Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

OPTIMUM FEATURE SELECTION BASED BREAST CANCER PREDICTION USING MODIFIED LOGISTIC REGRESSION MODEL

D. VETRITHANGAM[1], P.SHRUTI[2], B. ARUNADEVI[3], R. HIMABINDU[4], P. NARESH KUMAR[5], A. RAMESH KUMAR[6]

[1]Associate Professor, Chandigarh University, Department of CSE, Mohali, India

[2]Assistant Professor, Graphic Era Deemed to be University, Department of CSE, Dehradun, India

[3]Professor, Dr. N.G.P Institute of Technology, Department of ECE, Coimbatore, India

[4]Assistant Professor, Bapatla Engineering College, Department of Cyber Security, Andhra Pradesh, India

[5]Assistant Professor, K.G Reddy College of Engineering and Technology, Department of CSE, India

[6]Professor, K.S Rangasamy College of Technology, Department of Mechatronics Engineering, India

E-mail: [1]vetrigold@gmail.com, [2]shruti.p9753@gmail.com, [3]arunadevi@dmgpit.ac.in, [4]himabindu.r@becbapatla.ac.in, [5]pnrshkumar@gmail.com, [6]arameshkumaar@gmail.com

ABSTRACT

Patients with breast cancer are more likely to experience severe health issues and have a higher mortality rate. One of the main reasons for cancer-related deaths in women is breast cancer (BC). Early diagnosis of breast cancer enables patients to obtain proper care, enhancing their chance of survival. The main explanation could be that different breast densities and technical imaging quality issues cause radiologists to misinterpret concerning lesions, increasing the false-positive and negative) ratio. In this work, a new optimum feature selection-based model is developed to efficiently predict breast cancer using a modified logistic regression model. Our proposed model consists of two phases: a) feature selection and b) prediction. As a first step, preprocessing is done on the dataset to find the missing values and remove the unwanted noise, outliers, and so on. In this research work, the first dataset with 568 numbers of data and 30 numbers of features and the second dataset with 952 numbers of data and 26 numbers of features are considered for diagnosis and analysis. To select the features from the dataset's N features, an improved grey wolf population algorithm is used. Hence, 26 sets of features are selected for further processing. Our proposed model performed well on both datasets, with 92.9% and 93.38% accuracy for the first and second datasets, respectively. The novelty of this research work is to provide the best accuracy in disease diagnosis and prediction by selecting the optimum based on meaningful features.

Keywords: *Logistic Regression, Accuracy, Breast Cancer, Machine Learning, Prediction.*

1. INTRODUCTION

A cancerous tumor develops as a result of aberrant cell proliferation that infects the body's surrounding tissues. Tumors come in two varieties: benign and malignant. Noncancerous cells that grow locally but they do not spread across the body make up a benign tumor. On the other hand, cancerous cells make up a malignant tumor that has the capacity to proliferate uncontrollably, spread throughout the body, and invade the tissues. According to a 2021 estimate, the human body is made up of trillions of tiny cells, and breast cancer is one of the main reasons women die. During cellular reproduction, new cells continuously replace old ones. Normally, this process is homogeneous, but aberrant growth might result from excessive cell production. This unnatural growth may lead to cancer [1]. Cancer has been a fast-spreading, frequently lethal disease. Women are most likely to develop breast cancer. Therefore, a lower female death rate can result from the early detection of breast cancer. Breast cancer can be found with either mammography or a biopsy. The most often used type of technique for detecting breast cancer is mammography. This test, which is administered by radiologists to diagnose breast cancer, lowers mortality rates by 25%. Although

mammography is a useful tool, it can be difficult to interpret the images [2]. Typically used in histopathological image processing are machine learning techniques such as feature extraction, unsupervised learning, segmentation, and supervised learning [3]. This study focuses on breast cancer, one of many different types of cancer. Around 627,000 women worldwide pass away from breast cancer each year, making up roughly 15% of all cancer-related deaths [4]. According to reports, when breast cancer is detected earlier, the prognosis is favorable, and the endurance value has inevitably increased [5]. Prior to the specialists' visual review, the tissue removed during the biopsy is frequently stained with hematoxylin and eosin (H&E). Relevant areas of tissue scans on entire slides are evaluated during this process[6] The diagnosis process utilizing H&E-stained samples is not simple, and there is a 75 percent diagnostic agreement on average amongst specialists. Highly skilled pathologists must work extremely hard to manually examine histology images. The use of CAD has increased dramatically over the past ten years due to gains in processing power that are quite large and advancements in deep learning, particularly Convolutional Neural Networks, as well as the subjective nature of how morphological criteria are used in traditional classification (CNN)[7]. Detecting normal and abnormal mammograms using convolutional neural networks is suggested. Using the MIAS dataset for mammograms, this deep learning algorithm extracts features from subdivided aberrant classes and applies them to the normal class. In order to remove noise components that could reduce the accuracy of the overall network, several filter sizes and preprocessing algorithms were utilized on the real data [8][9]. The technique allows for the visualization of blood vessels and, consequently, of a tumor's optically induced angiogenesis contrast. Light remains the inquisitive energy, but unlike in DOT, photons are neither counted nor used to create the observed signal. Stress (acoustic) vibrations that are barely diffused and depreciated in soft tissue are instead monitored to provide greater clarity [10]. For the diagnosis of normal and atypical breast cancer, a deep learning technique was used with the ResNet50 network and VGG16, which worked well. The best classification accuracy result was achieved by VGG16, which had 94% accuracy [11]. Breast ultrasound lesion identification uses three different approaches, including a patch-oriented LeNet and a transfer learning technique with a pre-trained FCN-AlexNet [12].

The following are some of the research challenges in machine and deep learning-based breast cancer prediction:

Data quality and availability: The quality and quantity of available data determine the precision and dependability of machine learning and deep learning models. Due to privacy issues and data access restrictions, obtaining large, high-quality datasets is a challenge in breast cancer research.

Interpretability: In medical applications, where judgements made by machine learning models can have a big effect on health outcomes, interpretability is especially crucial.

Imbalanced datasets: Breast cancer datasets are frequently unbalanced, which means that there is not an equal amount of positive and negative samples in the dataset. As a result, models may be skewed to favour the dominant class while underperforming the minority class. An optimal feature selection-based breast cancer prediction model is developed to address the issues caused by interpretability issues and imbalanced datasets. Breast cancer prediction models may encounter a number of problems if their feature selection is subpar, such as:

Overfitting: If the model is too complicated, it may fit the noise in the data rather than the underlying patterns, which has a negative impact on how well it performs on new data. Poorly predicted accuracy may result from under fitting, where the model is too simplistic and fails to account for the intricate relationships in the data.

High dimensionality: The model could have a lot of pointless or redundant characteristics, which might make computation more difficult and make the model harder to understand.

Bias and instability: By incorporating noisy or irrelevant features, the model may be unstable or biased towards some features.

The ability to identify the most pertinent and instructive aspects connected to the disease makes optimal feature selection essential for breast cancer prediction. Breast cancer is a complicated illness with many contributing components; thus, figuring out the most important characteristics can help with prediction and the biology of the disease. A predictive model's interpretability can be improved, and the risk of overfitting can be decreased, by removing duplicate or irrelevant features and reducing the dimensionality of the data.

By determining the most pertinent and instructive features for the model, boosting predicted accuracy, reducing complexity, and improving model interpretability, optimal feature selection can assist in resolving these problems.

The objectives of this research are: a) to select optimal feature sets for disease diagnosis; b) to address the high-dimensionality issues in model computation; c) to improve the accuracy of disease diagnosis and prediction; d) to resolve the interpretability issues and make the model efficient in decision-making situations.

2. RELATED WORK

Numerous data points are input, and the machine learning model analyses the data, and then, using that trained model, we can forecast the future. Automated learning is called machine learning, and algorithms are created to gain knowledge from previous datasets. The key machine learning techniques for predicting breast cancer include the ones listed below:

2.1 Artificial Neural Network (ANN)

An algorithm used frequently in data mining is the artificial neural network. Input, hidden, and output layers are involved in creating a neural network. This method is employed to extract an excessively complex pattern [13]. In order to increase the model's performance and dependability, the BCNet was employed to recognize human peripheral blood cells [14].

In order to increase classification reliability, the classification system is constructed as a serial integration of two separate ensembles of arbitrary subspace classifiers. A group of SVM classifiers in the first group divides the original K-type classification into a count of K-2 type-class problems. A multi-layer perceptron ensemble makes up the second ensemble, which concentrates on the first ensemble's rejected samples [15].

2.2 Logistic Regression

The algorithm is supervised learning and has more dependent variables. This algorithm's output takes the form of a binary number. When applied to appropriate data, logistic regression might produce a continuous result. The statistical model used in this approach uses binary variables [16][17]. When the result variable is binary, logistic regression can be a potent analytical tool. The use of logistic regression has grown during the past ten years. This popularity can be attributed to the ease with which researchers can use sophisticated statistical programmers to carry out thorough studies of this method [16].

2.3 K-Nearest Neighbor (KNN)

The KNN algorithm is a technique for categorizing objects that depends on nearby training samples in the feature space. KNN is a form of instance-oriented learning where all computation is postponed until classification and the function is only locally approximated [18]. In order to recognize patterns, this method is utilized. It is an effective strategy for predicting breast cancer. Every class received the same amount of attention in order to spot the trend. K Nearest Neighbor extracts the related highlighted data from a large dataset [19].

2.4 Decision Tree (DT)

Decision trees are built using classification and regression models. There are fewer subsets of the data set. These smaller pieces of data can be used to make predictions with the best degree of accuracy possible by employing a decision tree that incorporates CART [20]. With a root node, branch node, and leaf node structure, the classification is built using the supervised decision tree technique. The decision tree is constructed sequentially by breaking the complete dataset into various subsets. J48 builds a decision tree using a top-down, greedy search of all potential branches [21].

2.5 Support Vector Machine

This approach to supervised learning is utilized for both classification and regression issues. It uses mathematical and theoretical functions to address the regression issue. When making predictions using a huge dataset, it offers the highest accuracy rate. It is a powerful machine learning algorithm that includes the concepts of 3D and 2D modeling [22]. The support vector machine was utilized to decrease variance and improve diagnosis accuracy in order to address the attainment limitations of individual models [23]. The SVM method for breast cancer assessment and diagnosis was put into action on the Wisconsin Diagnostic and Prognostic Breast Cancer datasets reported in the literature. The SVM technique

combines unusually well for both issues, with high accuracy, sensitivity, and specificity indices [24].

2.6 Support Vector Machine

Data can be divided into small groups using the clustering method K mean. Algorithms are used to determine the degree of similarity between various data points. To evaluate a large dataset, data points must precisely belong to at least one cluster[25]. The concept of resemblance is used to identify clusters. Similar data point clusters all belong to the same family. Each data point in the C-mean algorithm corresponds to a single cluster. Disease prediction and medical picture segmentation are its main uses [26].

2.7 Convolutional Neural Network

Deep learning is a type of feature depiction [27] that uses the raw data to automatically find features appropriate for a specific job. The extractors of features are task-specific in that they aren't always bound by the same set of rules [28]. A CNN-based method for classifying H & E-disordered histological sets of images for breast cancer is designed. The network learns all pertinent traits, minimizing the requirement for domain expertise. Images are divided into categories such as healthy tissue, lesions of benign origin, in situ, and aggressive malignancy. Another option is a binary classification as malignancy or non-malignancy. To do this, the network's architecture is built to excerpt data from many relaxant scales, such as nuclei and serious tissue conformation. The network is evaluated on a different collection of photos after being pre-trained on an enlarged area dataset. Scale-oriented network design and dataset augmentation have both been proven to be crucial for the approach's success [9]. CNN has been used to diagnose the malaria disease, and it produced a detection rate of 97.06 percent [29]. An integrated convolutional neural network was used to differentiate the COVID-19 images from the pneumonia images [30]

Artificial neural networks focus on providing an all-encompassing, easy way to learn real-valued, discrete-valued, and vector-valued functions by using examples (ANNs). Gradient descent is a technique used by algorithms like backpropagation to modify metrics to best suit an input-output training set. ANN learning has been effectively used to solve issues including understanding visual sceneries, speech recognition, and learning robot control strategies because it is robust to faults in the training data [13]. A breast cancer classification and learning task and learning framework for doing multi-tasks are being used because some breasts have both malignant and benign outcomes [31].

3. PROPOSED MODEL

3.1 Dataset Preprocessing and Feature Selection

We have considered two types of datasets in this work. The first dataset is called breast_cancer.csv, which is downloaded from Kaggle and available in the UCI Machine Learning Repository, and it consists of 568 numbers of data. It consists of various features, and the second dataset consists of 952 numbers of data. In data preprocessing, the missing values are checked in every row and column of the dataset. The ID number, diagnosis (M = malignant, B = benign), and attribute information are available in the dataset. Actual-valued elements calculated for every cell nucleus consist of the radius (mean distance from the center to a specific target value on the perimeter), texture (average grayscale value's standard deviation), area, perimeter, smoothness (varying locally in radius lengths), compactness (perimeter2/area-1.0), concavity (intensity of the contour's concave areas), concave points (how many of the contour's concave areas there are), symmetry, and fractal dimension ("coastline approximation"-1). For every picture, the mean, standard error, and "worst" or "worst" feature—the mean of the 3 numbers of the largest values—were computed, producing a count of 30 features. For instance, fields 3 and 13 represent the mean radius, the radius SE, and the worst radius, respectively. Each feature value has four significant digits added to it.

Our proposed model consists of two phases, which are: a) feature selection and b) prediction using a supervised machine learning algorithm. The dataset is preprocessed to remove noise and outliers, and then a certain number of features are selected using an improved grey wolf population algorithm. Feature selection (FS) is the selection of a small number of features without applying any transformations. Any classification system must prioritize pertinent features. As a result of the significant number of extracted characteristics, unnecessary features are frequently

created. The performance and strength of the system can be significantly impacted by these numerous characteristics, also referred to as "the curse of dimensionality." Furthermore, it is thought to be a significant threat to current learning strategies. By choosing pertinent characteristics, we can speed up training by streamlining the learned classifier. As choosing pertinent features is done in order to optimize the issue of feature selection, a specific fitness value can be considered an optimization problem. Numerous research studies have looked at this issue as an optimization issue. Some of them have the goal of maximizing the chosen attributes and using classification accuracy as a fitness function. Meta-heuristic algorithms that draw inspiration from nature are currently the most popular algorithms used to solve optimization problems.

Improved grey wolf population algorithm:

1: Set the initial parameters: Feature size (FS), maximum iterations (MI)
2: Generate the initial grey wolf population G_i (i = 1, 2 . . . n)
3: Assess the fitness of each search feature
4: Identify the best features based on their fitness
5: Ga = the best fitting search feature and its value
6: Gb = the second-level fittest search feature and its value
7: Gc = the third-level fittest search feature and its value
8. Gd = the Fourth level fittest search feature and its value
9: while (t < MI) do
10: For each search feature, do
11: Calculate the accuracy
12: Use Eq. (13) to re-position (update) the current search feature
13: end for
14: Evaluate the fitness of a new feature set of features or values.
15: Update the best four solutions: Ga,Gb,Gc & Gd
16: t = t + 1
17: end while
18: return the best result Ga.

We have used an improved grey wolf population technique for optimum feature selection to select the important and meaningful features among the N number of features in the dataset. Overall, there are 30 different features available in the dataset, and our grey wolf population algorithm selected 26 different features for further processing, such as training and testing and the prediction of breast cancer. Multiple times, the feature selection process is iterated, and the best four possible sets of features are selected. The breast cancer disease prediction is done using a modified logistic regression model, as shown in Figure 1. Python has a machine learning library called Scikit-learn (sklearn). It includes a number of methods for clustering, classification, and regression, including SVMs, k-means, gradient boosting, DBSCAN, and random forests. It is made to work with SciPy, Numpy, and Python. The dataset is trained and tested with a logistic regression model. A machine learning model is used to find out the relationship between the data. The dataset is split into training data and test data. The model is trained, and a trained model is called a trained logistic regression model.

All the required models and metrics are imported from Sklearn. The breast cancer dataset is read using Pandas pd. The terms "benign" and "malignant" are used. The first dataset with 568 records of data and 30 records of features is chosen and analyzed by applying our proposed model. The dataset consists of 212 malignant cases and 357 benign cases, and our proposed model produced a 92.98 percentage of accuracy. The second dataset, with 952 numbers of data and 26 numbers of features, is considered for diagnosis and analysis. In total, 28 columns are available, and 26 of those columns are featured. The first five rows of the data frame are listed as shown in figure 2. Here the data frame name is NC, and the first 10 values are the mean of the radius, concave points, compactness, area, perimeter, concavity, smoothness, and so on.

The diagnosis column is added to the label of the data frame because the diagnosis is the array that contains 0s and 1s. Figure 3 represents the last five rows of the data frame, which consist of 1206.0, 928.2, 1169, 602.4, and 1207.0 as area mean values, where 602.4 and 1207.0 are the minimum and maximum values, respectively. The maximum difference between the radius_worst values is 9.31. In terms of value, smoothness and compactness are less sensitive. Values that are not present but would have significance if they were are referred to as "missing data." A missing sequence, an incomplete feature, a missing file, an incomplete piece of information, or a mistake during data entry can all be considered missing data. In the current world, missing data is common in datasets. It is necessary to modify missing data fields before using them with data so that analysis

and modelling can be done on them. As shown in Figure 4, our dataset has no missing values.

```
NC.isnull().sum()
```

```
id                        0
diagnosis                 0
radius_mean               0
texture_mean              0
perimeter_mean            0
area_mean                 0
smoothness_mean           0
compactness_mean          0
concavity_mean            0
concave points_mean       0
symmetry_mean             0
fractal_dimension_mean    0
radius_se                 0
texture_se                0
perimeter_se              0
area_se                   0
smoothness_se             0
compactness_se            0
concavity_se              0
concave points_se         0
symmetry_se               0
fractal_dimension_se      0
radius_worst              0
texture_worst             0
perimeter_worst           0
area_worst                0
smoothness_worst          0
compactness_worst         0
concavity_worst           0
concave points_worst      0
symmetry_worst            0
fractal_dimension_worst   0
label                     0
dtype: int64
```

Figure 4: Representation of finding the missing values in the data set.

3.2 Prediction System

We have also built a prediction system that predicts whether a system has malignant (cancerous) cases or benign (non-cancerous) cases. The statistical measures for the dataset are provided by describe () method. The very first statistical measure is count, which describes how many columns are in each column.

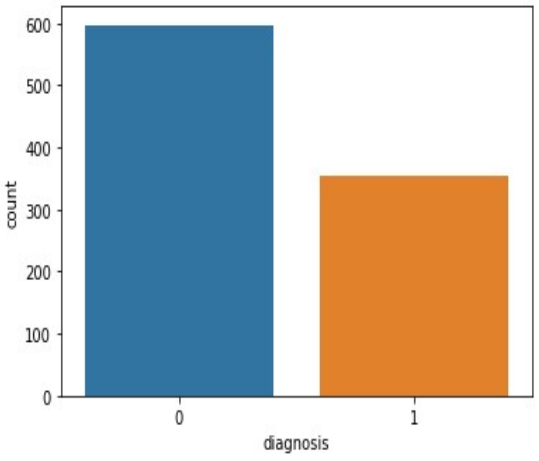

Figure 6: Count Plot for visualization of malignant and benign cases

The mean value is the overall mean value of every column in the dataset. The 25% and 50% describe that 25% and 50% of the values of every column have a specific range of values, as shown in Figure 5, which shows that 25% of the column has a value less than 8.667 and 50% of the column has a value less than 9.08. In order to find out the number of malignant (cancerous) cases or benign (not cancerous) cases in the chosen dataset, we assumed that benign (not cancerous) cases would be 0s and malignant (cancerous) cases would be 1s in our proposed work. Among the 952 numbers of cases in the dataset, 597 (not cancerous) numbers of benign cases are represented by 0's, and 355 numbers of malignant (cancerous) cases are represented by 1's, as shown in figure 6.

The values of the malignant (cancerous) cases are higher than the benign cases. For example, the radious_mean value of a malignant case is 17.43, which is greater than the radious_mean value of a benign (not cancerous) value of 12.14. As shown in Figure 7, the above values play a vital role in determining whether a person has a malignant tumor or a benign tumor. In our system, we have split our dataset into two different types, which are training data and testing data. Eighty percent of the dataset is considered training data, and twenty percent of the dataset is treated as a testing dataset. For example, our model has predicted that this is a benign (not cancerous) value for this set of values such as the radius mean, texture mean, perimeter mean, and so on, as represented by Figure 8.

4. RESULTS AND DISCUSSION

To achieve the objectives of the research work outlined in the introduction section, this work has used an improved grey wolf population technique for optimum feature selection to select the important and meaningful features among the N number of features in the dataset, and the dataset has been properly preprocessed to improve the quality of the dataset. Our dataset has a variety of properties, and multivariate visualization is crucial for understanding interactions between various attributes. It displays the relationships between the various dataset kinds. A correlation is a signal regarding how two variables have changed over time. A correlation between the features will help the model improve its interpretability, so the relationship between the features can be analyzed in various ways.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The third row and fourth column indicate a relationship between radius_mean in the y axis and texture_mean in the x axis with a value of 0.3. The fifth column and fifth row indicate that the relationship between the same variables or features is perimeter_mean with a value of 1.0. In this research work, we have considered 26 features or attributes. The relationships between the features are represented as shown in figure 9. To put it simply, skewness is a measure of how far a random variable's probability distribution deviates from the distribution. It is mostly used for uni-variant sets of observations and visualizes them through a histogram; as there is only one observation, we select that column of the dataset. The probability distribution with no skewness is known as the "normal distribution." The distribution with the tail on its right side is said to be favorably skewed. For a positively skewed distribution, the skewness value is higher than zero. The distribution with the tail on its left side is said to be negatively skewed. For a negatively skewed distribution, the value of skewness is less than zero.

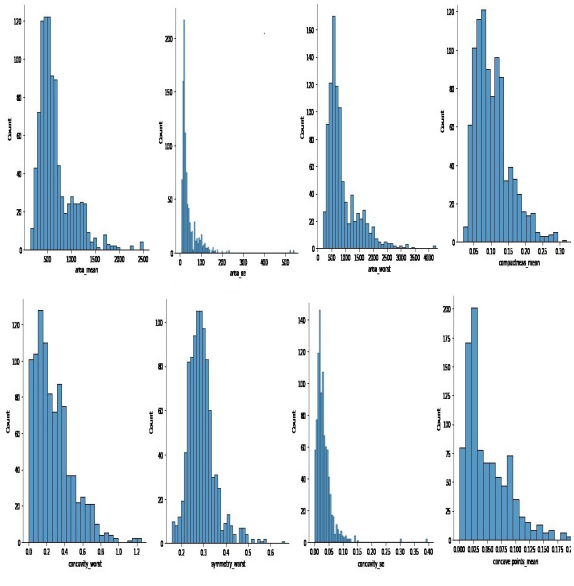

Figure 10: Visualization of features with the use of distribution plots

As shown in Figure 10, the majority of the features are right-skewed, which means that the majority of the data is on the left side and the mean value is greater than the median value. Many variables are available in the given dataset. But we have taken one variable at a time and analyzed it. Hence, it is called univariate analysis. In comparison to the median, the mean value is lower. Then it is called left-skewed data.

Box and whisker plots, also known as box plots, are an excellent graphic to use when showing how data points are distributed across a selected parameter. These graphs display the measurement ranges of the variables. This takes into account outliers, the median, the mode, and where most of the data points fit within the "box." These illustrations are useful for contrasting the distribution of various variables. Outliers are handled by the data science and feature engineering sections. The dataset affects how the logistic regression model behaves. Boxplots are a common method for visualizing data to determine whether or not outliers are present. Outliers generally appear as circles when visualized in a box and whisker plot. The horizontal lines accessible below the box represent the minimum values of the corresponding column, and the lines above and below the boxes are referred to as whiskers, as shown in figure 11.

For example, the texture_worst attribute has a minimum value of 12.02. The next box and its length are called the interquartile range (IQR). It is the range (dissimilarity) between the third quartile and the first quartile. The first quartile, 25 percent, is represented by this blue line that exists above the first horizontal line, which is termed Q1. The middle line represents the 50 percent, which is termed Q2, and the third quartile represents the 75 percent, which is termed Q3. The Q1, Q2, and Q3 values for the texture_worst column/attribute are 21.32, 25.48, and 30.06, as shown in figure 11. Similarly, the values of the min, max, Q1, Q2, and Q3 values of the texture mean are 9.71, 38.290, 16.33, 18.94, and 21.875. An outlier is a data point that lies outside the overall pattern in a distribution. The dissimilarity between Q3 and Q1 is called the interquartile range, or IQR. That is, IQR = Q3 - Q1, and any data point that falls outside of this range is considered an outlier and is treated accordingly at the lower bound: (Q1 - 1.5 * IQR) and the upper bound: (Q3 + 1.5 * IQR). An outlier is any data point that lies outside either the lower bound or the upper bound. We must understand the outliers, whether they are natural or incorrect data. If the outliers are natural, they should not be removed. The third column feature and the fourth column feature are represented using a scatter plot as shown in figure 12.

Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

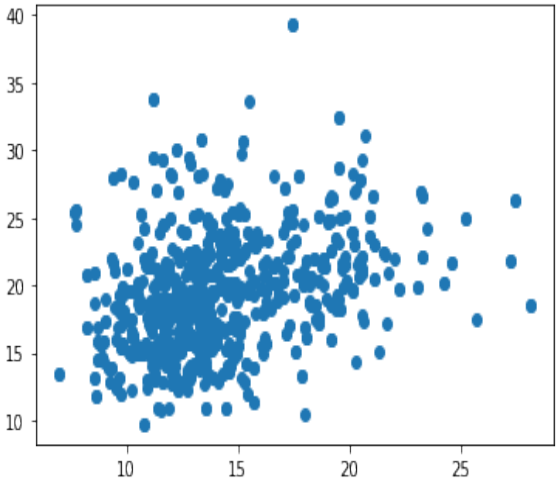

Figure 12: Visualization of two features by using a scatter plot

Our proposed model has done efficient, optimum-based feature selection, so the model has improved its interpretability in understanding the features and relationships between the features when it makes the decisions.

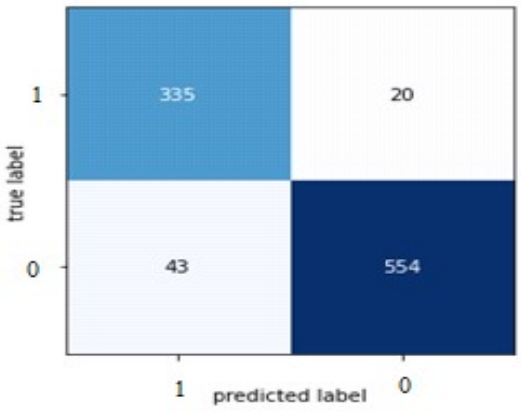

Figure 13: Confusion matrix for the results of the proposed model

Figure 13 shows the following values: true positive = 335, false positive = 20, false negative = 43, and true negative = 554. The proposed model predicted 335 cases as malignant (cancerous) cases, and they are actually malignant cases as well. The other 20 cases were predicted as cancerous, but they were actually non-cancerous or benign cases. All 554 cases predicted as benign (non-cancerous) cases are actually benign cases, while 43 cases are predicted as false negatives.

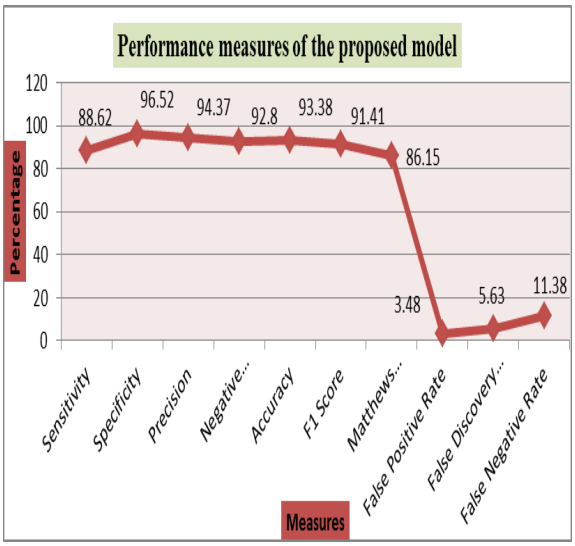

Figure 14: Performance of the proposed model in terms of performance matrix

The proposed model produced the sensitivity, specificity, precision, negative predicted value, accuracy, F1 score, Mathew correlation coefficient, false discovery rate, false positive rate, and false negative rate of 88.62%, 96.52%, 94.37%, 92.8%, 93.38%, 91.41%, 86.15%, 3.48%, 5.63%, and 11.38%, respectively, as shown in figure 14.

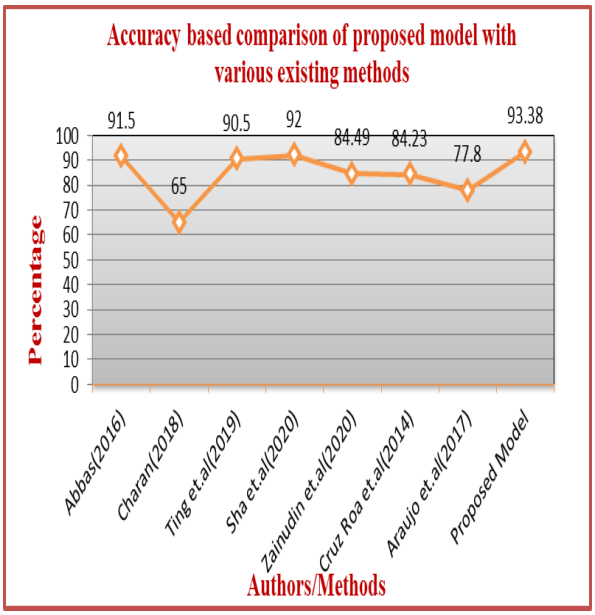

Figure 15: Comparison of the proposed model with existing model in terms of accuracy

The accuracy score is used to evaluate our model or to determine how many correct predictions are made by the newly proposed model. Abbas (2016)'s research group achieved 91.5% accuracy in his Deep-CAD system; the difference in accuracy value from the proposed model is 1.88%. Sha et al. (2020) achieved 92% accuracy with a 1.38% difference. In addition to that, the proposed model is compared with the existing methods in terms of sensitivity, specificity, precision, and AUC, as shown in figure 15. The sensitivity value of the proposed model is 88.62, which is less than other sensitivity values of existing research groups such as Abbas (2016), Charan (2018), Ting et al. (2019), and Sha et al. (2020), but the specificity of the proposed model is 96.52, which is greater than other existing methods listed above.

5. CONCLUSION AND FUTURE WORK

In this paper, a modified logistic regression model for improving the breast cancer disease prediction results on the breast cancer dataset is proposed. The purpose of this model is to help medical officers or doctors diagnose and predict breast cancer or tumors with high accuracy. The datasets were separated into two different classes: benign and malignant cases. The original dataset was preprocessed for noise removal, finding the missing values, and removing the unwanted data in the dataset. The improved grey wolf population algorithm played a vital role in selecting the features, so the modified logistic regression-based model achieved the best accuracy and specificity values when compared with other models. By doing effective dimensionality reduction on the dataset and optimum feature selection, computation difficulty has been reduced, and the model has not faced any difficulty in understanding the features and relationships between the features. The model has improved its interpretability in understanding the correlation between the features and making the decision.

Finally, it can be concluded that the feature selection-based modified logistic regression model achieved good performance compared with other existing approaches. The result showed 88.62% sensitivity, 96.52% specificity, 94.37% precision, 92.8% negative predicted value, 93.38% accuracy, 91.41% F1 score, 86.15% Mathew correlation coefficient, 3.48% false positive rate, 5.63% false discovery rate, and 11.38% false negative rate. In future work, the proposed method can be further integrated with other deep learning models to diagnose and predict other types of tumors with the highest accuracy.

REFERENCES:

[1] American Cancer Society Cancer Action Network, "Cancer Disparities: A Chartbook," Strategies, no. August, 2009.

[2] A. Chekkoury et al., "Automated malignancy detection in breast histopathological images," Med. Imaging 2012 Comput. Diagnosis, vol. 8315, p. 831515, 2012, doi: 10.1117/12.911643.

[3] J. de Matos, A. de S. Britto, L. E. S. Oliveira, and A. L. Koerich, "Histopathologic Image Processing: A Review," 2019, [Online]. Available: http://arxiv.org/abs/1904.07900.

[4] S. Manohar and M. Dantuma, "Current and future trends in photoacoustic breast imaging," Photoacoustics, vol. 16, no. April, p. 100134, 2019, doi: 10.1016/j.pacs.2019.04.004.

[5] J. K. Birnbaum, C. Duggan, B. O. Anderson, and R. Etzioni, "Early detection and treatment strategies for breast cancer in low-income and upper middle-income countries: a modelling study," Lancet Glob. Heal., vol. 6, no. 8, pp. e885–e893, 2018, doi: 10.1016/S2214-109X(18)30257-2.

[6] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological Image Analysis: A Review," IEEE Rev. Biomed. Eng., vol. 2, pp. 147–171, 2009, doi: 10.1109/RBME.2009.2034865.

[7] Y. Li, J. Wu, and Q. Wu, "Classification of Breast Cancer Histology Images Using Multi-Size and Discriminative Patches Based on Deep Learning," IEEE Access, vol. 7, pp. 21400–21408,2019,doi:10.1109/ACCESS.2019.2898044.

[8] S. Charan, M. J. Khan, and K. Khurshid, "Breast cancer detection in mammograms using convolutional neural network," 2018 Int. Conf. Comput. Math. Eng. Technol. Inven. Innov. Integr. Socioecon. Dev. iCoMET 2018 - Proc., vol. 2018-Janua, pp. 1–5, 2018, doi: 10.1109/ICOMET.2018.8346384.

[9] D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks," IEEE Access, vol. 6, pp. 24680–24693,2018,doi:10.1109/ACCESS.2018.2831280.

[10] S. Manohar and M. Dantuma, "Ac ce p us t," Biochem. Pharmacol., 2019, doi: 10.1016/j.pacs.2019.04.004.

[11] N. S. Ismail and C. Sovuthy, "Breast Cancer Detection Based on Deep Learning Technique," 2019 Int. UNIMAS STEM 12th Eng. Conf. EnCon 2019 - Proc., pp. 89–92, 2019, doi: 10.1109/EnCon.2019.8861256.

[12] M. H. Yap et al., "Automated Breast Ultrasound Lesions Detection Using Convolutional Neural Networks," IEEE J. Biomed. Heal. Informatics, vol. 22, no. 4, pp. 1218–1226, 2018, doi: 10.1109/JBHI.2017.2731873.

[13] Y. Uzun and G. Tezel, "Rule Learning With Machine Learning Algorithms and Artificial Neural Networks," J. Selcuk Univ. Nat. Appl. Sci., vol. 1, no. 2, p. pp 54-64, 2012.

[14] C. Chola et al., "BCNet: A Deep Learning Computer-Aided Diagnosis Framework for Human Peripheral Blood Cell Identification," Diagnostics, vol. 12, no. 11, p. 2815, 2022, doi: 10.3390/diagnostics12112815.

[15] Y. Zhang, B. Zhang, F. Coenen, and W. Lu, "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles," Mach. Vis. Appl., vol. 24, no. 7, pp. 1405–1420, 2013, doi: 10.1007/s00138-012-0459-8.

[16] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," J. Educ. Res., vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.

[17] Y. Yari, T. V. Nguyen, and H. T. Nguyen, "Deep learning applied for histological diagnosis of breast cancer," IEEE Access, vol. 8, pp. 162432–162448, 2020, doi: 10.1109/ACCESS.2020.3021557.

[18] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[19] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background," Int. J. Eng. Res. Appl., vol. 3, no. 5, pp. 605–610, 2013.

[20] H. Sharma and S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining," Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097,2016,doi:10.21275/v5i4.nov162954.

[21] P. Hamsagayathri and P. Sampath, "Priority based decision tree classifier for breast cancer detection," 2017 4th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2017, pp. 4–9, 2017, doi: 10.1109/ICACCS.2017.8014598.

[22] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 2049 LNAI, no. January 2001, pp. 249–257, 2001, doi: 10.1007/3-540-44673-7_12.

[23] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," Eur. J. Oper. Res., vol. 267, no. 2, pp. 687–699, 2018, doi: 10.1016/j.ejor.2017.12.001.

[24] E. Zafiropoulos, I. Maglogiannis, and I. Anagnostopoulos, "A support vector machine approach to breast cancer diagnosis and prognosis," IFIP Int. Fed. Inf. Process., vol. 204, pp. 500–507, 2006, doi: 10.1007/0-387-34224-9_58.

[25] Y. G. Li, "A clustering method based on K-means algorithm," Appl. Mech. Mater., vol. 380–384, pp. 1697–1700, 2013, doi: 10.4028/www.scientific.net/AMM.380 384.1697.

[26] L. V. Tilson, P. S. Excell, and R. J. Green, "A generalisation of the Fuzzy c-Means clustering algorithm," Remote sensing. Proc. IGARSS '88 Symp. Edinburgh, 1988. Vol. 3, vol. 10, no. 2, pp. 1783–1784, 1988, doi: 10.1109/igarss.1988.569600.

[27] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," 2015, doi: 10.1038/nature14539.

[28] Y. Jia et al., "Caffe: Convolutional architecture for fast feature embedding," MM 2014 - Proc. 2014 ACM Conf. Multimed., pp. 675–678, 2014, doi: 10.1145/2647868.2654889.

[29] K. Hemachandran et al., "Performance Analysis of Deep Learning Algorithms in Diagnosis of Malaria Disease," Diagnostics, vol. 13, no. 3, 2023, doi: 10.3390/diagnostics13030534.

[30] D. Vetrithangam, V. Indira, S. Umar, B. Pant, M. K. Goyal, and B. Arunadevi, "Discriminating the Pneumonia-Positive Images from COVID-19-Positive Images Using an Integrated Convolutional Neural Network," Math. Probl. Eng., vol. 2022, 2022, doi: 10.1155/2022/5643977.

[31] N. Wu et al., "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening," IEEE Trans. Med. Imaging, vol. 39, no. 4, pp. 1184–1194, 2020, doi: 10.1109/TMI.2019.2945514.

Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

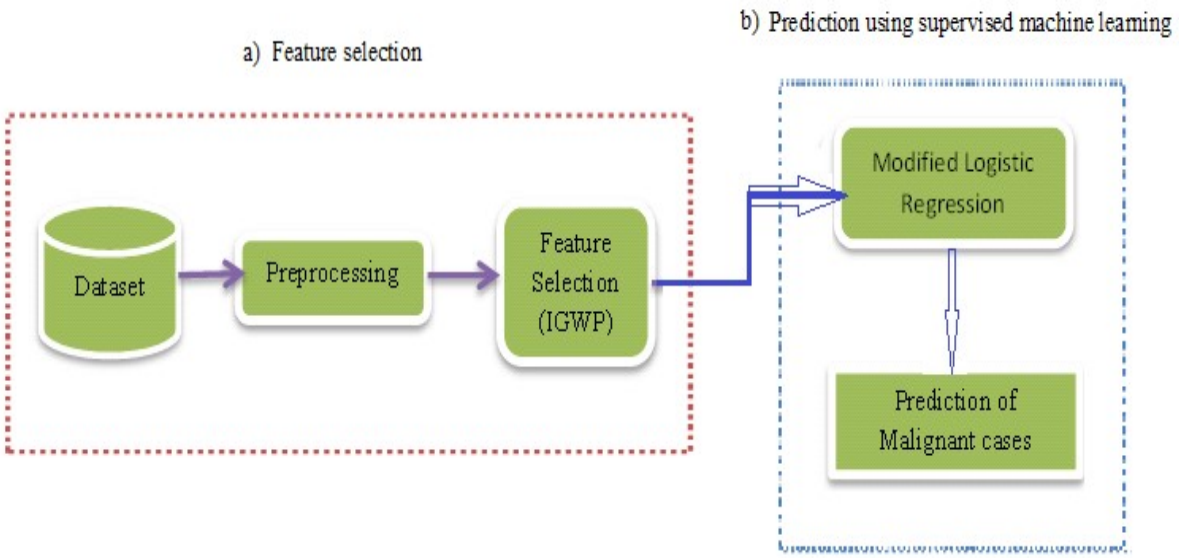

Figure 1: The flowchart of the proposed method; our prediction task is composed of two components: a) Feature selection b) Supervised machine learning algorithm

```
NC.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	radius_worst
0	842302.0	1	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	25.38
1	842517.0	1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	24.99
2	84300903.0	1	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	23.57
3	84348301.0	1	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	14.91
4	84358402.0	1	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	22.54

5 rows × 32 columns

texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890
23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902
25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758
26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300
16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678

Figure 2 : Representation of the first 5 rows of the data set.

```
NC.tail()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	radius_worst
947	885430.0	1	19.73	19.82	130.70	1206.0	0.1062	0.1849	0.24170	0.09740	...	25.28
948	8860703.0	1	17.30	17.08	113.00	928.2	0.1008	0.1041	0.12660	0.08353	...	19.85
949	886227.0	1	19.45	19.33	126.50	1169.0	0.1035	0.1188	0.13790	0.08591	...	25.70
950	886453.0	1	13.96	17.05	91.43	602.4	0.1096	0.1279	0.09789	0.05246	...	16.39
951	88649002.0	1	19.55	28.77	133.60	1207.0	0.0926	0.2063	0.17840	0.11440	...	25.05

5 rows × 32 columns

texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
25.59	159.8	1933.0	0.1710	0.5955	0.8489	0.2507	0.2749	0.12970
25.09	130.9	1222.0	0.1416	0.2405	0.3378	0.1857	0.3138	0.08113
24.57	163.1	1972.0	0.1497	0.3161	0.4317	0.1999	0.3379	0.08950
22.07	108.1	826.0	0.1512	0.3262	0.3209	0.1374	0.3068	0.07957
36.27	178.6	1926.0	0.1281	0.5329	0.4251	0.1941	0.2818	0.10050

Figure 3: Representation of the last five rows of the dataset

```
NC.describe()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_worst
count	9.520000e+02	952.000000	952.000000	952.000000	952.000000	952.000000	952.000000	952.000000	952.000000	952.000000	...	952.000000
mean	3.152921e+07	0.372899	14.109875	19.397332	91.865893	653.467857	0.096675	0.104819	0.089374	0.048991	...	25.795074
std	1.298229e+08	0.483830	3.535587	4.329331	24.386027	354.021714	0.013907	0.051777	0.079437	0.038500	...	6.137810
min	8.670000e+03	0.000000	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	...	12.020000
25%	8.667148e+05	0.000000	11.680000	16.330000	75.135000	418.325000	0.086750	0.066230	0.029988	0.020688	...	21.327500
50%	9.080545e+05	0.000000	13.390000	18.940000	86.600000	552.050000	0.096035	0.095090	0.061950	0.033870	...	25.480000
75%	8.810572e+06	1.000000	15.750000	21.872500	103.700000	768.325000	0.105400	0.130500	0.130700	0.073400	...	30.060000
max	9.113205e+08	1.000000	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	...	49.540000

8 rows × 33 columns

perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	label
952.000000	952.000000	952.000000	952.000000	952.000000	952.000000	952.000000	952.000000	952.000000
107.112521	877.730987	0.132956	0.255527	0.275159	0.114947	0.289686	0.084099	0.372899
33.442874	564.920887	0.022488	0.154001	0.208028	0.065035	0.061637	0.017642	0.483830
50.410000	185.200000	0.071170	0.027290	0.000000	0.000000	0.156500	0.055040	0.000000
84.080000	514.000000	0.117875	0.147800	0.118100	0.065280	0.250150	0.071980	0.000000
97.980000	689.100000	0.131600	0.217000	0.230600	0.101500	0.282450	0.080200	0.000000
125.100000	1055.000000	0.146400	0.341700	0.385300	0.161000	0.317675	0.092110	1.000000
251.200000	4254.000000	0.222600	1.058000	1.252000	0.291000	0.663800	0.207500	1.000000

Figure 5: Computation of values Min, Max, Mean, Standard deviation, and quartile values

```
NC.groupby('label').mean()
```

label	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean
0	2.670658e+07	0.0	12.136233	18.094238	78.034640	462.171524	0.093089	0.081227	0.047295	0.026136
1	3.963939e+07	1.0	17.428930	21.588732	115.125775	975.169014	0.102707	0.144493	0.160137	0.087426

2 rows × 32 columns

radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_w
13.381834	23.744941	87.02526	559.249749	0.126048	0.185363	0.170964	0.075417	0.269912	0.07
21.069127	29.242761	140.89307	1413.317746	0.144573	0.373522	0.450383	0.181424	0.322940	0.09

Figure 7: Computation of mean values for all the features

radius_ mean	texture_ mean	perimeter _mean	area_ mean	smoot hness_ mean	compactn ess_mean	concavity _mean	concave points_ mean	symmetry _mean	fractal_di mension_ mean	radius_ se	texture _se	perimeter_ se	area_se	smoothness_ se
19.07	24.81	128.3	1104	0.091	0.219	0.2107	0.09961	0.231	0.06343	0.9811	1.666	8.83	104.9	0.006548

compactness _se	concavity _se	concave points_se	symmetry _se	fractal_di mension_ se	radius_w orst	texture_ worst	perimeter _worst	area_wor st	smoothne ss_worst	compactn ess_worst	concavity _worst	concave points_ worst	symmetry _worst	fractal_di mension_ worst
0.1006	0.09723	0.02638	0.05333	0.007646	24.09	33.17	177.4	1651	0.1247	0.7444	0.7242	0.2493	0.467	0.1038

Figure 8: An example prediction of a proposed model for a certain set of features

Journal of Theoretical and Applied Information Technology
30th April 2023. Vol.101. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

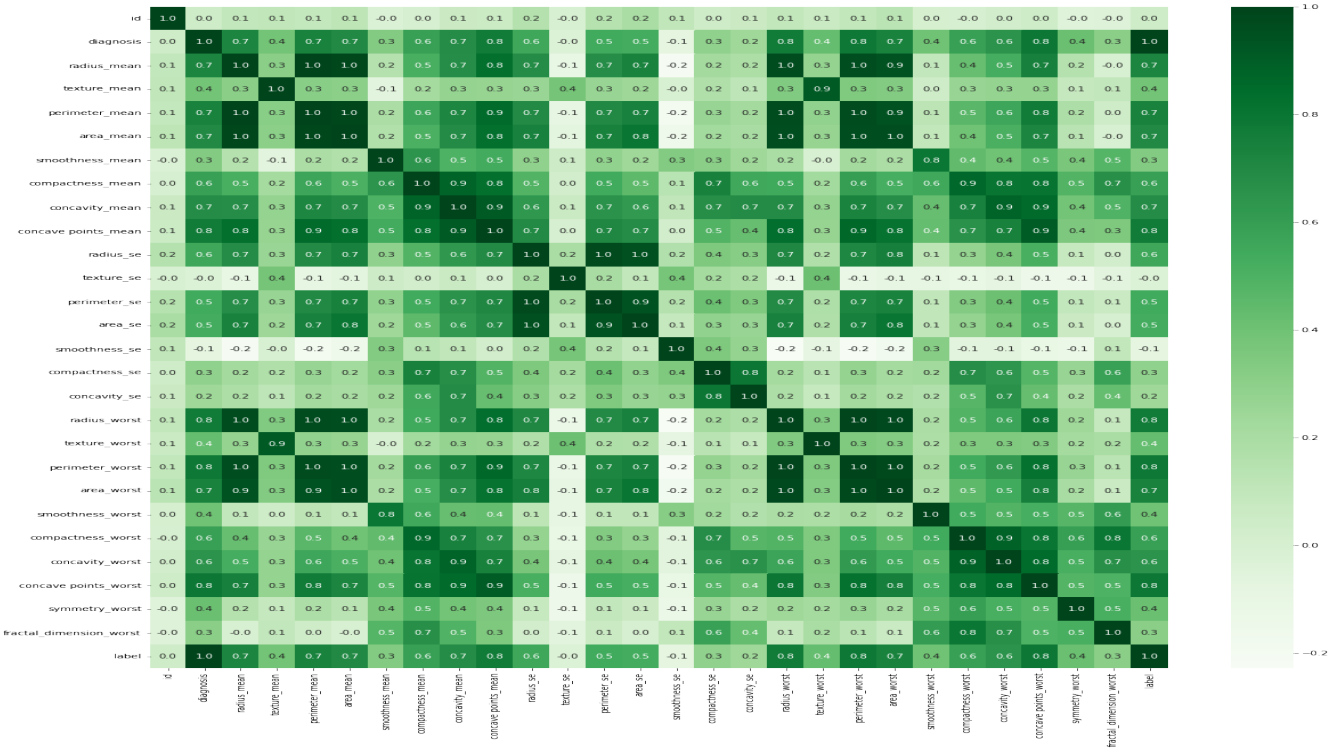

Figure 9: The correlative matrix of the set of features

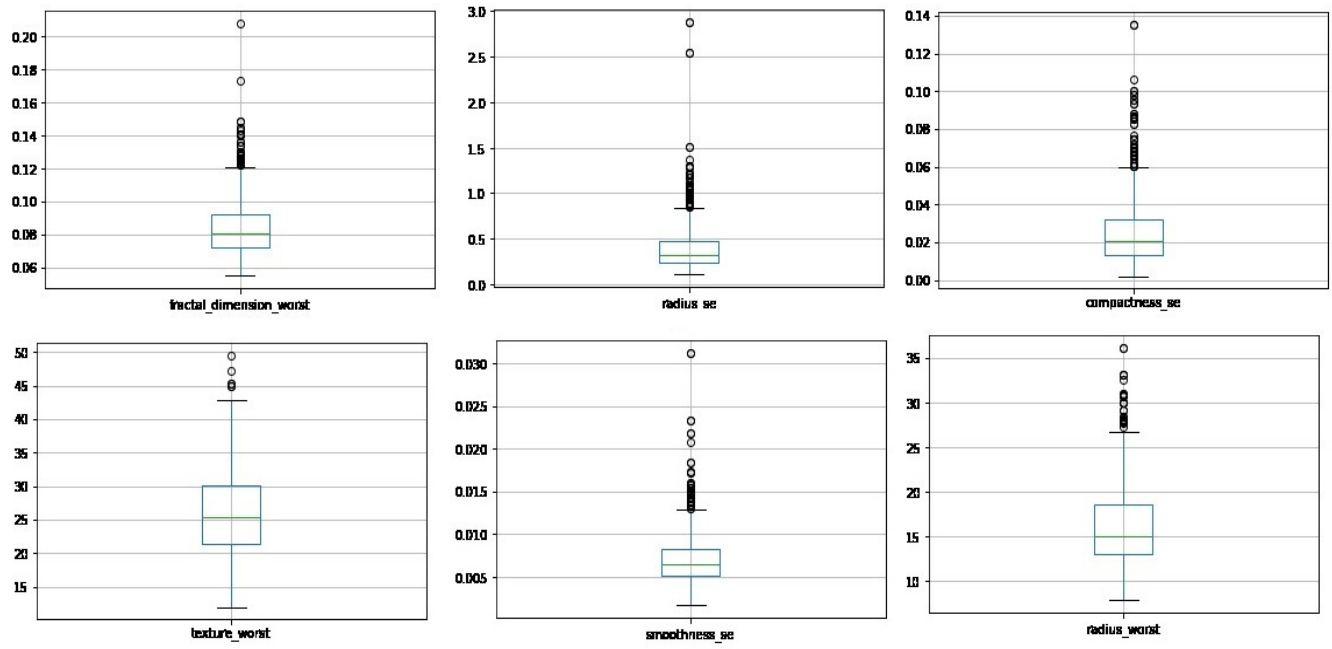

Figure 11: Visualization of features by using Box and whisker plots