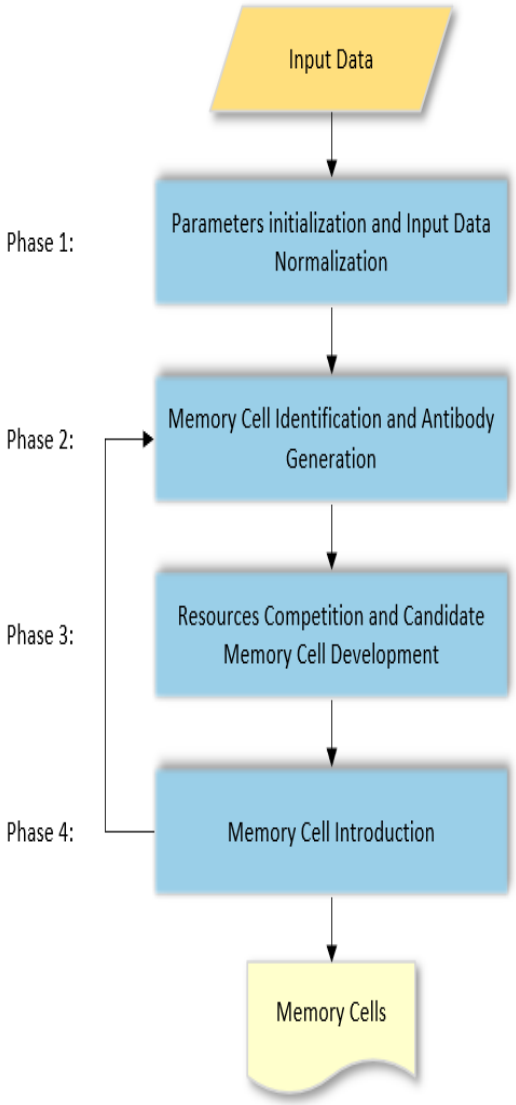

Fig. 3: Flowchart of the artificial immune recognition system algorithm

4 Results And Discussion

4.1 Dataset

Choosing a real dataset is necessary for building a robust ML model. In the proposed model, the used dataset is obtained from the "UCI" ML repository for BC disease [18] which contains 116 instances ("64" for BC patients and "52" for healthy controls). Typically, the dataset is divided into "70 %" for the training set and "30 %" for the testing set. Ten predictors indicate the presence or absence of BC. Those are anthropometric data and parameters that can be collected during routine blood tests. Collected data involves "Age, Body Mass Index (BMI), Glucose, Insulin, Homeostasis Model Assessment (HOMA), Leptin, Adiponectin, Resistin, and Chemokine Monocyte Chemoattractant Protein (MCP-1)" as shown in Table 1. The heatmap correlation matrix in Fig. 4 indicates that glucose is the most important predictor for BC. Fig. 5 shows the data representation for each value in the data set which indicates that the HOMA and MCP-1 are the best predictors for classification results.

Table 1: UCI breast cancer dataset description

Parameters	Measure unit	Range
Age	Years	24 .. 89
BMI	kg/m2	18.37 .. 38.5788
Glucose	mg/dL	60 .. 201
Insulin	μU/mL	2.432 .. 58.46
HOMA	(Glucose* Insulin) / 405	0.467409 .. 25.0503
Leptin	ng/mL	4.311 .. 90.28
Adiponectin	μg/mL	1.65602 .. 38.04
Resistin	ng/mL	3.21 .. 82.1
MCP-1	pg/dL	45.843 .. 1698.44

4.2 Results

In this section, to study the performance of K-star, NB, CLONALG, and AIRS algorithms, several experiments have been done over the training data. Here, K-star algorithm is compared with other algorithms before and after applying filtering in terms of accuracy, sensitivity, specificity, and AUC in case of splitting data to "70%" for training and "30%" for testing, as shown in Tables 2 and 3, respectively. Furthermore, Fig. 6 presents a comparison between the mentioned algorithms in their performance metrics: accuracy, sensitivity, specificity, precision, and AUC. Experiments are performed to analyse the performance between our work and the existing revised methods in Table 4. In Table 5, Confidence Intervals (CIs) with a "95%" CI have been measured in the test set for accuracy, sensitivity, and specificity values. For each classifier, predictive models were created considering more significant variables as predictors. Variables are, as follows: Glucose, Resistin, Age, BMI, HOMA, Leptin, Insulin, Adiponectin, and MCP-1 with variables v1 up to V9, respectively. The Table indicates that the K-star classifier algorithm has got the best results in accuracy, sensitivity, and specificity in almost all variables.

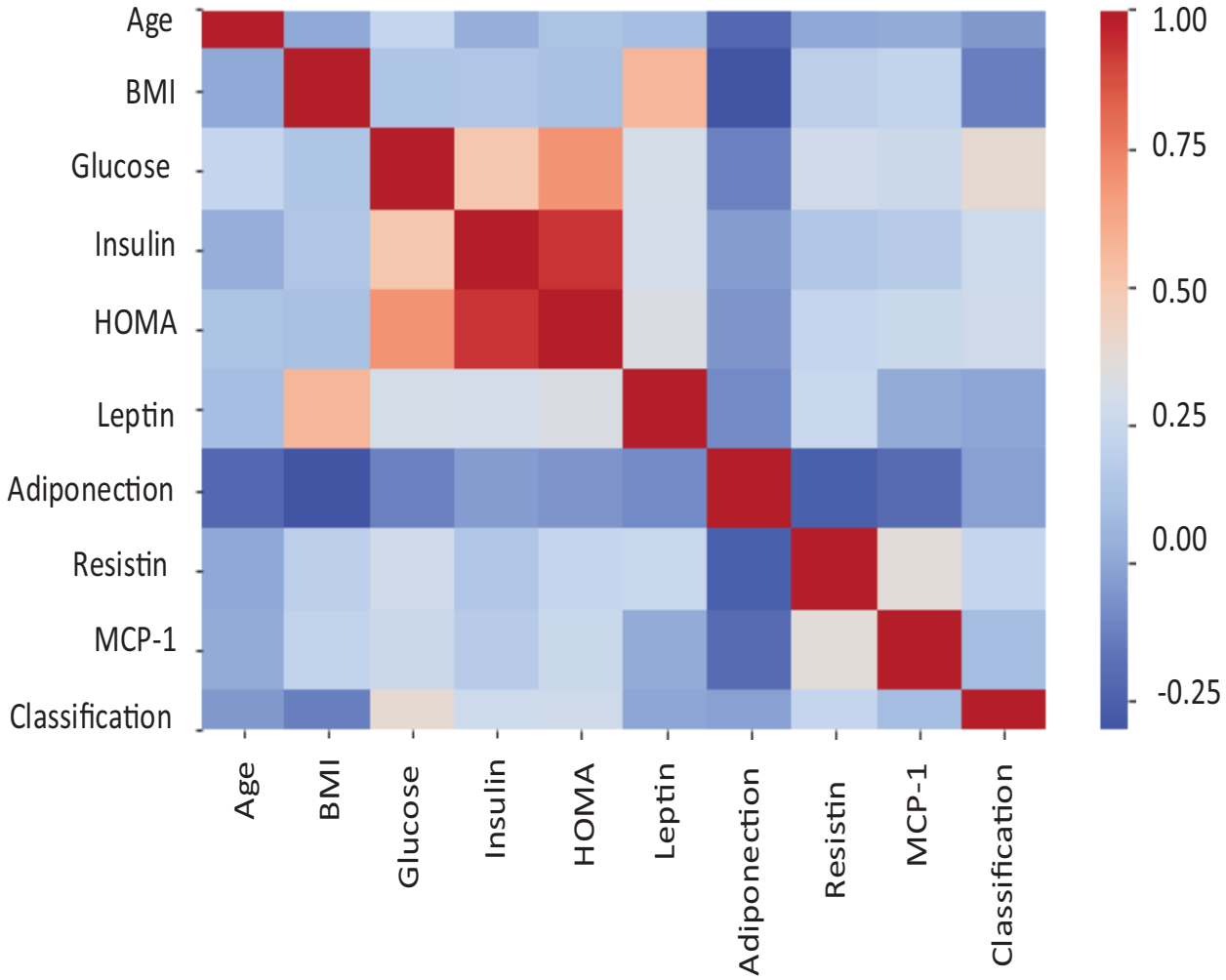

Fig. 4: A correlation heatmap matrix for the dataset parameters representation

Table 2: The overall classification performance before applying partition membership filtering .

Classifier	Classifier Performance				
	Accuracy(%)	**Sensitivity(%)**	**Specificity(%)**	**Precision**	**AUC**
K-star	94.93	100.00	91.67	0.88	0.995
NB	57.14	82.14	40.48	0.479	0.697
CLONALG	58.57	25.00	80.95	0.466	0.53
AIRS	72.85	82.14	66.67	0.62	0.74

5 Conclusion

Depending on the parameters of Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP-1, the presence and absence of BC could be indicated. A case study of BC from the UCI repository was used to test the discussed classification algorithms. The results of the experiments showed that the performance of K-star was of a high level compared to the NB, CLONAL, and AIRS classifiers. The classification with the K-star classifier would reach 97.14 % if data were pre-processed by partition membership filtering. Accuracy, sensitivity, and specificity of the presence of BC could be predicted between [86.01,

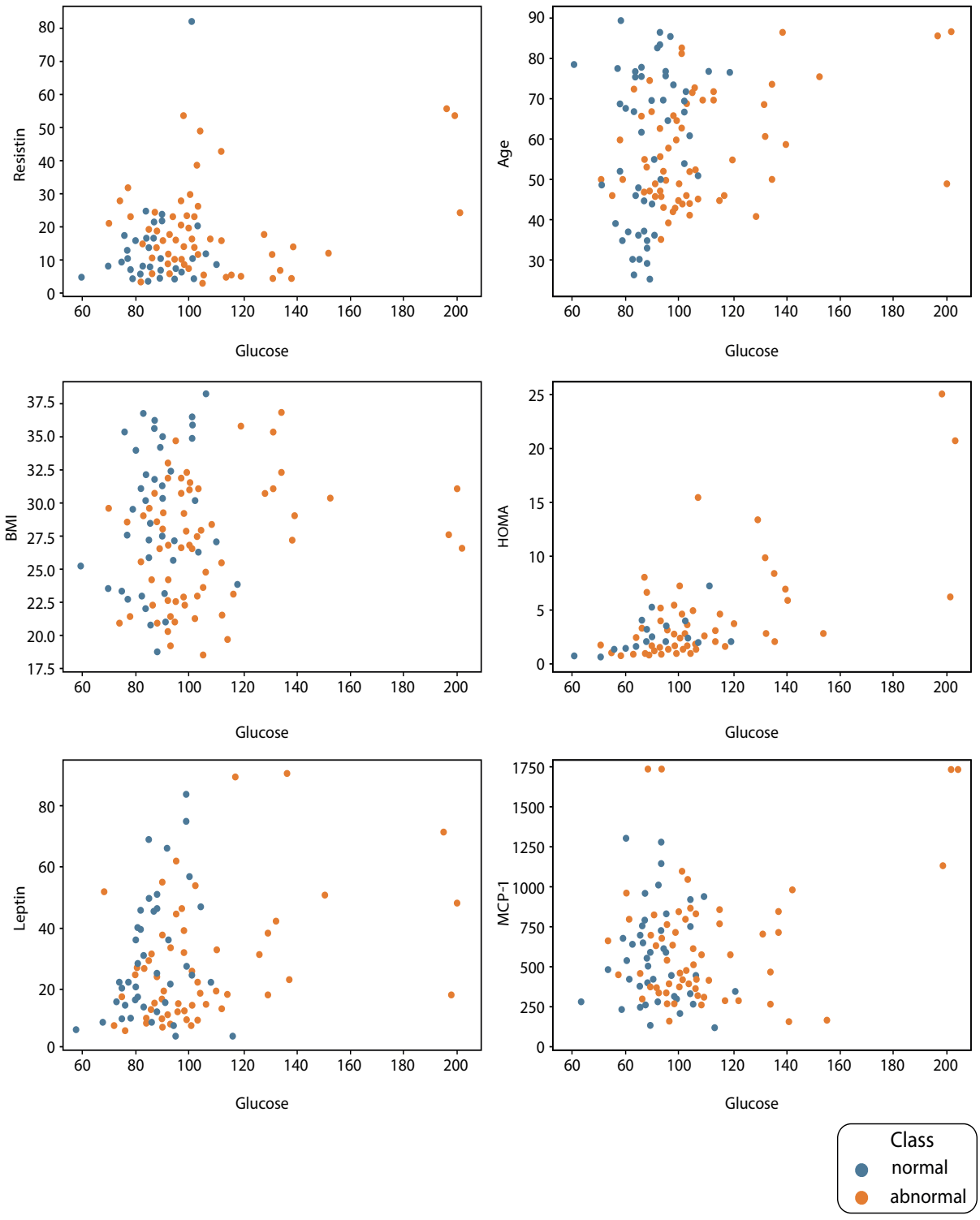

Fig. 5: A data representation between glucose and other parameters in the dataset

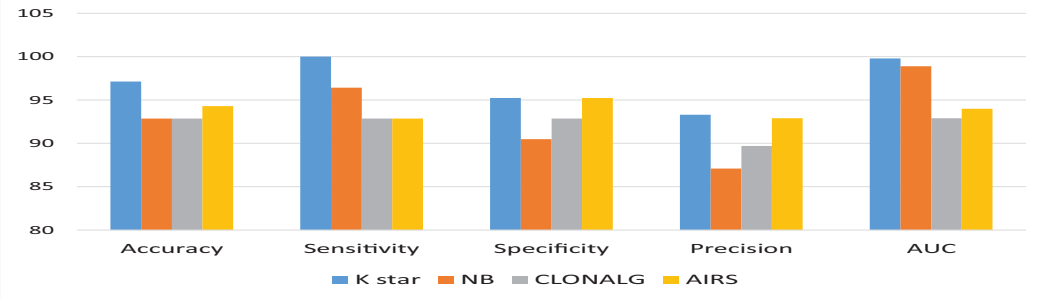

Fig. 6: Comparison between the mentioned algorithms in five metrics

Table 3: The overall classification performance after applying partition membership filtering.

Classifier	Classifier Performance				
	Accuracy(%)	**Sensitivity(%)**	**Specificity(%)**	**Precision**	**AUC**
K-star	97.142	100.00	95.24	0.933	0.998
NB	92.857	96.43	90.48	0.871	0.989
CLONALG	92.86	92.86	92.86	0.897	0.929
AIRS	94.29	92.86	95.24	0.929	0.94

Table 4: Comparison between our model and related model.

(95%CI)	**Sensitivity(%)**	**Specificity(%)**
Our model	[81.65, 99.91]	[80.52, 98.50]
PATRÍCIO (2018)	[82.00, 88.00]	[85.00, 90.00]

Table 5: A total comparison between the proposed algorithms for each variable.

Variables	Figures of interest	Classifier (95% CI)			
		K-star(%)	**NB(%)**	**CLONALG(%)**	**AIRS(%)**
V1-V2	Accuracy	[73.62, 91.89]	[65.55, 86.33]	[62.44, 83.99]	[59.38, 81.60]
	Sensitivity	[51.33, 86.78]	[51.33, 86.78]	[63.11, 93.94]	[81.65, 99.91]
	Specificity	[80.52, 98.50]	[65.88, 91.40]	[52.91, 82.38]	[38.67, 70.15]
V1-V3	Accuracy	[73.62, 91.89]	[65.55, 86.33]	[41.94, 66.26]	[70.34, 89.72]
	Sensitivity	[44.07, 81.36]	[59.05, 91.70]	[87.66, 100.00]	[59.05, 91.70]
	Specificity	[87.43, 99.94]	[60.55, 87.95]	[12.05, 39.45]	[68.64, 93.03]
V1-V4	Accuracy	[68.73, 88.61]	[68.73, 88.61]	[68.73, 88.61]	[47.59, 71.53]
	Sensitivity	[59.05, 91.70]	[59.05, 91.70]	[59.05, 91.70]	[0.00, 12.34]
	Specificity	[65.88, 91.40]	[65.88, 91.40]	[65.88, 91.40]	[91.59, 100.00]
V1-V5	Accuracy	[75.29, 92.93]	[59.38, 81.60]	[33.74, 58.06]	[35.09, 59.45]
	Sensitivity	[47.65, 84.12]	[24.46, 62.82]	[81.65, 99.91]	[87.66, 100.00]
	Specificity	[87.43, 99.94]	[77.38, 97.34]	[3.98, 25.63]	[3.98, 25.63]
V1-V6	Accuracy	[70.34, 89.72]	[70.34, 89.72]	[53.40, 76.65]	[25.89, 49.52]
	Sensitivity	[63.11, 93.94]	[63.11, 93.94]	[4.03, 32.67]	[15.88, 52.35]
	Specificity	[65.88, 91.40]	[65.88, 91.40]	[91.59, 100.00]	[25.63, 56.72]
V1-V9	Accuracy	[86.01, 98.42]	[70.34, 89.72]	[41.94, 66.26]	[50.48, 74.11]
	Sensitivity	[81.65, 99.91]	[63.11, 93.94]	[18.64, 55.93]	[10.69, 44.87]
	Specificity	[80.52, 98.50]	[65.88, 91.40]	[50.45, 80.43]	[74.37, 96.02]

98.42], [81.65, 99.91], and [80.52, 98.50], respectively based on Glucose and MCP-1. Therefore, Glucose and MCP-1 may be considered a good predictor for BC biomarkers to perform screening tests.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article

References

[1] Cruz, Joseph A., and David S. Wishart. Applications of machine learning in cancer prediction and prognosis, Cancer informatics, 2 ,59-78(2006).
[2] BC Homepage, https://www.dosomething.org/us/facts/11-facts-about-breast-cancer, last accessed 10/2/2020.
[3] Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer, BMC cancer, 18, 1-8 (2018).
[4] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal, 13, 8-17(2015).
[5] Saber, Abeer, Aya M. Al-Zoghby, and Samir Elmougy. Big-data aggregating, linking, integrating and representing using semantic web technologies, in Proc.AISI2018, 331-342, (2018).
[6] Lee, J.H., Lee, S.H., Baek, C., Chun, H., Ryu, J.H., Kim, J.W., Deaton, R. and Zhang, B.T. In vitro molecular machine learning algorithm via symmetric internal loops of DNA, Biosystems, 158, 1-9 (2017).
[7] Chaurasia, V. and Pal, S. A novel approach for breast cancer detection using data mining techniques, International Journal of Innovative Research in Computer and Communication Engineering, 2, 2456-2465 (2017).
[8] Agrawal, S., Singh, B., Kumar, R., & Dey, N. Machine learning for medical diagnosis: A neural network classifier optimized via the directed bee colony optimization algorithm, U-Healthcare monitoring systems, 1, 197-215 (2019).
[9] Hussien, Abdelazim G., Mohamed Amin, and Mohamed Abd El Aziz. A comprehensive review of moth-flame optimisation: variants, hybrids, and applications, Journal of Experimental & Theoretical Artificial Intelligence,32, 705–725 (2020).
[10] Jain, Indu, Vinod Kumar Jain, and Renu Jain. Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification, Applied Soft Computing, 62, 203-215 (2018).
[11] Lötsch, J., Sipilä, R., Tasmuth, T., Kringel, D., Estlander, A.M., Meretoja, T., Kalso, E. and Ultsch, A. Machine-learning-derived classifier predicts absence of persistent pain after breast cancer surgery with high accuracy, Breast cancer research and treatment, 171, 399-411 (2018).
[12] Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V. and Akbani, R. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer, Cell, 173(2), 291-304 (2018).
[13] Mahmood, D.Y. and Hussein, M.A. Intrusion detection system based on K-star classifier and feature set reduction, International Organization of Scientific Research Journal of Computer Engineering (IOSR-JCE), 15(5), pp.107-112 (2013).
[14] Thangaraju, P. and Mehala, R. Performance analysis of PSO-KStar classifier over liver diseases, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 4(7), 3132-3137 (2015).
[15] Mahmood, Deeman Y., and Mohammed A. Hussein. Intrusion detection system based on K-star classifier and feature set reduction, International Organization of Scientific Research Journal of Computer Engineering (IOSR-JCE), 15(5), 107-112 (2013).
[16] Guney, K., Babayigit, B. İ. L. A. L., & Akdagli, A. Position only pattern nulling of linear antenna array by using a clonal selection algorithm (CLONALG), Electrical Engineering, 90(2), 147-153 (2007).
[17] Zare, A., Jahromi, M. Z., & Boostani, R. An adaptive-distance artificial immune recognition system, Neural Network World, 20(5), 637-650 (2010).
[18] Dataset, https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#, last accessed 10/2/2020.

Abeer Saber was born in Damietta, Domyat, Egypt, in 1992. She received the B.Sc. degree in computer science and the M.Sc. degree in computer science from Mansoura University, Egypt, in 2013 and 2018, respectively. She is currently an Assistant Lecturer of computer science with the Faculty of Computers and Information, Kafr El-Sheikh University. She has published many research articles published in prestigious international conferences and reputable journals. Her current research interests include big data analysis, semantic web, linked open data, data mining, and machine learning.

Osama M. Abo-Seida was born in Kotor, Tanta, Egypt, in 1968. He received the B.Sc. degree in mathematics, the M.Sc. degree in mathematics, and the Ph.D. in mathematics from Tanta University, Egypt, in 1990 ,1994, and 1997 respectively. He is currently a Professor of mathematics and the Dean of the faculty of Computers and Information with Kafr El-Sheikh University. He has authored or coauthored many articles in international repute journals. His research interests include Wave Propagation, Applied Mathematics, Computional Electromagnetics, Magnetical Networks, and Machine Learning.

Mohamed Sakr was born in Shebin El-Kom, Menoufia, Egypt, in 1990. He received the B.Sc. degree in computer science, the M.Sc. degree in computer science, and the Ph.D. in computer science from Menoufia University, Egypt, in 2011, 2014,and 2019 respectively . He is currently a Lecturer of computer science with the Faculty of Computers and Information, Menoufia University. He has published many research articles published in prestigious international conferences and reputable journals. His current research interests include big data analysis, anomaly detection, data mining, and machine learning.

Arabi E. keshk received the B.Sc. in Electronic Engineering and M.Sc. in Computer Science and Engineering from Menoufia University, Faculty of Electronic Engineering in 1987 and 1995, respectively and received his PhD in Electronic Engineering from Osaka University, Japan in 2001. His research interest includes software testing, software engineering, distributed system, cloud computing, IoT, big Data analytics, and bioinformatics.