Medical Technologies Journal, Volume: 4, Issue: 2, April-June 2020, Pages:535-544.
Doi : https://doi.org/10.26415/2572-004X-vol4iss2p535-544

# Breast cancer classification using machine learning techniques: a comparative study

**Type of article: Original**

Djihane Houfani[1], Sihem Slatnia[1], Okba Kazar[1], Noureddine Zerhouni[2], Hamza Saouli[1], Ikram Remadna[1]

[1] LINFI Laboratory, Department of Computer Science, University of Biskra, Algeria.
[2] Institut FEMTO-ST, UMR CNRS 6174- UFC / ENSMM / UTBM, Besanon, France

**Abstract:**

**Background:** The second leading deadliest disease affecting women worldwide, after lung cancer, is breast cancer. Traditional approaches for breast cancer diagnosis suffer from time consumption and some human errors in classification. To deal with this problems, many research works based on machine learning techniques are proposed. These approaches show their effectiveness in data classification in many fields, especially in healthcare.

**Methods:** In this cross sectional study, we conducted a practical comparison between the most used machine learning algorithms in the literature. We applied kernel and linear support vector machines, random forest, decision tree, multi-layer perceptron, logistic regression, and k-nearest neighbors for breast cancer tumors classification. The used dataset is Wisconsin diagnosis Breast Cancer.

**Results:** After comparing the machine learning algorithms efficiency, we noticed that multilayer perceptron and logistic regression gave the best results with an accuracy of 98% for breast cancer classification.

**Conclusion:** Machine learning approaches are extensively used in medical prediction and decision support systems. This study showed that multilayer perceptron and logistic regression algorithms are performant ( good accuracy specificity and sensitivity) compared to the other evaluated algorithms.

**Keyword:** Breast cancer, Classification, Accuracy, Comparative study, Machine learning.

**Corresponding author:** Djihane Houfani, LINFI Laboratory, Department of Computer Science, University of Biskra, Algeria., email: djihane.houfani@univ-biskra.dz.

Received: July 12 2020. Reviewed: August 17 2020. Accepted: October 3 2020. Published: October 20 2020.
Medical Technologies Journal subscribes to the principles of the Committee on Publication Ethics (COPE).

## 1. Introduction

Breast cancer (BC) is among the major deadliest diseases affecting women around the world [1]. It occurs because of the uncontrolled growth of the cells in breast tissue. BC diagnosis based on histopathological data can provide inaccurate outcomes. In last decade, machine learning (ML) techniques are widely used in diagnosis of BC to help pathologists and physicians in early detection, decision making process and giving a successful plan for treatment.

In the literature, several algorithms for breast cancer diagnosis and prognosis are proposed. In this paper we provide a practical comparison between kernel and linear support vector machines (K-SVM, L-SVM respectively), random forest(RF), decision tree (DTs), multi-layer perceptron (MLP), logistic regression (LR), and k-nearest neighbors (k-NN) which are the most used algorithms in several researches [2-4]. The goal of this study is to evaluate the performance of these algorithms in terms of effectiveness, efficiency and accuracy,. We conduct this comparative study to find out the best approach to be used in learning models, to apply it on new datasets and improve its performance by combining it with other technologies such as fuzzy learning, convolutional neural networks, genetic algorithms …etc.

In the rest of the paper, we will explain our experiment, the materials and the methods, in Section 2. Then, we will present the obtained results in Section 3, and finally, we present our conclusions and future works in Section 4.

# 2. Materials and Method

## A. Related works

Classification is one of the most crucial machine learning tasks. It is applied in many research works using several medical datasets in order to classify breast cancer cells. In this section, we present some works that apply ML techniques for early BC diagnosis.

Abdel-Zaher and Eldeib [5] proposed a computer aided diagnosis (CAD) scheme for BC detection. Deep belief network unsupervised learning and back propagation supervised learning are used in this system. It is evaluated using the Wisconsin Breast Cancer Dataset (WBCD) and gave an accuracy of 99.68% on.

Thein and Khin [6] presented an approach for BC classification. The proposed system applied the island-based training method on the Wisconsin Diagnostic and Prognostic Breast Cancer data sets. This approach gave good accuracy and low training time by using and analyzing two migration topologies.

Ibrahim and Siti [7] applied MLP neural network and enhanced non-dominated sorting genetic algorithm (NSGA-II) for BC automatic classification. Compared to other methods, this work improved classification accuracy by optimizing the ANN parameters and network structure.

Guan et al. [8] proposed breast tumor classifier. They used Wisconsin Breast Cancer Dataset to evaluate their diagnostic model called self-validation cerebellar model articulation controller (SVCMAC) neural network. The advantages of this method are simple computation, fast learning, and good generalization capability.

Kumar et al. [9] proposed an ensemble voting classifier. It combines J48, Naïve Bayes, and SVM on WBCD to improve the decision-making approaches in the prediction of BC survivability. The dataset is preprocessed. Then, it was trained and tested using 10-fold cross validation. The combined model gave good accuracy.

Mittal et al. [10] presented a hybrid classifier for BC diagnosis. The classifier is a combination of self-organizing maps (SOM) and stochastic gradient descent (SGD) on WBCD. The proposed system improved the accuracy compared to other works in state of the art ML techniques.

Haifeng et al. [11] proposed an SVM-based ensemble learning model for BC diagnosis. The proposed model includes C-SVM and ν–SVM structures, and six types of kernel functions. It was tested using two datasets: the WBCD (original and diagnosis) datasets, and the Surveillance, Epidemiology, and End Results (SEER) dataset. The system presented a good accuracy compared to works based on single SVM.

Emina et al. [12] proposed a BC classifier applying several machine learning algorithms. Logistic Regression, Decision Trees, RF, Bayesian Network, MLP, Radial Basis Function Networks (RBFN), SVM, Rotation Forest and genetic algorithm-based feature selection were compared and the Rotation Forest model with GA-based 14 features The system gave the best results (accuracy 99.48%). The system was evaluated using diagnosis and original WBCD datasets.

Zheng et al. [13] applied K-means and support vector machine (K-SVM) algorithms to develop a hybrid system for breast tumors classification. The method is tested on WDBC dataset and gave an accuracy of 97.38%.

Arpit and Aruna [14] developed a genetically optimized neural network (GONN) for BC classification. They improved the neural network architecture by introducing a new crossover and mutation operators. The proposed approach is evaluated by using WBCD and presented good accuracy.

By this study, we aim to employ MLP, L-SVM, K-SVM, DTs, RF, KNN, NB, LR and NB on the Wisconsin Breast Cancer (Diagnosis) dataset to compare their performance (effectiveness and efficiency).

## B. Experiments

In this section, a presentation of the experiments is given.

### a. Dataset

In the literature, many studies used The Wisconsin Breast Cancer dataset (diagnosis). It is available in the UCI Machine Learning Repository. It has 569 instances (Benign: 357 Malignant: 212), 2 classes (37.3% malignant and 62.7% benign), and 32 attributes which are: ID number, Diagnosis (M = malignant, B = benign), Radius (mean of distances from center to points on the perimeter), Texture (standard deviation of gray-scale values), Perimeter (mean size of the core tumor), Area, Smoothness (local variation in radius lengths), Compactness (perimeter^2 / area - 1.0), Concavity (severity of concave portions of the contour), Concave points (number of concave portions of the contour, Symmetry, Fractal dimension (coastline approximation – 1), the mean, standard error and "worst" or largest (mean of the three largest values) of these attributes are computed for each image, resulting in 32 attributes) [15].
In this work, the ML algorithms are evaluated using WBCD.

### b. Data Normalization

The z-score standardization method is used to normalize the dataset. The used equation for calculating the z-score is given in (1), where, $\mu_m$ is the mean value of the attribute, $\delta_m$ is the standard deviation, $x_m$ is the raw data, and $x'_m$ is the normalized data [16].

$$x'_m = \frac{x_m - \mu_m}{\delta_m} \qquad (1)$$

### c. Methods description

Supervised learning algorithms are algorithms that learn on a labeled dataset, given training on input and output parameters [17]. In this case, the goal of computers is to learn a general formula which maps inputs to outputs. Predictive models developed in this type of learning are achieved using classification and regression techniques. Classification methods predict discrete variable however, regression techniques provide continuous variables.

- **Naïve Bayes (NB)**

NB is a probabilistic ML method. It calculates probabilities of different classes given some observed evidence [18]. It uses the maximum likelihood method for parameter estimation and is appropriate for high dimensionality inputs. Equation (2) gives the probability of a class given predictor $P(c|x)$, where $P(c|x)$ is posterior probability, $P(c)$ is the class prior probability, $P(x|c)$ is the likelihood, and $P(x)$ is the probability of predictor [ 4]:

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)} \qquad (2)$$

- **Support Vector Machine (SVM)**

SVM algorithm consists of finding a hyperplane which separates classes [17]. It is suited for high dimensional inputs and memory efficient because it uses support vectors. SVM is a powerful algorithm; however its storage and computational grow up with the number of training vectors [18].

- **Decision tree (DT)**

DT is a diagram having a tree structure, where each node represents a test on an attribute, each branch denotes an outcome of the test, and each terminal node detains a class label. It is used to classify input data points or predict output values given inputs [19]. It is efficient and capable of fitting complex datasets.

- **Random forest (RF)**

RF is a large number of decision trees which are ensemble [18]. In this method, each individual tree builds an output class then the average of predictions is taken. The final result is generated by taking the mode of classes found separately [11].

- **Logistic regression (LR)**

LR algorithm can be applied for classification and regression tasks [19]. It is a statistical method for data analyzing. It aims to obtain the best fitting model which describes the relation between inputs and outputs.

- **K-Nearest Neighbor (K-NN)**

K-NN is the simplest machine learning method. It is non-parametric method used for both classification and regression. It consists of calculating the distance between the test data and the input and gives the prediction accordingly [18].

- **Multilayer Perceptron (MLP)**

MLP is a feed forward supervised neural network for data classification. It is composed of many layers as a directed graph between the input and output layers. For the training task, MLP uses backpropagation method [18].

# 3. Results and discussion

The goal of this work is to compare the performance of NB, SVM, DT, LR, RF, k-NN and MLP. We used split method to divide the dataset: a training set (75%) to train the model, and a testing set (25%) to evaluate it.

*(75 × 569) / 100 = 426 → training data*
*569 – 426 = 143 → testing data*

### A. Efficiency

A confusion matrix is a performance measurement that provides information about real and predicted values resulting from a classification system. Table 1 is a description of a confusion matrix for a two class classifier.

TABLE I. CONFUSION MATRIX REPRESENTATION

| **True value** | **Predicted value** | |
|---|---|---|
| | ***Positive*** | ***Negative*** |
| Positive | TP | FN |
| Negative | FP | TN |

Where:

- **TP:** malignant tumors (M) correctly predicted as malignant;
- **FP:** benign tumors (B) incorrectly predicted as malignant tumors (M);
- **FN:** malignant tumor (M) incorrectly identified as benign tumor (B);
- **TN:** benign tumor (B) correctly identified as benign.

To compare true classes and predicted results, we use confusion matrices shown in figure 1. We note that both of MLP and LR correctly predict 140 instances from 143 instances (87 benign instances correctly predicted benign and 53 malignant instance effectively malignant), and 3 instances incorrectly predicted (2 benign instances predicted as malignant and 1 malignant instance predicted as benign). As a result, MLP and LR give the best accuracies.

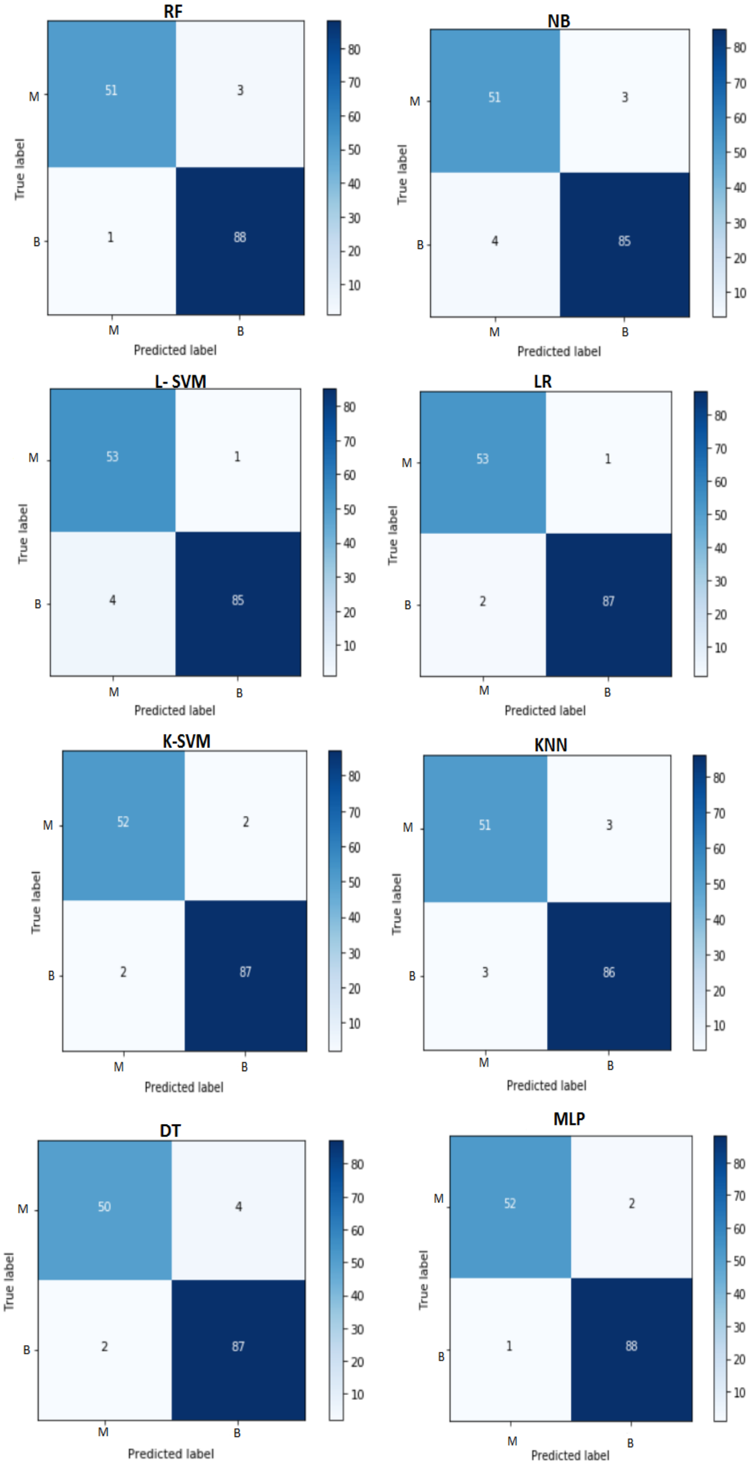

**Fig. 1.** Confusion matrices

To check how our models are efficient, we build the ROC (receiver operating characteristic) curve presented in figure 2. ROC curve is used with binary classifiers, it is used to understand the performance of a ML algorithm; it plots the TPR against the FPR [17]. The TPR and FPR are given in equation (3) and (4) [20]:

$$TPR = \frac{TP}{(TP+FN)} \qquad (3)$$

$$FPR = \frac{FP}{(FP+TN)} \qquad (4)$$

TPR and FPR values are given in table 2.

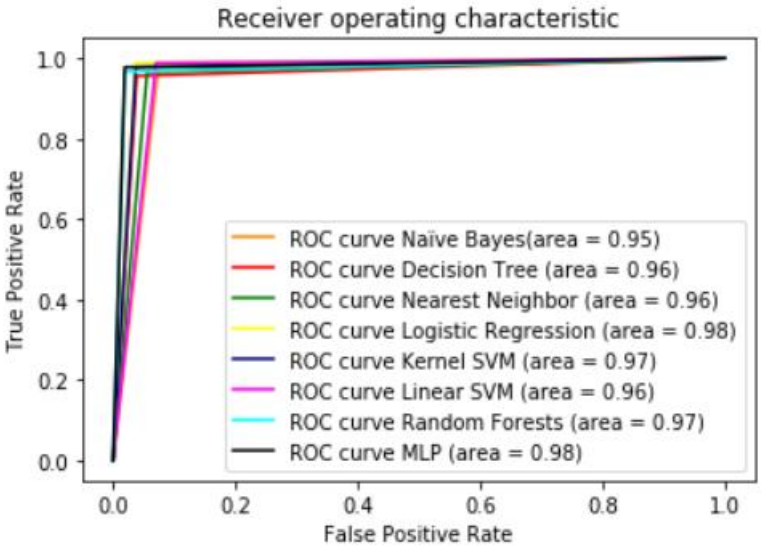

**Fig. 2.** ROC curve

We can easily observe that MLP and LR are the best classifiers followed by other algorithms.

## B. Effectiveness

To measure the performance of used algorithms, we conduct a comparison based on accuracy, correctly and incorrectly classified instances.

Accuracy is a metric used for evaluating classification models. It gives the ratio of the total number of the correct predictions, its equation is given in (5) [21]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \qquad (5)$$

Table 2 and figure 3 show the obtained results.

TABLE II. PERFORMANCE MEASUREMENTS

| Evaluation criteria | Classifiers | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | MLP | L-SVM | K-SVM | DT | RF | LR | KNN | NB |
| Correctly classified instances | 140 | 138 | 139 | 137 | 139 | 140 | 137 | 136 |
| Incorrectly classified instances | 3 | 5 | 4 | 6 | 4 | 3 | 6 | 7 |
| Training accuracy | 0,988 | 0,99 | 0,988 | 1 | 0,957 | 0,988 | 0,98 | 0,93 |
| Testing accuracy | 0,979 | 0,96 | 0,97 | 0,958 | 0,972 | 0,979 | 0,958 | 0,95 |
| TPR | 0.963 | 0.981 | 0.963 | 0.926 | 0.944 | 0.981 | 0.944 | 0.944 |
| FPR | 0.011 | 0.045 | 0.022 | 0.022 | 0.011 | 0.022 | 0.034 | 0.045 |

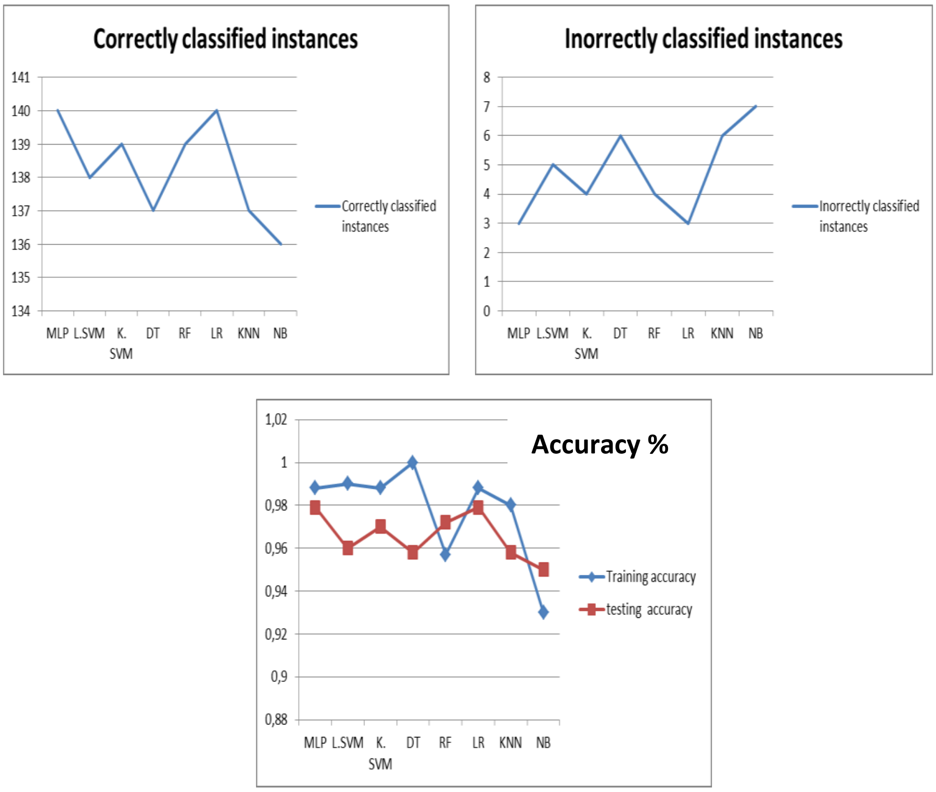

**Fig. 3.** comparative graphs of used classifiers

We can notice from figure 3, that accuracy obtained by MLP and LR (98%) is the best compared to accuracy obtained by KNN, DTs, RF, L-SVM, K-SVM and NB which vary between 95% and 97%. We can also easily see that both of MLP and LR reach the best value of correctly classified instances and the lower value of incorrectly classified instances compared to the other classification methods. It is noted that the DTs performance was the highest in the training phase, but it wasn’t in the testing phase; this proofs that DTs can learn accurately in the training phase, however it can be weak in generalization.

In summary, both of MLP and LR algorithms provide a good performance (effectiveness and efficiency, accuracy, sensitivity and specificity) compared to the other algorithms. In this study, we achieve the highest accuracy (98%) in classifying breast tumors.

# 4. Conclusion

Machine learning techniques are revolutionizing the field of bio-medical and healthcare. One of the most important challenges of ML is to provide computationally efficient and accurate classifiers for healthcare field. In the last decade, many research works have been conducted in medical field for this reason. ML techniques have played a crucial role in improving classification and prediction accuracy. Although several algorithms have achieved a very good accuracy using WBCD, the development of new algorithms is still essential.

In this paper, we employed MLP, L-SVM, K-SVM, DTs, RF, KNN, LR and NB on WBCD dataset. We compared their performance in terms of effectiveness and efficiency to find the highest classification accuracy. In this experimental study, we achieved the best accuracy (98%) in classifying BC dataset using MLP and LR. In conclusion, MLP and LR have shown their efficiency in BC classification.

For the future work, we plan to apply deep reinforcement learning and genetic algorithms on new datasets to boost the breast cancer diagnosis and further improve prognostic accuracy.

## 5. Conflict of interest statement

We certify that there is no conflict of interest with any financial organization in the subject matter or materials discussed in this manuscript.

## 6. Authors' biography

**Djihane Houfani** received her Master degree in Computer Science from University of Biskra, Algeria in 2017. She is now a PhD student in artificial intelligence at the University of Biskra and her current research interest includes medical prediction, deep learning, multi-agent systems and optimization.

**Sihem Slatnia** was born in the city of Biskra, Algeria. She followed her high studies at the university of Biskra, Algeria at the Computer Science Department and obtained the engineering diploma in 2004 on the work "Diagnostic based model by Black and White analyzing in Background Petri Nets", After that, she obtained Master diploma in 2007 (option: Artificial intelligence and advanced system's information), on the work "Evolutionary Cellular Automata Based-Approach for Edge Detection". She obtained PhD degree from the same university in 2011, on the work "Evolutionary Algorithms for Image Segmentation based on Cellular Automata". Presently she is an associate professor at computer science department of Biskra University. She is interested to the artificial intelligence, emergent complex systems and optimization.

**Okba Kazar** professor in the Computer Science Department of Biskra, he helped to create the laboratory LINFI at the University of Biskra. He is a member of international conference program committees and the "editorial board" for various magazines. His research interests are artificial intelligence, multi-agent systems, web applications and information systems.

**Noureddine Zerhouni** holds a doctorate in Automatic-Productivity from the National Polytechnic Institute of Grenoble (INPG), France, in 1991. He was a lecturer at the National School of Engineers (ENI, UTBM) in Belfort. Since 1999, he is Professor of Universities at the National School of Mechanics and Microtechnics (ENSMM) in Besançon. He is doing his research in the Automatic department of the FEMTO-ST Institute in Besançon. His areas of research are related to the monitoring and maintenance of production systems.

**Hamza Saouli** received the Master and Doctorate degrees in Computer Science from University of Mohamed KhiderBiskra (UMKB), the Republic of Algeria in 2010 and 2015, respectively. He is a university lecturer since 2015 and his research interest includes artificial intelligence, web services and Cloud Computing.

**Ikram Remadna** received her Master degree in Computer Science from University of Biskra, Algeria in 2016. She is now a PhD student in artificial intelligence at the University of Biskra and her current research interest includes Prognostics and Health Management and Deep learning.

## 7. References

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 19992008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute (2012).

[2] BF Cruz, JT de Assis, VV Estrela, A Khelassi, A compact SIFT-based strategy for visual information retrieval in large image databases, Medical Technologies Journal 3 (2), 402-412, 2019 https://doi.org/10.26415/2572-004X-vol3iss2p402-412

[3] Q Memon, (2019), On assisted living of paralyzed persons through real-time eye features tracking and classification using Support Vector Machines, Medical Technologies Journal 3 (1), 316-333 https://doi.org/10.26415/2572-004X-vol3iss1p316-333

[4] Devi I., Karpagam G.R. and Vinoth Kumar B (2017), A survey of machine learning techniques. International Journal of Computational Systems Engineering. 3 (4): 203-212. https://doi.org/10.1504/IJCSYSE.2017.10010099

[5] Abdel-Zaher Ahmed M. and Eldeib Ayman M. (2016), Breast cancer classification using deep belief networks. Expert Systems with Applications. ELSEVIER;46:139-144. https://doi.org/10.1016/j.eswa.2015.10.015

[6] Thein HTT. and Khin MMT. (2015), An Approach for Breast Cancer Diagnosis Classification Using Neural Network. Advanced Computing. An International Journal (ACIJ). 6 (1): 1-11. https://doi.org/10.5121/acij.2015.6101

[7] Ashraf O. I. and Siti, M. S. (2018), Intelligent breast cancer diagnosis based on enhanced Pareto optimal and multilayer perceptron neural network. International Journal of Computer Aided Engineering and Technology. Inderscience. 10 (5): 543-556. https://doi.org/10.1504/IJCAET.2018.10013710

[8] Guan J., Lin L., Ji G., Lin C., Le T., Imre JR. (2016), Breast Tumor Computer-aided Diagnosis using Self-Validating Cerebellar Model Neural Networks. Acta Polytechnica Hungarica. 13 (4): 39-52. https://doi.org/10.12700/APH.13.4.2016.4.3

[9] Karthik Kumar U., Sai Nikhil M.B. and Sumangali K. (2017), Prediction of Breast Cancer using Voting Classifier Technique. IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM); 2017 2 - 4 August; Veltech Dr.RR & Dr.SR University, Chennai, T.N., India. 108-114 https://doi.org/10.1109/ICSTM.2017.8089135

[10] Mittal D., Gaurav D. and Sanjiban SR. (2015), An Effective Hybridized Classifier for Breast Cancer Diagnosis. IEEE International Conference on Advanced Intelligent Mechatronics (AIM); 2015 July 7-11. Busan, Korea. https://doi.org/10.1109/AIM.2015.7222674

[11] Haifeng W., Bichen Z., Sang W.Y., Hoo S. K. (2017), A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis. European Journal of Operational Research. Elsevier: 1-33.

[12] Emina A., Abdulhamit S. (2015), Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput & Applic. Springer.

[13] Zheng B., Sang WY., Sarah SL. (2013), Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications. Elsevier: 1-7.

[14] Arpit, B., Aruna, T. (2015), Breast Cancer Diagnosis Using Genetically Optimized Neural Network Model. Expert Systems with Applications. Elsevier: 1-15.

[15] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), last accessed September 20, 2019.

[16] E. Kreyszig (1979), Advanced Engineering Mathematics (Fourth ed.). Wiley, ISBN 0-471-02140-7.

[17] Aurélien Géron (2017), Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Published by O'Reilly Media.

[18] Amit K. and Bikash KS. (2017), A case study on machine learning and classification. International Journal Information and Decision Sciences 9 (2): 97-208 https://doi.org/10.1504/IJIDS.2017.084885

[19] Francois Chollet (2018), Deep Learning with Python. Published by Manning Publications.

[20] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning - ICML '06. https://doi.org/10.1145/1143844.1143874

[21] Piri, S., Delen, D., & Liu, T. (2018). A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets. Decision Support Systems, 106, 15-29. https://doi.org/10.1016/j.dss.2017.11.006

[22] Razmjooy N., Estrela VV., Loschi HJ. (2019), A study on metaheuristic-based neural networks for image segmentation purposes, Data Science Theory, Analysis and Applications, Taylor and Francis, Abingdon, UK, 2019. https://doi.org/10.1201/9780429263798-2

[23] Razmjooy N., N, Estrela V.V., Loschi H.J., Farfan W.S. (2019), A Comprehensive Survey of New Metaheuristic Algorithms, Wiley.

[24] Karim CN., Mohamed O, Ryad T. (2018), A new approach for breast abnormality detection based on thermography. Medical Technologies Journal, 2(3):245-254. https://doi.org/10.26415/2572-004X-vol2iss3p245-254

[25] Hemanth J., Estrela V.V. (2017), Deep Learning for Image Processing Applications. Advances in Parallel Computing, Vol. 31, IOS Press, Amsterdam, Netherlands. ISSN: 978-1-61499-822-8. https://www.iospress.nl/book/deep-learning-for-imageprocessing-applications/

[26] Souadih K., Belaid A, Ben Salem D. (2019), Automatic Segmentation of the Sphenoid Sinus in CT-Scans Volume with Deep Medics 3D CNN Architecture, Medical Technologies Journal, Vol. 3, no. 1, pp. 334-46, https://doi.org/10.26415/2572-004X-vol3iss1p334-346

.