International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 3, June 2021, pp. 2631~2639
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i3.pp2631-2639

Expert cancer model using supervised algorithms with a LASSO selection approach

Pronab Ghosh[1], Asif Karim[2], Syeda Tanjila Atik[3], Saima Afrin[4], Mohd. Saifuzzaman[5]
[1,3,4,5]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
[2]College of Engineering, IT and Environment, Charles Darwin University, Casuarina, NT, Australia

Article Info

Article history:

Received Jul 5, 2020
Revised Sep 11, 2020
Accepted Oct 13, 2020

Keywords:

Breast cancer
Decision tree
SVM

ABSTRACT

One of the most critical issues of the mortality rate in the medical field in current times is breast cancer. Nowadays, a large number of men and women are facing cancer-related deaths due to the lack of early diagnosis systems and proper treatment per year. To tackle the issue, various data mining approaches have been analyzed to build an effective model that helps to identify the different stages of deadly cancers. The study successfully proposes an early cancer disease model based on five different supervised algorithms such as logistic regression (henceforth LR), decision tree (henceforth DT), random forest (henceforth RF), Support vector machine (henceforth SVM), and K-nearest neighbor (henceforth KNN). After an appropriate preprocessing of the dataset, least absolute shrinkage and selection operator (LASSO) was used for feature selection (FS) using a 10-fold cross-validation (CV) approach. Employing LASSO with 10-fold cross-validation has been a novel steps introduced in this research. Afterwards, different performance evaluation metrics were measured to show accurate predictions based on the proposed algorithms. The result indicated top accuracy was received from RF classifier, approximately 99.41% with the integration of LASSO. Finally, a comprehensive comparison was carried out on Wisconsin breast cancer (diagnostic) dataset (WBCD) together with some current works containing all features.

Corresponding Author:

Pronab Ghosh
Department of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
Email: pronab1712@gmail.com

1. INTRODUCTION

The second-highest deadly disease over the world and a significant reason for women's deaths in this contemporary time is breast cancer. It creates a significant challenge to women's health in the world today. As per the statistics of the world health organization (WHO), 2.1 million women are getting affected due to breast cancer annually. The rate of death is approximately 15% among all women [1], a reported number of 627,000 females died due to cancer-related issues in 2018. It has been predicted that 127.5 are diagnosed, whereas 20.6 females have died per 100,000 each year [2]. A survey was taken from Globocan which ensures that 87,090 females have died in the year of 2018 [3] but 58% of deaths were noticed in developed countries as per the report of 2008 [4]. Considering the number of death records, India has achieved the first rank whereas Thailand has the fourth most, with 5,902 deaths in the same year [3].

The key concept behind the proposed research is to develop a framework for breast cancer diagnosis that is completely based on machine learning. The study aims to address different algorithms such as LR, DT,

RF, SVM, and KNN for identification the affected people of breast cancer. The most reliable selection technique such as LASSO is also used to determine the most relevant and strongly associated attributes that show considerable influence on the predicted feature. This analysis is carried out using 10-fold CV for making it more reliable. Different performance assessment metrics such as confusion matrix, accuracy (Acc), precision (P score), recall (R score), specificity (Spe), negative predictive value (NPV), false discovery rate (FDR), false negative rate (FNR), and false positive rate (FPR) have been used to properly evaluate classifier performance. Besides, the procedure of data preprocessing is applied to the dataset of breast cancer. The key objectives of the present research are:

- All features have been prepossessed with the help of standard scaler technique to keep the values in the range of [0, 1].
- The evaluation processes of various models have been experimented with the separation of 80: 20 by LASSO with 10-fold cross-validation.
- The study carried out a comprehensive comparison of the performance of LASSO selected features and current existing works to identify the affected cancer patients, which highlights the performance of the proposed intelligent system. We have used the default settings available at scikit-learn for LASSO, and have not optimized any specific parameter for performance tuning. The default setting provided good enough results.

Several machine learning approaches have been evaluated to predict an accurate outcome on the dataset of breast cancer. Some of them are explained to show the researchers' findings. Latchoumi *et al.* [5] explained a weighted particle swarm optimization technique using a smooth support vector method to give this research novelty. Earning a low error rate was the main target of this research technique. Besides, it successfully generates 98.42% accuracy using this algorithm. Distinctive machine learning calculations have been contemplated and used to anticipate the early detection of breast cancer in the investigation of J. Rohit *et al.* [6] they also worked on breast cancer dataset by using different predictive models and identified an optimal solution considering various stages of cancer. The models were assessed separately on the basis of their deployment strategy. Alicovic *et al.* used a genetic algorithm [7] based on FS and multiple classifications to make their research more specific. That research helps to find out the Individual accuracy and diversity with very sensitive accuracy. The observation has been deployed for identifying no class or different classes. A multi-layered algorithm has been generated by K. Arutchelvan *et al.* [8] through the combination of DT techniques. This algorithm helps to diagnose the prediction of cancer risk and other critical diseases. Under the supervision of data mining approaches, it easily detects whether a patient has cancer or not.

The researchers collected the clinical dataset to evaluate the grammatical problem using machine learning tools. Kumar *et al.* showed [9] an idea on the breast cancer dataset which helped to eradicate the early death risk because of their comprehensive research. Different data mining tools were developed to make a prediction system for Breast cancer by A. F. M. Agarap [10]. Six approaches were performed on this system to achieve specificity and recall results. The reported accuracy exceeds 90% in their proposed system. In the research work of Nauck *et al.* [11] 95.57% of accuracy has been shown with the fuzzy clustering technique. P. Gupta *et al.* explained a technique on the cancer dataset of UCI Irvine machine learning repository combining three algorithms [12] (CART, RF, and KNN) to show the predicted performance. KNN model, however, provides 97% accuracy among all of them. Y. Khourdifi *et al.* [13] analyzed the early prediction rate of breast cancer depending on various classifiers, such as NB, RF, SVM, and KNN. In their study, the best result was obtained from SVM around 97.9%. C. Shravya *et al.* [14] suggested a diagnosis approach on the basis of cancer dataset combining three classifiers such as LR, SVM, and KNN to evaluate the performance. They also illustrated different performance indices in terms of accuracy, precision, and sensitivity. The highest predictable accuracy of 92.07% was generated by SVM model. N. M. Ali *et al.* [15] suggested a model that was implemented by SVM and LG algorithms based on Boruta and LASSO FS techniques. In their experiment, the best performance was noticed on LASSO by LG (98.61%) algorithm. Three strategies for the identification of cancer stages were tested over the breast cancer dataset by V. Chaurasia *et al.* with the completion of the pre-processing step, an average result was achieved among all of them through a simple logistic method [16], but it was about to 74.47%.

In the study of [12-14], they evaluated the overall results with some classifiers without using any FS approach. Among all of them, the best result was obtained from SVM in [13] which was 97.9%. Another study [17] illustrated two FS techniques including LASSO and Boruta and got 98.61% by LR. However, our system has obtained 99.41% by RF using the LASSO FS approach which outperforms other previous studies.

2. OUR PROPOSED ALGORITHMS

The learning technique of a machine is a solution that supports a distinct estimation. It divides the data into training to predict the best parameter of the model and gathered outcomes are applied to the test data. The learning algorithm [18] keeps on upgrading itself optimum prediction and interpretation of new data.

2.1. Logistic regression

Logistic Regression, which is also called a statistical approach, helps to classify a classified variable [19]. By estimating the probability of a particular class, LR generates a model that differentiates between samples. Having two different outcomes, 1 represents true and 0 represents false. Figure 1 shows the working approach [20].

Input : Dataset, D = {(x1, y1), (x2,y2),……..(xp,yp)}
Base learner L
Number of base learners, B
Process:
For b = 1 to B:
$h_b = L\ (D_b)$ // Train a base learner h_b
end
Output : $H(x) = \left(\sum_{b=1}^{B} h_b(x)\right) / B$

Figure 1. An algorithm of logistic regression

Three different entities have been selected as input such as data set, base learner, and the number of base learners. After the input, base learner is selected for training and it continues till it reaches the upper limit. As a result, the expected output is gained through the H(x) variable where base leaner is divided by number of base learners (B).

2.2. Decision tree

Decision Tree uses hierarchical tree approaches in where each node illustrates a feature, a branch illustrates the rule of decision, and leaves show an outcome [21]. It can be applied for classification and regression trees problem. The dataset has been divided into two subgroups to show the illustration process of CART. An overall idea is illustrated in the equations [22].

$$\text{Gain of information (IG), I (N, P)} = -\left(\frac{N}{N+P}\log 2\ \frac{N}{N+P} + \frac{p}{N+P}\log 2\ \frac{p}{N+P}\right) \quad (1)$$

$$\text{Value of Entropy, E (A)} = \sum_{i=1}^{n}\left(\frac{Pi+Ni}{N+P}\right) * \text{Gain of information} \quad (2)$$

There are two different probabilities (P and N), that have been successfully utilized to produce information gain (IG). The summations of probabilistic outcomes are calculated to get the IG value in (1). The value of entropy (E (A)) is calculated in (2) based on IG value. IG value is multiplied by the probabilistic outcomes and it is continued until to get n value.

2.3. Random forest

Random forest has been deployed as an ensemble model that is a common tool for classification and regression. To improve accuracy along with overfitting problems, it combines multiple DT [23] in a single unit. Each tree is normally built to achieve a sample of new training data with an averaged value of final prediction.

2.4. Support vector machine

Support Vector Machine is used as a training method to study classification and regression rules from a large number of data. An optimal hyperplane [24] is being produced to categorize test data by SVM provided a set of labeled training is available. SVM is extensively used to identify the stage of cancers in histopathology images. At n-levels, a hyperplane is described in (3) [25].

$$B0 + \beta 1X1 + \beta 2\ X2 + \ldots + \beta nXn = 0 \quad (3)$$

where $\beta 0, \beta 1, \beta 2$ ... βn represents hypothetical values and Xn shows data points in the sample space of n dimension. The initial intention for developing SVM was to solve 2-class classification problem, later on, it was tuned for multi-class problems. The algorithm takes a 1-vs-rest approach where it attempts to separate a single class from all other classes. At the time of testing, the class label of z of a class pattern y is determined as:

$$z = \begin{cases} n, & if\ d_n(y) + t_l > 0 \\ 0, & if\ d_n(y) + t_l < 0 \end{cases} \quad (4)$$

where dn (y) $= max\ \{\mathrm{dn}\}_{i=1}^{Nl}$, di (y) is the distance from y to the SVM hyperplane corresponding to class i, and t_l is the classification threshold.

2.5. K-Nearest neighbor

KNN is the most commonly used algorithms in machine learning because of its flexibility. Besides, the learning stage is not necessary like other algorithms [26]. KNN is called a classified and a lazy algorithm in data mining. Euclidean distance is shown in the equation is: [27] Two data points are taken for Euclidean distance.

$$D\ (x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (5)$$

The coordinate of y point is subtracted from the coordinator of x point under the root square value. The obtainable result is called Euclidean distance like D(x, y). The following data set is categorized using training data.

3. RESEARCH METHODOLOGY

Research methodology helps to obtain a logical knowledge of the research work. Research subject and instrumentation have been explained to aid in the establishment of a clear concept. Since data is the most significant part of machine learning work, a conceptual description has been added to the data collection.

3.1. Dataset collection

In our proposed system, Wisconsin breast cancer (diagnostic) dataset (WBDC) [28] gathered from UCI machine learning repository has been evaluated to predict the accuracy rate. Most of the attributes were in numeric values, except diagnosis, which has been in categorical value. As a result, we have converted this categorical value into numeric value for making a prediction. In 569 instances with 32 attributes, 357 are benign class (B), and 212 are malignant (M) class. Two classes of cancer disease are shown in Table 1.

Table 1. Attribute, and type of values

Serial Numbers	Attributes of Breast Cancer	Serial Numbers	Attributes of Breast Cancer	Serial Numbers	Attributes of Breast Cancer
1	ID number	12	fractal_dimension_mean	23	radius_worst
2	Diagnosis	13	radius_se	24	texture_worst
3	radius_mean	14	texture_se	25	perimeter_worst
4	texture_mean	15	perimeter_se	26	area_worst
5	perimeter_mean	16	area_se	27	smoothness_worst
6	area_mean	17	smoothness_se	28	compactness_worst
7	smoothnes_mean	18	compactness_se	29	concavity_worst
8	compactness_mean	19	concavity_se	30	concave points_worst
9	concavity_mean	20	concave points_se	31	symmetry_worst
10	concave_points_mean	21	symmetry_se	32	fractal_dimension_worst
11	symmetry_mean	22	fractal dimension_se		

3.2. Data preprocessing

The applied dataset is picked from the UCI machine learning repository to detect the stages of cancer disease and standard scaler [29] approach has been addressed to keep them in the range of [0, 1]. Afterward, we have to convert one categorical feature such as 'Diagnosis' into numbers by using label encoding [30] technique. For example, we label the 'Benign' and 'Malignant' as 0 and 1 respectively.

3.3. An expected outcome on a selected feature selection algorithm

Feature selection [31] is a technique that helps to select the appropriate features for getting the highest outcomes based on the gathered dataset. Before performing a data experiment, the selection process of the function must examine the dataset. The selection of features in this framework is only used for improving model efficiency, and also helps to reduce execution time. We used one of the embedded methods such as the LASSO strategy. Using a randomly generated subset of keywords from the corresponding sub-region, the efficiency of this function can be improved by repeating the above procedure. It is called the randomized LASSO function that was introduced Wang [32]. In addition, LASSO is considered the most significant feature contained in qi which symbolizes the vector of the similar sub-region keys in Figure 2.

3.3.1. Least absolute shrinkage and selection operator feature selection algorithm (LASSO)

The LASSO functionality of the operator is dependent upon updating the absolute value of the function coefficient. Various coefficient ranks of the characteristics are zero, and those characteristics with negative coefficients are removed from the subset of features. The LASSO performs well with low coefficiency function values. A subset of desired functions including irrelevant features may be selected in the LASSO approach [33]. The LASSO selects closely related characteristics that are to be viewed as true, and the rest as false. After completing the evaluation process of LASSO, We get four essential features that have been clearly shown in Figure 3. Overall, texture_worst contains the highest score (0.01748).

Input: Data = $\{X_i\ j\ , Y_i\ j\}$ = 1, 2, . . ., N_i; Sampling ratio ε ∈ (0,1);
1. number of randomizations T ∈ M; threshold H ∈ M
2. Output : relevant features $\bar{q}$ i of the grid
3. for k = 1, 2, . . . , T:
4. Data = sampling with replacement from Data with ratio ε
5. q_i = LASSO-based fingerprint selection using Data
6. end for
7. frequency of selection of each feature is calculated according to q_i, k = 1, 2, . . . , T
8. Return $\bar{q}$ i: the set of features selected most frequently

Figure 2. The working procedure of LASSO algorithm

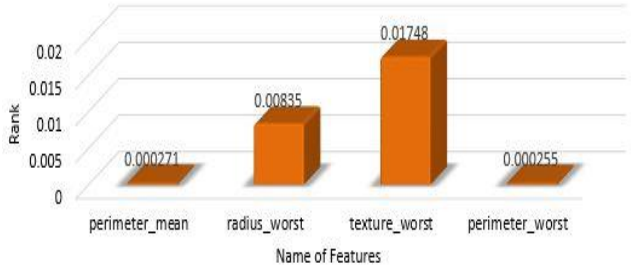

Figure 3. The graphical view of selective features by LASSO algorithm

3.4. Graphical representation of proposed model

The cancer dataset has been collected through an online repository to detect the diagnose rate of cancer. Since the collected dataset of cancer has no missing value, it is directly transmitted to the preprocessing technique. In this technique, 10-fold CV approach is taken for an experiment to deal with overfitting and underfitting issues. After successfully applying five algorithms to the given dataset, the most suitable outcome is received from RF-based on validation dataset among all algorithms. The overall description process using pseudocode has been shown as a graphical format in Figure 4.

```
Step 1: Function Pre-processing ()
Step 2:     import dataset
Step 3:     Convert dataset to csv format
Step 4:     Label encoder for categorical values
Step 5: End Pre-processing ()
Step 6: Function SpecificFeaturesSelection ()
Step 7:                 Go to Step 1
Step 8:                    Choose selection model
Step 9:                  Apply LASSO selection algorithm on preprocessed dataset
```

```
Step 10: End SpecificFeaturesSelection ()
Step 11: Function TrainTestSplit ()
Step 12:                   Go to Step 7
Step 13:                 Choose 10-fold cross-validation method
Step 14: End TrainTestSplit ()
Step 15: Function BestClassifiersSelectionApproach ()
Step 16:               Choose 5 algorithms (LR, DT, RF, SVM, and KNN)
Step 17:                 For i = 0: 5
Step 18:                     Go to Step 12
Step 19:                   Predict class of data
Step 20:               Evaluate result
Step 21:                 Comparison among overall outcomes
Step 22:                 Recommend the best model over the dataset by LASSO technique
Step 23: End BestCLassifiersSelectionApproach ()
```

3.5. Validation technique of classifiers

In k-fold cross-validation, the collected data was broken down into k equal parts where k-1 categories have been chosen to train the models and other parts are evaluated in each phase to test performance. The process of validation is iterated for k-times. The classifier's efficiency is calculated by the results of k. Various values of k are chosen for CV. We've only used k = 10 in our experiment because their output is good [34]. In the 10-fold Cross-validation system, data is 80% allocated for training and the remaining for experimental purposes. For each fold, the process was carried over for 10 times; every data points in the testing and training sets were randomly distributed over the entire dataset before the selection training and testing of new sets for each iteration. After the completion of 10-fold period, averages of all performance metrics are calculated.

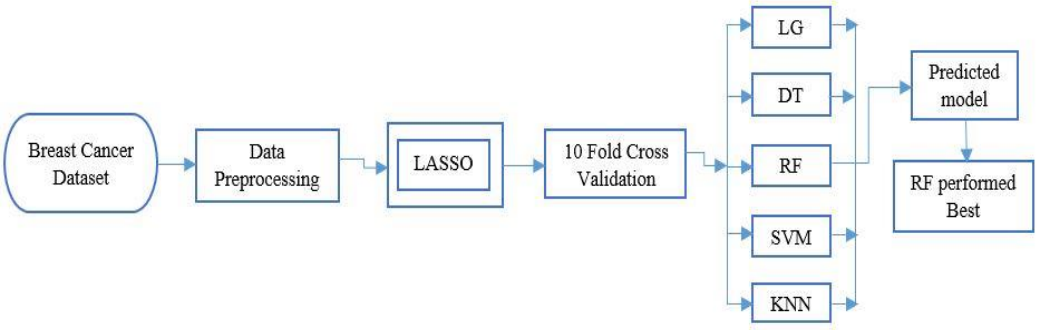

Figure 4. Skeleton of the proposed cancer detection model

3.6. Performance measure indices

The performance and correctness have been surveyed due to do measuring some performance indices to reduce the death risk [35] using machine learning methods. This formula has been applied to find Acc, P_score, R_score, and Spe. [36].

4. EXPERIMENTAL RESULTS AND DISCUSSION

Five algorithms are used to predict breast cancer outcomes on selected features (perimeter_mean radius_worst, texture_worst, perimeter_worst) by the LASSO FS algorithm that helps to predict the best outcome over the extracted columns. To evaluate various methods, some performance metrics such as Confusion Matrix, Acc, FPR, FNR, Spe, and P_Score. have been used to evaluate various models of algorithms. Among them, RF has provided the highest accuracy for both training and testing data.

4.1. Experimental consequence among several algorithms

LR which is called statistical machine learning technique, on the other hand, SVM can be used in both classification and regression. We have obtained from both 98.24% acc, 97% R_score, 98% score for both Spe and P_score, 0% is achieved by FPR and FDR. DT works like a binary tree where every data point denotes an attribute, and KNN is considered a non-parametric algorithm along with classification and regression usage. We have achieved a similar result, where acc is found at 96.75%. Other performance indices such as NPV, R_score, P_score, and FNR are shown 97.41%, 98%, 98%, and 2% respectively for disease

prediction. RF has got the highest Acc (97.41%) with the lowest error rate [17] comparatively than others. This is the highest predictable score regarding breast cancer dataset. A short explanation is added in Table 2.

Table 2. Machine learning classifiers performance on 10-fold cross-validation technique

Dimension	Logistic Regression	Decision Tree	Random Forest	Support Vector Machine	K-Nearest Neighbor
Confusion matrix	[61.6% 0%] [3% 35.4%]	[61.7% 0%] [2% 36.3%]	[61.7% 0%] [1% 37.3%]	[61.6% 0%] [3% 35.4%]	[61.7% 0%] [2% 36.3%]
Acc	98.24%	98.88%	99.41%	98.24%	98.88%
P_score	98%	98%	99%	98%	98%
R_score	97%	98%	99%	97%	98%
Spe	98%	99%	99%	98%	99%
NPV	98%	97.41%	98.41%	98%	97.41%
FDR	0%	0%	0%	0%	0%
FPR	0%	0%	0%	0%	0%
FNR	2.9%	2%	0.9%	2.9%	2%

4.2. Comparing with existing systems accuracy

Table 3 provides a comparative view in terms of Accuracy achieved by various algorithms with existing systems. We can easily see from the table that the prediction rate of our proposed system is very high rather than previous works explained referring to [12-15]. In the table, both LR and SVM show a similar accuracy of 98.24% from our system. LR provides 92.10% for [14], 73.61% and 98.61% for [35] that is noticed from Bortua and LASSO FS techniques, whereas SVM is found at about 97.9% [13], 92.78% [14], 69.44% and 59.72% for [35] that is taken from Bortua and LASSO FS techniques respectively. Concerning RF, our 10-fold CV technique has achieved a better prediction rate of 99.41%, which is the highest output of our model, compared to recent works of [12] (96.47%). Regarding DT and KNN, our system outperforms than all given existing techniques. The predicted accuracy is received 98.88%, on the contrary, 92.35% is found for DT in [12] and the outcomes of KNN is obtained 97% [12], 96.1% [13] and 92.23% [14]. Finally, we can easily say that individuals with these features contain high risks of being affected by breast cancer which has been briefly described in the overall context.

Table 3. Comparison of the proposed system with existing systems

Models	Accuracy of 10-fold cross validation method	Existing Systems [12]	Existing Systems [13]	Existing Systems [14]	Existing Feature Selection techniques [36]
LR	98.24%	-	-	92.10%	Boruta (73.61%), LASSO (98.61%)
DT	98.88%	92.35 %	-	-	-
RF	99.41%	96.47 %	96%	-	-
SVM	98.24%	-	97.9 %	92.78%	Boruta (69.44%), LASSO (59.72%)
KNN	98.88%	97 %	96.1 %	92.23%	-

5. CONCLUSION AND FUTURE WORK

We have studied the use of different ML tools to predict the early diagnosis rate. All of these techniques through the LASSO feature selection have been evaluated for getting a more optimistic result. After the effective feature selection steps, a rather promising outcome has been obtained from RF algorithm with 99.41% accuracy in comparison to other all techniques. To address probable overfitting issues, we are already trying to collect a large number of datasets for calculating the performance even more precisely. After all, our proposed technique has been succeeded in generating more secure and efficient results with very low error rates. In future, we will develop an online android app to show the relevant symptoms of breast cancer at the earliest as a tool for early detection of such type of cancer.

ACKNOWLEDGEMENTS

The authors are grateful to Daffodil International University, Dhaka, Bangladesh for all the support.

REFERENCES

[1] "World Health Organization (WHO)," 2020. [Online]. Available: https://www.who.int/cancer/prevention/diagnosisscreening/breastcancer/en/.
[2] "Women Breast Cancer," 2020. [Online]. Available: https://seer.cancer.gov/statfacts/html/breast.html.

[3] World Health Organization International Agency for Research on Cancer, "Global Cancer Observatory," 2020. [Online]. Available: http://gco.iarc.fr/.
[4] "Breast Cancer Dataset Survey by WHO," [Online]. Available: who.int/cancer/detection/breastcancer /en/index1.html
[5] L. Latchoumi, et al., "Abnormality Detection Using Weighed Particle Swarm Optimization and Smooth Support Vector Machine," *Biomedical Research*, vol. 28, no. 11, pp. 4749-4751, 2017.
[6] J. Rohit, et al., "Stage-specific Predictive Models for Breast Cancer Survivability," *International Journal of Medical Informatics*, vol. 97, pp. 304-311, 2017.
[7] E. Aličković and A. Subasi, "Breast Cancer Diagnosis Using GA Feature Selection and Rotation Forest," *Neural Computing and Applications*, vol. 28, pp. 753-763, 2017.
[8] K. Arutchelvan and R. Periyasamy, "Cancer Prediction System Using Datamining Techniques," *International Research Journal of Engineering and Technology (IRJET)*, vol. 2, no. 8, pp. 1179-1183, 2015.
[9] M. Kumar, S. S. Tomar and B. Gaur, "Mining Based Optimization for Breast Cancer Analysis: A Review," *International Journal of Computer Applications*, vol. 119, no. 13, pp. 1-6, 2015.
[10] A. F. M. Agarap, "On Breast Cancer Detection: an Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset," *ICMLSC '18: Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, pp. 5-9, 2018.
[11] Nauck, et al., "Obtaining Interpretable Fuzzy Classification Rules from Medical Data," *Artificial Intelligence in Medicine*, vol. 16, no. 2, pp.149-169, 2015.
[12] P. Gupta and L. Shalini, "Analysis of Machine Learning Techniques for Breast Cancer Prediction," *International Journal of Engineering and Computer Science*, vol. 7, no. 5, pp. 23891-2389, 2018.
[13] Y. Khourdifi, et al., "Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification," *2018 International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS)*, Kenitra, 2018, pp. 1-5.
[14] C. Shravya, K. Pravalika, and S. Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 6, pp. 1106-1110, 2019.
[15] N. M. Ali, N. A. A. Aziz and R. Besar, "Comparison of Microarray Breast Cancer Classification Using Support Vector Machine and Logistic Regression with LASSO and BORUTA Feature Selection," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 20, no. 2, pp. 712-719, 2020.
[16] V. Chaurasia, et al., "Data Mining Techniques: to Predict and Resolve Breast Cancer Survivability," *International Journal of Computer Science and Mobile Computing IJCSMC*, vol. 3, no. 1, pp. 10-22, 2017.
[17] S.Yuanhang, et al., "On Extended Long Short-term Memory and Dependent Bidirectional Recurrent Neural Network," *Neurocomputing*, vol. 356, pp. 151-161, 2019.
[18] F. M. J. M. Shamrat, M. Asaduzzaman, P. Ghosh, D. Sultan and Z. Tasnim, "A Web Based Application for Agriculture: "Smart Farming System," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 6, pp. 2309-2320, 2020.
[19] S. Bharati, M. A. Rahman and P. Podder, "Breast Cancer Prediction Applying Different Classification Algorithm with Comparative Analysis using WEKA," *2018 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), Dhaka, Bangladesh*, 2018, pp. 581-584.
[20] F. Muratore, et al., "Correlations between Histopathological Findings and Clinical Manifestations in Biopsy-proven Giant Cell Arteritis," *Journal of Autoimmunity*, vol. 69, pp. 94-101, 2016.
[21] P. Ghosh, M.Z. Hasan, O.A. Dhore, A.A. Mohammad and M. I. Jabiullah, "On the Application of Machine Learning to Predicting Cancer Outcome," *Proceedings of the International Conference on Electronics and ICT – 2018, organized by Bangladesh Electronics Society (BES)*, 2018.
[22] F. M. J. M. Shamrat, Z. Tasnim, P. Ghosh, A. Majumder and Md. Z. Hasan, "Personalization of Job Circular Advertisement to Candidates Using Decision Tree Classification Algorithm," *2020 IEEE International Conference for Innovation in Technology*, 2020.
[23] S. Raghavendra, et al., "Performance Evaluation of Random Forest with Feature Selection Methods in Prediction of Diabetes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 353-359, 2020.
[24] M. Srivenkatesh, "Prediction of Breast Cancer Disease Using Machine Learning Algorithms," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9, no. 4, pp. 2868-2878, 2020.
[25] M. M. Islam, H. Iqbal, M. R. Haque and M. K. Hasan, "Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors," *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dhaka, 2017, pp. 226-229.
[26] S. A. Medjahed, A. Benyettou and T. A. Saadi, "Breast Cancer Diagnosis by Using K-Nearest Neighbor with Different Distances and Classification Rules," *International Journal of Computer Applications*, vol. 62, no. 1, pp. 1-5, 2013.
[27] P. Ghosh, M. Z. Hasan and M. I. Jabiullah, "A Comparative Study of Machine Learning Approaches on Dataset to Predicting Cancer Outcome," *Bangladesh Electronic Society*, vol. 18, pp. 01-05, pp. 81-86, 2018.
[28] "Breast Cancer Wisconsin (Diagnostic) Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+ Cancer+Wisconsin+ (Diagnostic).
[29] "Standard Scaler Technique," [online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing. StandardScaler.html.

[30] "Encoding Categorical Data to Prepare Dataset," 2020. [Online]. Available: https://towardsdatascience.com/smarter-ways-to-encode-categoricaldata-for-machine-learning-part-1-of-3-6dca2f71b159.
[31] Y. Su, et al., "Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis," *2018 24th International Conference on Pattern Recognition (ICPR)*, Beijing, 2018, pp. 585-590.
[32] C. Zhou, and A. Wieser, "Jaccard Analysis and LASSO-Based Feature Selection for Location Fingerprinting with Limited Computational Complexity," *LBS 2018: 14th International Conference on Location Based Services*, pp. 71-87, 2018.
[33] H. Dhahri, et al., "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms," *Journal of healthcare engineering*, vol. 2019, 2019.
[34] A. U. Haq, et al., "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems*, vol. 2018, 2018.
[35] G. Saranya, et al., "A Comprehensive Study on Disease Risk Predictions in Machine Learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4217-4225, 2020.
[36] F. M. J. M. Shamrat, P. Ghosh, M. H. Sadek, A. Kazi, and S. Shultana, "Implementation of Machine Learning Algorithms to Detect the Prognosis Rate of Kidney Disease," *2020 IEEE International Conference for Innovation in Technology*, 2020.

BIOGRAPHIES OF AUTHORS

Pronab Ghosh acquired his Bachelor of Science degree in Computer Science and Engineering (CSE) Department from Daffodil International University in the year of 2019. He has been involved extensively in collaborative research activities, particularly in the fields of machine learning, cloud computing, and IoT. In addition, he has written and published different articles in conferences and international journals.

Asif Karim is a Ph.D. researcher at Charles Darwin University and lives in Darwin's port city. His research area includes Cryptographic Communication and Artificial Intelligence. He is presently aiming to build a robust and sophisticated email filtering framework using Machine Learning models. Asif has a broad background in the IT industry, especially in software engineering.

Syeda Tanjila Atik has completed her both Bachelor and Masters in IT, both from the IIT, Jahangirnagar University in 2016 and 2018, respectively. She has been recently serving as a lecturer at Daffodil International University since April 2017. She works on edge computing, data mining, IoT, and machine learning.

Saima Afrin completed her Bachelor of Science degree in the field of Computer Science and Engineering from the renowned Daffodil International University in the year of 2019. At present she is a lecturer of Daffodil International University in CSE department. Her research area contains Machine Learning and IoT.

Mohd. Saifuzzaman has obtained his M.Sc. in CS from Jahangirnagar University and B.Sc. in CSE from Daffodil International University. He currently works as a lecturer in the CSE Department at Daffodil International University in Bangladesh since September 2017. He prefers IoT work and Data Mining work. He is also an active member of DIU-NLP and Machine Learning Research Lab