World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:17, No:8, 2023

Classification of Potential Biomarkers in Breast Cancer Using Artificial Intelligence Algorithms and Anthropometric Datasets

Aref Aasi, Sahar Ebrahimi Bajgani, Erfan Aasi

Open Science Index, Bioengineering and Life Sciences Vol:17, No:8, 2023 publications.waset.org/10013232.pdf

***Abstract*—**Breast cancer (BC) continues to be the most frequent cancer in females and causes the highest number of cancer-related deaths in women worldwide. Inspired by recent advances in studying the relationship between different patient attributes and features and the disease, in this paper, we have tried to investigate the different classification methods for better diagnosis of BC in the early stages. In this regard, datasets from the University Hospital Centre of Coimbra were chosen, and different machine learning (ML)-based and neural network (NN) classifiers have been studied. For this purpose, we have selected favorable features among the nine provided attributes from the clinical dataset by using a random forest algorithm. This dataset consists of both healthy controls and BC patients, and it was noted that glucose, BMI, resistin, and age have the most importance, respectively. Moreover, we have analyzed these features with various ML-based classifier methods, including Decision Tree (DT), K-Nearest Neighbors (KNN), eXtreme Gradient Boosting (XGBoost), Logistic Regression (LR), Naive Bayes (NB), and Support Vector Machine (SVM) along with NN-based Multi-Layer Perceptron (MLP) classifier. The results revealed that among different techniques, the SVM and MLP classifiers have the most accuracy, with amounts of 96% and 92%, respectively. These results divulged that the adopted procedure could be used effectively for the classification of cancer cells, and also it encourages further experimental investigations with more collected data for other types of cancers.

***Keywords*—**Breast cancer, health diagnosis, Machine Learning, biomarker classification, Neural Network.

I. Introduction

BC is a disease in which certain cells in the breast become abnormal and multiply uncontrollably to form tumors. BC is considered as a leading cause of death among women, and the second most diagnosed cancer worldwide [1]-[4]. According to reports and the World Health Organization (WHO), in 2020, there were over 2 million women diagnosed with BC and approximately seven hundred thousand died globally because of it [5], [6].

BC screening is of great importance to follow the early diagnosis and may increase the likelihood of obtaining a good outcome in treatment [7]-[9]. One of the primal methods for early BC detection is the breast self-examination and finding any abnormal masses [10], another traditional way is to collect a piece of tissue or a sample of cells from the area for testing and analyzing in a laboratory [11]. However, finding the BC biomarkers, and introducing of cheaper, more efficient, and noninvasive methods would be beneficial [12].

Biomarkers of the disease are being used in screening and diagnosis and monitoring of disease progression [13]-[16]. There have been many reports regarding potential candidates for BC biomarkers [17]-[22]. For example, it has been shown that the level amount of resistin and glucose as biomarkers are higher in women with BC [23], [24]. In [25] authors studied the correlation of resistin, and adiponectin with BC risk among 41 BC patients. They disclosed that high level of resistin and low level of adiponectin could be responsible for risk of BC.

In another work [24], authors examined 150 women consisting of 82 BC patients and 68 healthy controls in the hospital to be able to investigate the links between the clinicopathological features and the BC risk. They adopted the enzyme-linked immunosorbent assay (ELISA) and found that serum resistin, leptin, adiponectin, and visfatin levels could be considered risk factors and biomarkers for BC. There is no doubt that assessment of data obtained from patient and doctor evaluation is the most valuable elements in diagnosis.

Recently, applying statistical analyses like ML techniques have gained considerable attention in different areas [26], [27] and they have been used for the correct diagnosis and the classification of the BC dataset [28]. Data processing and classification seem to be imperative ways to classify datasets of patients into malignant or benign groups [29]-[32]. For instance, in [29], Random Forest and Genetic Algorithm (GA) methods were applied on the BC datasets from WDBC (Wisconsin Diagnostic Breast Cancer database). The authors selected features of BC and classified them with SVM method.

Other scholars [33] used the breast cancer dataset from WDBC and emphasized that feature selection is a key role to build a BC classifier for the preventive diagnosis. In this regard, kernel-based Bayesian classifier was employed to select the BC features. In another survey [34], datasets from WDBC were utilized and seven deep learning-based techniques such as Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU) were used to analyze the dataset. The authors claimed that from performances, LSTM and GRU can provide useful results. Moreover, new datasets from blood analysis have been reported which was conducted in the

Aref Aasi is with Department of Mechanical Engineering, Worcester Polytechnic Institute, Worcester, Massachusetts 01609, United States (e-mail: aaasi@wpi.edu).

Sahar Ebrahimi Bajgani is with Department of Business School, Worcester Polytechnic Institute, Worcester, Massachusetts 01609, United States.

Erfan Aasi is with Department of Mechanical Engineering, Boston University, Boston, Massachusetts 02215, United States.

Open Science Index, Bioengineering and Life Sciences Vol:17, No:8, 2023 publications.waset.org/10013232.pdf

Gynaecology Department of the University Hospital Centre of Coimbra (CHUC) [35]. In this clinical dataset, nine features and characteristics such as age, body mass index (BMI), glucose, resistin etc. were gathered. The research group [36] used regression and SVM models to determine the presence of BC based on the acquired features.

Motivated by the aforementioned dataset (CHUC), and by the high performance of ML based models in BC diagnosis, we have comprehensively studied different classifier methods. For this purpose, features have been selected carefully by utilizing Random Forest (RF), then selected features were used as input for classification by the DT, KNN, XGBoost, LR, NB, and SVM algorithms. Moreover, we have applied neural the network-based classifier (MLP-Classifier) to the dataset and the accuracy has been compared with other methods.

The remaining part of the paper is oriented as follows. Section II outlines our method, presents the data we used, and discusses model training and performance evaluation. Section III represents the implementation and our main results analysis. Section IV discusses the results and draws some conclusions.

II. Materials and Methods

Dataset

The database used is the CHUC [35]. The dataset provides naïve data, i.e., collected before surgery and treatment. A total of 166 participants were enrolled, and several clinical features were measured, consisting of age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP-1. It should be mentioned that 64 women with BC and 52 healthy volunteers were included in the present study.

There are 10 predictors, all quantitative attributes, and a binary dependent variable, referring the presence or absence of BC (Labels: 1 = Healthy controls, 2 = Patients). The predictors are anthropometric data and parameters which were gathered in routine blood analysis.

K-Nearest Neighbor

The idea is to memorize the training set and classify a new data sample based on the labels of its nearest neighbors in the training set. Given a positive integer k (a hyper-parameter) and a test observation x_0, the initial step of the KNN classifier involves finding a set of K points in the training data that are most similar to x_0, denoted as N_0. Subsequently, the classifier calculates the likelihood of class j by determining the proportion of points in N_0 that have response values equal to j:

$$\mathrm{P}\,(\mathrm{Y} = j \mid X = x_0) = \frac{1}{K}\sum_{i\epsilon N_0}^{N} I\,(y_i = j)$$

where I is the indicator function, so that I = 1 when $y_i = j$. Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

Selection of the most significant and informative features and removal of the remaining features (or in other words compression of original feature set to smaller set) are one of the most important tasks in design of the efficient classification model.

Support Vector Machine

SVM is widely used in biology, due to its high-speed accuracy in multi-dimensional space [37]. SVM finds the best hyperplane to classify data samples with different classes, also the parameters need to be tuned to obtain the most accurate results. Given a labeled dataset, SVM finds the best hyperplane to classify data samples with different classes. Different types of kernel functions have been developed for SVM classifiers, where some of the common ones are linear, Radial Basis Function (RBF), polynomial, and sigmoid functions. For example, in the case of linear separating function, $\mathrm{g(x)} = \mathrm{sign(f(x))}$ where f(x) is the separating function; $\mathrm{f(x)} = \omega^T x + b$. Here, f(x) is a hyperplane, which can be used to separate data.

Decision Tree

A DT classifier is a decision support tool that uses a tree structure built using input features. The aim of this approach is to produce a tree-like structure for the inputs and creates a unique output at every leaf. In order to decrease uncertainty or disorders from the dataset, methods such as entropy, and Gini index can be used.

XGBoost

The Gradient Boosting Classifier depends on a loss function. Gradient boosting systems have two other necessary parts: a weak learner and an additive component.

Naïve Bayes

The NB classifier makes the assumption that the impact of a predictor's value (x) on a certain class (c) is unrelated to the values of other predictors.

Logistic Regression

Data could be categorized into discrete classes using logical regression, which examines the relationship between one or more independent variables. Mathematically, LR estimates a multiple linear regression function defined as:

$$\mathrm{logit(p)} = \log\left(\frac{p(y=1)}{1-p(=1)}\right) = \beta_0 + \beta_1 x$$

where p is probability, β is constant. If log is linearly related to x, then the relation between x and p is nonlinear, and has the form of the S-shaped curve.

Statistical Evaluation

Performance of all classifier's approaches can be measured by different methods, of the most popular metrics are confusion matrix, accuracy, sensitivity (Sens), specificity (Spec), receiver operating characteristic curve (ROC), and Area under the ROC Curve (AUC). Basically, a confusion matrix represents the behavior of our classification model. There are four main factors in the confusion matrix: True Negative (TN), True Positive (TP) (correct classification) and False Negative (FN), False Positive (FP) (incorrect

classification). For example, using these factors, the performance measurements are defined as:

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

$$Sens = \frac{TP}{(TP+FN)}$$

$$Spec = \frac{TN}{(TN+FP)}$$

In aggregate, an excellent model has a performance value near to the 1 which means it has a good measure of separability. A poor model has a performance value near 0 which means it has the worst measure of separability.

The workflow of the procedure that was taken in this paper is exhibited in Fig. 1.

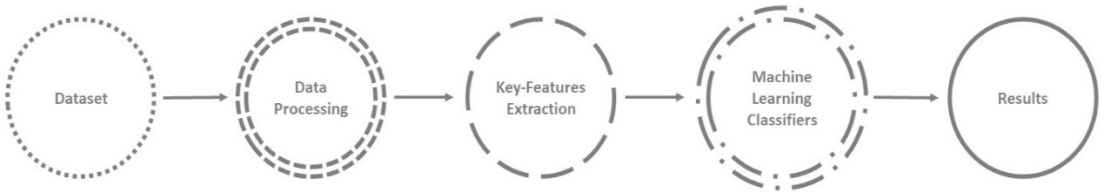

Fig. 1 The graphical procedure framework

III. Results

Data Preprocessing

The number of healthy and patient cases from CHUC (Classification) dataset are demonstrated in Fig. 2.

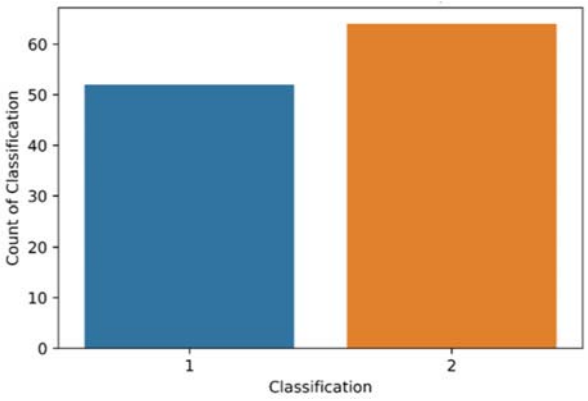

Fig. 2 Number of Healthy (1), and BC patients (2)

It is apparent that the number of healthy controls with label 1 are lesser than BC patients with label 2. To make the dataset suitable for the ML models, some preprocessing actions need to be done on the dataset.

Standardization, normalization, scaling, replacing missing values, and removing unwanted information before introducing the data to the ML model are called preprocessing [38]. Standardization is a common estimator for many ML models, and it can give values centered around zero. StandardScaler was employed to compute the standard deviation. It was noted that the standard deviation (std) can be a robust preprocessor to dataset. So, a preliminary step was taken by applying the std to the dataset.

Over the last few years, data normalization is gaining huge interest in the era of ML in medical applications [39]. Later, data normalization was applied by the min-max normalization method. The min-max normalization was done using Euclidean distance defined as:

$$D(c, e) = \sqrt{\sum_{i=1}^{n}(c_i - e_i)^2}$$

where $c = \{c_1, c_2, \dots, c_n\}$ are the centers of the data, $e = \{e_1, e_2, \dots, e_n\}$ are the real-valued data, and D is the distance. Moreover, all the target datasets (Classification) were converted to binary values 0, and 1 for healthy controls, and patients, respectively.

Feature Selection

Feature selection is the process of selecting the subset of the most informative attributes to improve the performance of the model. RF has been used to distinguish the most favorable features by the feature importance. The feature importance is a way to select the desired and relevant features. To that end, RF select a subset of features.

Random Forest

RF is an ensemble learning method consisting of multiple DTs, and each tree of this method can calculate the importance of features. This algorithm benefits from the bootstrap aggregation method which can improve the performance of each DT. A DT with M leaves divides the feature space into N regions and the function f for prediction is:

$$f(x) = \sum_{n=1}^{N} C_n \pi(x, Rn)$$

Here R_n is a region appropriate to n; C_n is a constant suitable to n:

$$\pi(x, Rn) = \begin{cases} 1 & if\ x \epsilon R_n \\ 0 & otherwise \end{cases}$$

It should be noted that the last conclusion is made from the majority vote of all trees.

After applying the RF, it collects the feature importance values via the feature_importances attribute. To interpret easy, the plot of the importances was depicted in Fig. 3. It is worth mentioning that in this method the relative values of the computed importances are taken into account.

Open Science Index, Bioengineering and Life Sciences Vol:17, No:8, 2023 publications.waset.org/10013232.pdf

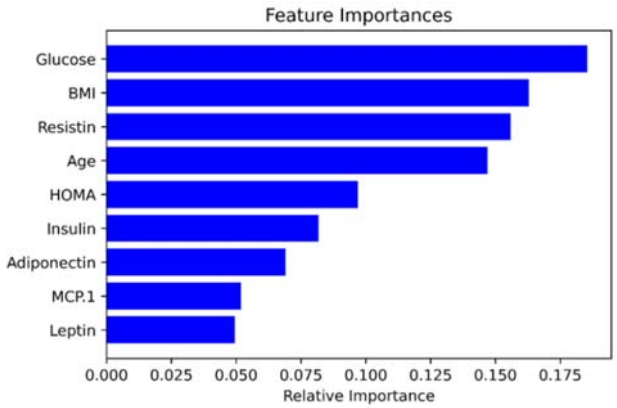

Fig. 3 Feature importances using RF model

As it was plotted in Fig. 3, among the features (attributes), glucose has the highest importance over the others. The four highest features had been selected for further investigations which means glucose, BMI, resistin, and age.

AI-Based Classifiers

In this stage, ML-based classifiers were implemented on the selected data. In this regard, the most popular ML methods have been studied: DT, KNN, XGBoost, SVM, NB, and LR. Furthermore, NN-based classifier (MLP-Classifier) was also investigated. First, DT was applied to the dataset, and the rest of the methods have been applied. The dataset was divided into training and testing data, the data were split with a test size of 20% of the whole dataset.

Decision Tree

The two Gini and Entropy indexes were selected as criteria. The results showed that there is accuracy of 70.8, and 83.3% for entropy, and gini indexes, respectively.

KNN

The KNN is non-parametric method that can be used for classification, and regression. In this method, with using of distance, and proximity, the neighbors of a point are established. The confusion matrix was obtained and displayed in Fig. 4.

The number of neighbors was optimized and set to 7. The accuracy of the KNN was calculated and accuracy of 83% was found for the designed algorithm.

Naïve Bayes

Bayesian classifier is an example of statistical classifier. Evaluation of the posterior probability value of P(y|x), each of the class of y, with an object x have been investigated. The confusion matrix was depicted in Fig. 4, and accuracy of 66.6% was found.

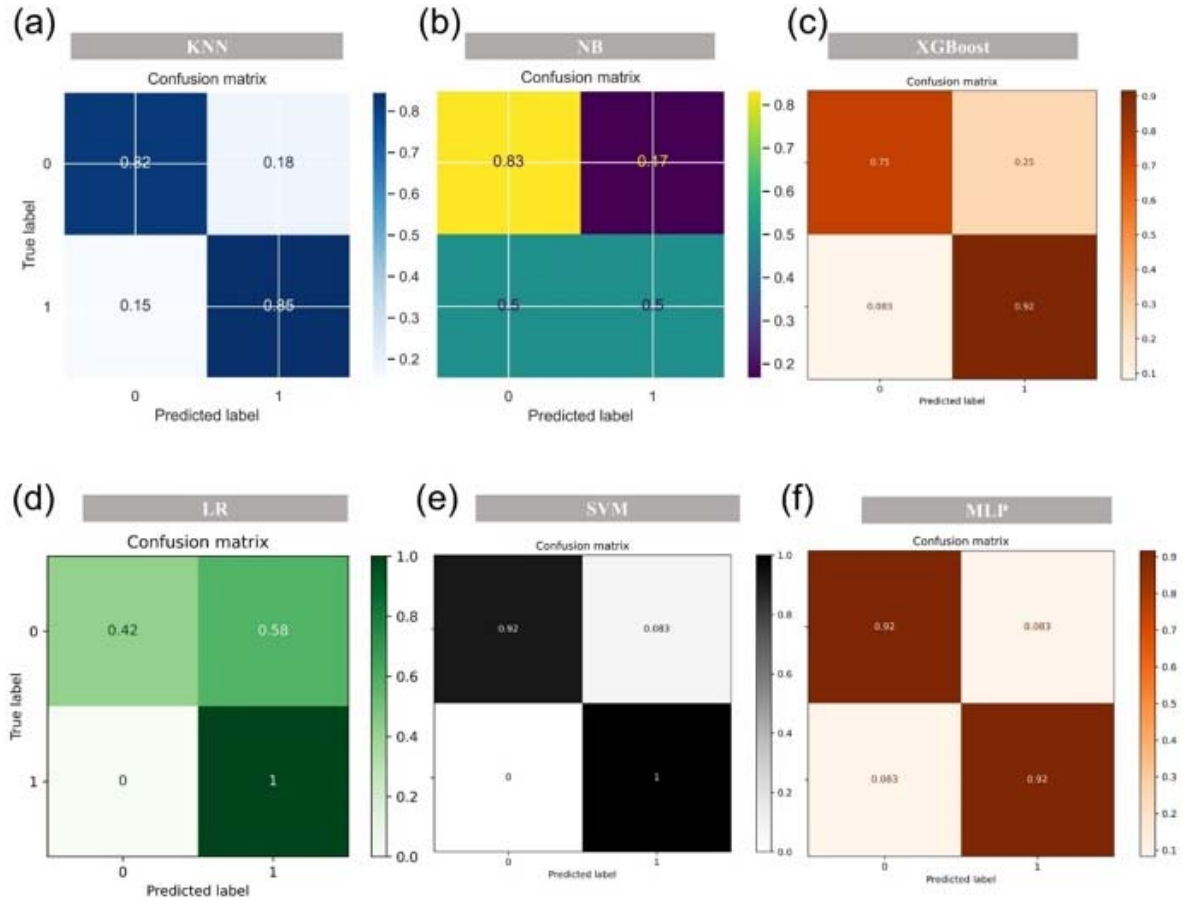

Fig. 4 Confusion matrix for different algorithms: (a) KNN, (b) NB, (c) XGBoost, (d) LR, (e) SVM, and (f) MLP

Open Science Index, Bioengineering and Life Sciences Vol:17, No:8, 2023 publications.waset.org/10013232.pdf

Open Science Index, Bioengineering and Life Sciences Vol:17, No:8, 2023 publications.waset.org/10013232.pdf

XGBoost

The XGBoost is a highly scalable end-to-end tree boosting system used in ML for classification. The confusion matrix was obtained and exhibited in Fig. 4, and accuracy of 83.3% was found.

Logistic Regression

The LR model, also known as logit is one of the common models for classification. The confusion matrix was illustrated in Fig. 4, and accuracy of 86.1% was found.

SVM

SVM is a powerful method for classification, among the kernel functions for the model, RBF was selected and the accuracy of 96% was disclosed for the testing dataset. The confusion matrix was demonstrated in Fig. 4.

NN Classifier

The MLP classifier as one of the supervised machines learning based methods was employed. For the given data, the confusion matrix was achieved and shown in Fig. 4. Moreover, the accuracy of 91.6% was calculated.

With the aim to compare results obtained by different classification models, Table I is illustrated to provide a statistical comparison between the studied methods.

TABLE I
COMPARISON OF ACCURACY OF THE STUDIED ALGORITHMS

Methods	Accuracy (100%)
DT	83.3
KNN	83
NB	66.6
XGBoost	83.3
LR	86.1
SVM	96
MLP	91.6

All in all, the general results obtained from the algorithms showed that, the SVM, MLP, and LR methods have the highest accuracy, respectively. Moreover, it was perceived that selecting the appropriate features is of great importance and could improve the accuracy. For this purpose, RF was utilized to extract the favorable attributes out of all nine features.

IV. CONCLUSIONS

Due to the complication and high mortality of BC among females, diagnosis precision is critical. By using preprocessing methods such as standard deviation, min-max normalization, a preliminary step was applied to the dataset. Next, features have been selected using the RF algorithm, and four attributes that have the highest importances were achieved. Four features; glucose, BMI, resistin, and age, were chosen for computations. To classify the processed data, different AI-based classifiers have been selected. The classification was performed with DT, KNN, NB, XGBoost, LR, SVM, and MLP algorithms. The results divulged that SVM can give the highest accuracy with a value of 96%, followed by the MLP method with 91.6% accuracy. At last, but not least, the studied procedure could provide a practical way to diagnose BC in its early stages, and it can open up a new avenue to study different cancers.

REFERENCES

[1] J. M. Jerez-Aragonés, J. A. Gómez-Ruiz, G. Ramos-Jiménez, J. Muñoz-Pérez, and E. Alba-Conejo, "A combined neural network and decision trees model for prognosis of breast cancer relapse," *Artificial intelligence in medicine,* vol. 27, no. 1, pp. 45-63, 2003.

[2] L. A. Torre, R. L. Siegel, E. M. Ward, and A. Jemal, "Global cancer incidence and mortality rates and trends—an update," *Cancer Epidemiology and Prevention Biomarkers,* vol. 25, no. 1, pp. 16-27, 2016.

[3] C.-W. Chou, Y.-M. Huang, Y.-J. Chang, C.-Y. Huang, and C.-S. Hung, "Identified the novel resistant biomarkers for taxane-based therapy for triple-negative breast cancer," *International journal of medical sciences,* vol. 18, no. 12, p. 2521, 2021.

[4] R. Roslidar *et al.*, "A review on recent progress in thermal imaging and deep learning approaches for breast cancer detection," *IEEE Access,* vol. 8, pp. 116176-116194, 2020.

[5] A. V. Berumen, G. J. Moyao, N. M. Rodriguez, A. M. Ilbawi, A. Migliore, and L. N. Shulman, "Defining priority medical devices for cancer management: a WHO initiative," *The Lancet Oncology,* vol. 19, no. 12, pp. e709-e719, 2018.

[6] O. Ginsburg *et al.*, "Breast cancer early detection: A phased approach to implementation," *Cancer,* vol. 126, pp. 2379-2393, 2020.

[7] M. d. F. O. Baffa and L. G. Lattari, "Convolutional neural networks for static and dynamic breast infrared imaging classification," in *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2018: IEEE, pp. 174-181.

[8] H.-J. Chiu, T.-H. S. Li, and P.-H. Kuo, "Breast cancer–detection system using PCA, multi-layer perceptron, transfer learning, and support vector machine," *IEEE Access,* vol. 8, pp. 204309-204324, 2020.

[9] S. Y. Siddiqui *et al.*, "IoMT cloud-based intelligent prediction of breast cancer stages empowered with deep learning," *IEEE Access,* vol. 9, pp. 146478-146491, 2021.

[10] K. Kerlikowske, D. Grady, S. M. Rubin, C. Sandrock, and V. L. Ernster, "Efficacy of screening mammography: a meta-analysis," *Jama,* vol. 273, no. 2, pp. 149-154, 1995.

[11] R. Greenberg, Y. Skornick, and O. Kaplan, "Management of breast fibroadenomas," *Journal of general internal medicine,* vol. 13, no. 9, pp. 640-645, 1998.

[12] P. V. de Campos Souza, Y.-K. Wang, and E. Lughofer, "Knowledge extraction about patients surviving breast cancer treatment through an autonomous fuzzy neural network," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2020: IEEE, pp. 1-8.

[13] H. Pham and D. H. Pham, "A novel generalized logistic dependent model to predict the presence of breast cancer based on biomarkers," *Concurrency and Computation: Practice and Experience,* vol. 32, no. 1, p. e5467, 2020.

[14] A. Aasi, S. E. Bajgani, and B. Panchapakesan, " A first-principles investigation on the adsorption of octanal and nonanal molecules with decorated monolayer WS_2 as promising gas sensing platform," *AIP Advances,* vol. 13, no. 2, p. 025157, 2023.

[15] A. Aasi, E. Aasi, S. Mehdi Aghaei, and B. Panchapakesan, "Green Phosphorene as a Promising Biosensor for Detection of Furan and p-Xylene as Biomarkers of Disease: A DFT Study," *Sensors,* vol. 22, no. 9, p. 3178, 2022.

[16] A. Aasi, S. Mehdi Aghaei, and B. Panchapakesan, "Noble Metal (Pt or Pd)-Decorated Atomically Thin MoS_2 as a Promising Material for Sensing Colorectal Cancer Biomarkers Through Exhaled Breath," *International Journal of Computational Materials Science and Engineering,* p. 2350014, 2023, doi: https://doi.org/10.1142/S2047684123500148

[17] R. T. Chlebowski *et al.*, "Predicting risk of breast cancer in postmenopausal women by hormone receptor status," *JNCI: Journal of the National Cancer Institute,* vol. 99, no. 22, pp. 1695-1705, 2007.

[18] A. W. Opstal-van Winden *et al.*, "A bead-based multiplexed immunoassay to evaluate breast cancer biomarkers for early detection in pre-diagnostic serum," *International journal of molecular sciences,* vol. 13, no. 10, pp. 13587-13604, 2012.

[19] K. D. Cole, H. J. He, and L. Wang, "Breast cancer biomarker measurements and standards," *PROTEOMICS–Clinical Applications,* vol. 7, no. 1-2, pp. 17-29, 2013.
[20] J. G. Santillán-Benítez *et al.*, "The tetrad BMI, leptin, leptin/adiponectin (L/a) ratio and CA 15-3 are reliable biomarkers of breast cancer," *Journal of clinical laboratory analysis,* vol. 27, no. 1, pp. 12-20, 2013.
[21] M. Dalamaga, G. Sotiropoulos, K. Karmaniolas, N. Pelekanos, E. Papadavid, and A. Lekka, "Serum resistin: a biomarker of breast cancer in postmenopausal women? Association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters," *Clinical biochemistry,* vol. 46, no. 7-8, pp. 584-590, 2013.
[22] G. Khakpour, A. Pooladi, P. Izadi, M. Noruzinia, and J. Tavakkoly Bazzaz, "DNA methylation as a promising landscape: A simple blood test for breast cancer prediction," *Tumor Biology,* vol. 36, no. 7, pp. 4905-4912, 2015.
[23] J. Crisostomo *et al.*, "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," *Endocrine,* vol. 53, no. 2, pp. 433-442, 2016.
[24] A. Assiri, H. F. Kamel, and M. F. Hassanien, "Resistin, visfatin, adiponectin, and leptin: risk of breast cancer in pre-and postmenopausal Saudi females and their possible diagnostic and predictive implications as novel biomarkers," *Disease markers,* vol. 2015, 2015.
[25] J.-H. Kang, B.-Y. Yu, and D.-S. Youn, "Relationship of serum adiponectin and resistin levels with breast cancer risk," *Journal of Korean medical science,* vol. 22, no. 1, pp. 117-121, 2007.
[26] H. Kobeissi, S. Mohammadzadeh, and E. Lejeune, "Enhancing mechanical metamodels with a generative model-based augmented training dataset," *Journal of Biomechanical Engineering,* vol. 144, no. 12, p. 121002, 2022.
[27] S. Mohammadzadeh and E. Lejeune, "Predicting mechanically driven full-field quantities of interest with deep learning-based metamodels," *Extreme Mechanics Letters,* vol. 50, p. 101566, 2022.
[28] H. Pham and D. H. Pham, "A Median-Based Machine-Learning Approach for Predicting Random Sampling Bernoulli Distribution Parameter," *Vietnam Journal of Computer Science,* vol. 6, no. 01, pp. 17-28, 2019.
[29] E. Aličković and A. Subasi, "Breast cancer diagnosis using GA feature selection and Rotation Forest," *Neural Computing and applications,* vol. 28, no. 4, pp. 753-763, 2017.
[30] M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2017: IEEE, pp. 226-229.
[31] N. Khuriwal and N. Mishra, "Breast cancer diagnosis using deep learning algorithm," in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, 2018: IEEE, pp. 98-103.
[32] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in *2018 International Conference on Robots & Intelligent System (ICRIS)*, 2018: IEEE, pp. 157-160.
[33] Q. Wuniri, W. Huangfu, Y. Liu, X. Lin, L. Liu, and Z. Yu, "A generic-driven wrapper embedded with feature-type-aware hybrid Bayesian classifier for breast cancer classification," *IEEE Access,* vol. 7, pp. 119931-119942, 2019.
[34] P. Ghosh, S. Azam, K. M. Hasib, A. Karim, M. Jonkman, and A. Anwar, "A performance based study on deep learning algorithms in the effective prediction of breast cancer," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021: IEEE, pp. 1-8.
[35] M. Patricio, J. Pereira, J. Crisostom, P. Matafome, R. Seiça, and F. Cramelo, "Breast Cancer Coimbra Data Set," *Web site: https://archive.ics.uci.edu/ml/datasets/Breast+ Cancer+ Coimbra,* 2018.
[36] V. J. Silva Araújo, A. J. Guimarães, P. V. de Campos Souza, T. S. Rezende, and V. S. Araújo, "Using resistin, glucose, age and BMI and pruning fuzzy neural network for the construction of expert systems in the prediction of breast cancer," *Machine Learning and Knowledge Extraction,* vol. 1, no. 1, pp. 466-482, 2019.
[37] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics,* vol. 16, no. 10, pp. 906-914, 2000.
[38] R. Molania *et al.*, "Removing unwanted variation from large-scale RNA sequencing data with PRPS," *Nature Biotechnology,* pp. 1-14, 2022.
[39] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Computational intelligence and neuroscience,* vol. 2021, 2021.

Open Science Index, Bioengineering and Life Sciences Vol:17, No:8, 2023 publications.waset.org/10013232.pdf