A Novel Twdlnn And Mining Based Breast Cancer Prediction System in A Big Data Environment

Shahina Parveen M[1], U. Sakthi[2], Mrs. Thamari Thankam[3], T. Anjikumar[4], P. John Augustine[5], Raja Sarath Kumar Boddu[6]

[1]Associate Professor, Information Science & Engineering Department, CMR Institute of Technology, Bengaluru, Karnataka, India.

[2]Professor, Department of CSE, St.Joseph's Institute of Technology Semmancheri, Chennai, Tamil Nadu, India.

[3]Lecturer, Department of Health Science (Nursing), Bulehora University, Ethiopia.

[4]Assistant Professor, Department of CSE, Satya Institute of Technology and Management, Gajularega, Vizianagaram, Andhra Pradesh, India.

[5]Associate Professor, Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India.

[6]Professor and Principal, Department of CSE, Lenora College of Engineering, Rampachodavaram, Andhra Pradesh, India

shahinaparveenm@gmail.com[1], sakthi.ulaganathan@gmail.com[2], thangamsudhakar2000@gmail.com[3], anji5678@gmail.com[4], pjohnaugustine@gmail.com[5], Iamsarathphd@gmail.com[6]

Abstract

Amongst women, Breast Cancer (BC) has turned out to be the main reason for mortality. Predicting the BC early can help in saving the women as of the severe stage of cancer. Though most existing research has been utilizing disparate algorithms for prediction, they still lack in some areas, like accurate prediction and the execution speed. Thus, to trounce such cons, this paper proposed a novel Target Weight based Deep Learning Neural Network (TWDLNN) and mining based BC prediction system on a Big Data (BD) environment. The proposed paper totally comprises '4' steps: i) pre-processing, ii) Feature Selection (FS), iii) rule mining, and iv) classification. First, the Hadoop Distributed File Systems (HDFS) Map-Reduce (MR) function removes the redundant data, and also the missing attributes are swapped in the pre-processing step. Then, the Levy Flight based Chickens Swarm Optimizations (LFCSO) selects the vital features. Subsequently, the Associations Rule Mining (ARM) process is executed, wherein the CFI is attained. Next, the closed frequent itemset (CFI) is inputted to the TWDLNN algorithm that classifies the inputted data into a normal or cancer patient. In the experimental investigation, the proposed TWDLNN's performance is contrasted with the existing DLNN, ANN, SVM, along with RF-centred on the accuracy as well as execution time metrics.

Key words: Target Weight based Deep Learning Neural Network (TWDLNN), Levy Flight based Chicken Swarm Optimization (LFCSO) algorithm, Hadoop Distributed File System (DFS) and Closed Frequent Itemset (CFI).

1. INTRODUCTION

Mining with BD or BD mining has turned out to be a functional research field [1]. Additional data is offered in BD, like HealthCare Data (HCD), business information, etc. Here, HCD is a significant one. In this HCD, information related to cancer has attained the most views since cancer is the utmost feared diseases on the planet [2], and mostly female are affected by BC. BC has taken as the 2nd most leading cancer amongst female universally, which is observed to be augmented each year because of factors, like inheritance, lifestyle, along with dietary habits [3]. The disease should be identified at earlier stages [4] to avert the women's death. These days, BC analysis is trending and demanding subject [5].

For trouncing the concerns, several tests are generated to predict the BC and also some conventional techniques for the BC diagnosis [6]. Nowadays, artificial intelligence is attaining increasing attention to gather and analyze medical data. Due to the augmenting cancer cases, doctors find it hard to study every detail of cancer, and thus, encompassing automatic data analyzing equipment as support would help doctors to make an effective choice [7]. There are abundant algorithms for classification along with the prediction of BC results [8]. Nevertheless, classification algorithms suffer from less accurate predictions. Thus, this document exhibited a TWDLNN along with mining centred BC prediction system.

The paper is set as: Section 2 evaluates the associated work regarding the proposed work. Section 3 displays a concise discussion about the proposed BC prediction system. Section 4 analyses the proposed system's performance. Last, of all, section 5 wrapped up the paper.

2. RELATED WORK

Moloud Abdar along with Vladimir Makarenkov [9] offered a data mining method for a precise BC prediction. The scheme utilized the CWV-BANNSVM, joint boosting ANN (BANN) along with '2' SVM. In the experiential examination, the implemented techniques' performance was estimated centred on numerous well-known metrics as well as the CWV-BANN-SVM that had enhanced performances. However, it utilized the SVM along with the usual ANN. The ANN had the weight propagation issue, thus, it might produce poor accuracy.

Moloud Abdar *et al.* [10] established the nested ensembles system that utilized the Stacking as well as Vote as the classifiers amalgamation techniques on the ensemble techniques to detect the benign BC as of malignant cancers. In an investigational assessment, the '2'-layer nested ensembles classifiers having single classifiers (BayesNet together with Na¨ıve Bayes). Outcomes illustrated that the system had an enhanced outcome.

Suganthi Jeyasingh and Malathi Veluchamy [11] established a Modified Bat Algorithms (MBA) for FS to eradicate unrelated features as of a dataset. Primarily, the MBA was utilized for FS, along with the chosen features were inputted to the Random Forests (RF) classification algorithm. The scheme attained an improved outcome centred on numerous metrics.

Heng Kong *et al.* [12] offered a Jointly Sparse Discriminant Analysis (JSDA) to search the key features in BC and extricated the key features for enhancing the accurateness in diagnosing along with prediction. Experimental outcomes on BC datasets signified that JSDA trounced some popular sub-space learning algorithms on prediction accuracy.

Bichen Zheng *et al.* [13] established a hybrid of K-mean along with a support vectors machine (K-SVM) for BC prediction. The K-means was utilized to identify the concealed manner of the benign along with malignant tumors independently. After that, an SVM was employed to attain the classifier to differentiate the tumors. Centred on the experimental outcome, the K-SVM lessened the calculation time considerably lacking lesser diagnosis accuracy.

A number of experiments were performed to comprehend the prediction of BC utilizing data mining methods and disparate machine learning algorithms, however, those techniques have an issue in the accurate identification along with the speed of the prediction system (i.e, the prevailing techniques took the entire data for the prediction of BC, which in turn consumes more time). To trounce this issue of the prevailing research method, this paper utilizes the TWDLNN classifier centred BC prediction system in a BD environment.

3. BREAST CANCER PREDICTION SYSTEM USING TWDLNN CLASSIFIER IN BIG DATA ENVIRONMENT

BD is the compilation of data that is immense and still augmenting exponentially with time. Nowadays, BD is more preferred in healthcare for disease prediction. However, the accuracy and speed of the prediction process are the major issues that one faces during disease prediction. Thus, these issues are resolved by utilizing this TWDLNN and mining based BC prediction system. In the initial step, the HDFS removes the redundant data, and this system encompasses the map and reduces function. Centred on this function, the redundancy is shunned. Next, the LFCSO algorithm selects the features. After that, the CFI is extracted as of these selected features by utilizing the ARM. Lastly, the CFI is inputted to the TWDLNN for classifying the patient as cancer or normal patient. The structural design for the proposed work is exhibited in Figure 1,

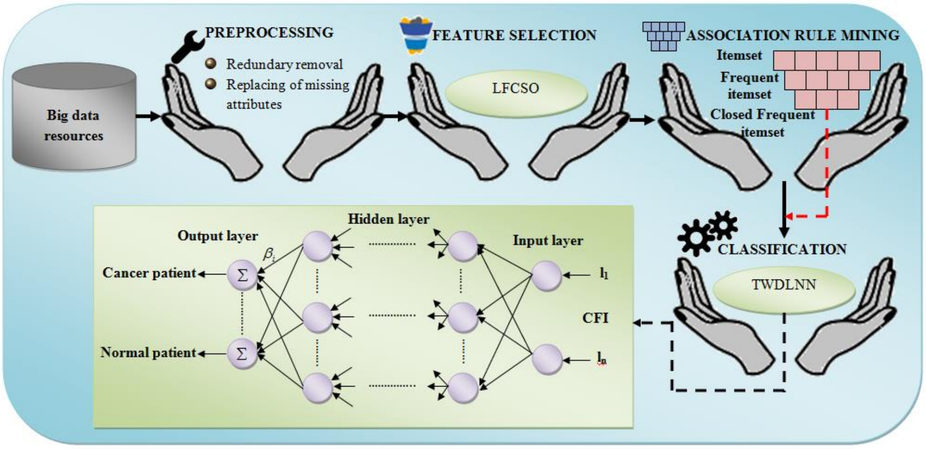


Figure 1: Block diagram for the proposed methodology

3.1 Pre-processing

Primarily, the BD resources (i.e., dataset) $\{(G_s = g_1, g_2, ...,g_n),\ or\ \ g_i,\ i = 1,2,....n\}$are pre-processed since BD comprises unstructured data. Therefore, the data ought to be preprocessed with '2' steps: i) removal of redundancy and ii) replace missing attributes.

Primarily, the HDFS Map-Reduce function removes the redundant information. The MR function is the best, simple, as well as parallel computing method, which is normally employed for the analyzation of huge data. The BC dataset encompasses a big quantity of data; therefore, this MR function is most useful for this BC prediction system. HDFS doesn't change the file once it is written. Therefore, if any changes are needed to be made, then the complete file must be rewritten. The HDFS has '2' phases for resolving the redundant data, that are i) map and ii) reduced function. The former function, initially, maps the entire data on the dataset, and then, the latter function reduces the data centred on the mapping.

In map function, the inputted data are split into tuples (key/ value pairs), i.e., the inputted data are transferred to the mapper function in a line by line manner. The map function m_f is signified as,

$$m_f = map(g_i), \qquad i = 1,2,......,n \tag{1}$$

Wherein, $map()$ implies the function, which takes care of mapping. Next, the map output is inputted to the reduce function r_f which unites those tuples centered on the key, which is expressed as,

$$r_f = reduce(m_f) \tag{2}$$

Wherein, $reduce()$ signifies the function that reduces the recurring values. Subsequent to the removal of redundant data, the proposed system replaces the missing values by means of taking the average means of the entire given values, which is expressed as,

$$\mu = \frac{\left(\sum g_i\right)}{n} \tag{3}$$

Wherein, μ implies the missing attribute, and n signify the number of data.

3.2 Feature Selection

Subsequent to pre-processing, the LFCSO is exhibited to choose the needed features since the dataset encompasses n-number of features. Chicken Swarm Optimization (CSO) stands as a bio-enthused algorithm centred on the hierarchy order as well as the movement of a swarm of chickens amid their food searching activities. This algorithm encompasses '3' search movements, say rooster, hen, together with chick movements. In the hen movement, the arbitrary number is utilized. The arbitrary selection doesn't render an effective outcome (explicitly, over searching issue happens). Thus, the proposed work regards the levy flight for arbitrary selection. Here, the rooster, hen, chicks, together with mother are regarded as features. Therefore, the '3' movements are elucidated as,

Rooster movement: As per the social behavior of rooster, the update technique of the roosters can well be designed as:

$$H_{i,j}^{t+1} = H_{i,j}^{t} .\left(1 + u_d\right) \tag{4}$$

Wherein, u_d signifies a Gaussian distribution, $H_{i,j}^{t+1}$ implies the updated solution, and $H_{i,j}^{t}$ signifies the current solution.

Hen movement: Hens goes around the group-mate roosters for foraging, and this condition can well be formulated as:

$$H_{i,j}^{t+1} = H_{i,j}^{t} + .\exp\left(\frac{\omega_i - \omega_{v1}}{\left|\omega_i\right| + cons}\right).\delta.\left(H_{v1,j}^{t} - H_{i,j}^{t}\right) + \exp\left(\omega_{v2} - \omega_i\right).\delta.\left(H_{v2,j}^{t} - H_{i,j}^{t}\right),\ \left(v_1 \neq v_2\right) \tag{5}$$

$$\delta = Levy \sim y^{-\alpha}\left(1 < \alpha \leq 3\right) \tag{6}$$

Wherein, δ signifies the levy flight function α implies the levy distribution function, y signifies the random function, v_1 denotes an index of the rooster, which is the i^{th} hens group mate, whilst v_2 implies an index of a chicken.

Chick Movement: Chicks can only forage around their mother hens and the update technique of the chicks is elucidated as:

$$H_{i,j}^{t+1} = H_{i,j}^{t} + e_l.\left(H_{m,j}^{t} - H_{i,j}^{t}\right) \tag{7}$$

Wherein, $H_{m,j}^{t}$ signifies the position of the i^{th} chick's mother in order that $m \in \left[1, N\right]$, e_l represents how much speed a chick runs to follow its mother. Lastly, the chosen features are,

$$F_s = \left\{k_1, k_2, k_3, \ldots\ldots\ldots\ldots k_n\right\}\ \ (or)\ \ k_i = 1,2,3,\ldots\ldots\ldots\ldots,n \tag{8}$$

Wherein, F_s implies the chosen feature set and k_n implies the n-number of chosen features.

3.3 Association Rule Mining

Subsequent to FS, the itemset, Frequent Itemset (FI), along with the CFI are extracted as of the chosen features. Therefore, the chosen features k_i are implied as the itemset. After that, the FI is extracted as of the itemset. The frequency of an itemset is basically the number of incidences of a specific item on the item-set. Subsequent to FI extraction, the CFI is extracted. CFI aim at discovering FI in a labeled test case. CFI is the highest number of occurrences of a specific item in the FI. The last output of CFI from the itemset and FI is expressed as,

$$L_c = l_i, \qquad i \in N \tag{9}$$

Wherein, L_c implies the CFI and l_i signifies the total CFI in the chosen features of the dataset.

3.4 Classification using TWDLNN Algorithm

Here, utilizing TWDLNN, the BC is predicted. The CFI’s output is inputted to the TWDLNN. The normal neural network comprises one hidden layer, in this deep learning classifier, more number of hidden layers is employed. Generally, the neural networks affected by the weight adjustments process (i.e., weight adjustment consumes more time and possible to render an inaccurate outcome). Thus, the proposed methodology utilizes TWDLNN. The TWDLNN comprises an input layer, n-number of hidden layers, along with an output layer. First, the CFI is given to the inputted layer. Next, the inputs layer's output is rendered to the hidden one. The hidden layer unit is computed as,

$$C_i = bias + \sum_{i=1}^{n} l_i \,.\, \beta_i \tag{10}$$

Wherein, C_i implies the hidden layer, and β_i signifies the weight value of the specific layer. Next, calculate the output unit that is computed as,

$$o_t = bias + \sum_{i=1}^{n} C_i \,.\, \beta_i \tag{11}$$

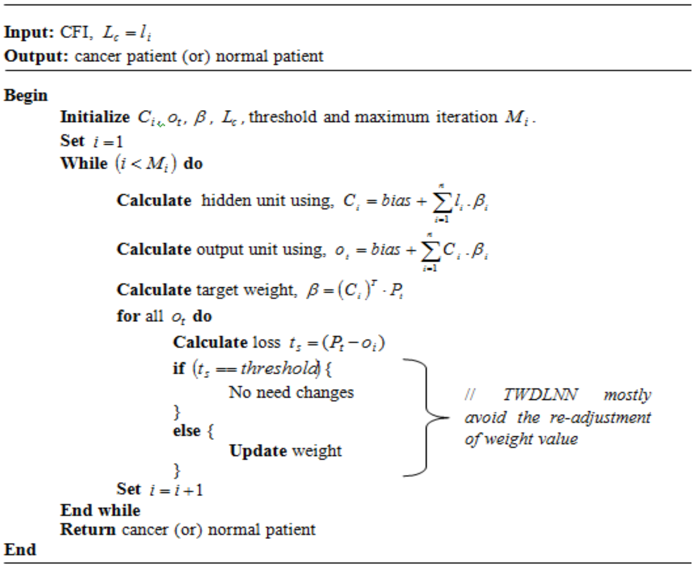
Wherein, o_t signifies the output unit, here, the weight value β_i is computed by the technique of target weight calculation that means the inverse value is computed for the hidden layer outputs, then the inverse hidden layer output is multiplied with the target value, which is denoted as,

$$\beta = \left(C_i\right)^T \cdot P_t \tag{12}$$

Wherein, P_t signifies the target value, and then, compute the loss function for identifying how much loss happened within the process utilizing the equation (13),

$$t_s = (P_t - o_t) \tag{13}$$

Wherein, P_t implies the target output of the system and t_s implies the loss function. The TWDLNN gives less loss function since it chooses the weight value centred on the target. The TWDLNN's Pseudocode is exhibited in Figure 2,



```
Input: CFI, $L_c = l_i$
Output: cancer patient (or) normal patient

Begin
    Initialize $C_i, o_t, \beta, L_c$, threshold and maximum iteration $M_i$.
    Set $i = 1$
    While $(i < M_i)$ do
        Calculate hidden unit using, $C_i = bias + \sum_{i=1}^{n} l_i \cdot \beta_i$
        Calculate output unit using, $o_i = bias + \sum_{i=1}^{n} C_i \cdot \beta_i$
        Calculate target weight, $\beta = (C_i)^T \cdot P_i$
        for all $o_t$ do
            Calculate loss $t_s = (P_t - o_i)$
            if $(t_s == threshold)$ {                // TWDLNN mostly
                No need changes                      avoid the re-adjustment
            }                                        of weight value
            else {
                Update weight
            }
        Set $i = i + 1$
    End while
    Return cancer (or) normal patient
End
```

Figure 2: Pseudocode for TWDLNN

In Figure 2, the hidden along with output unit of the TWDLNN is elucidated and the targeted weight also elucidated. The CFI of the patient HCD is taken as the inputs. By utilizing this TWDLNN, HCD is effectively categorized as normal and BC patient data.

4. RESULT AND DISCUSSION

Here, the proposed system's performance is examined. The proposed work is executed in the JAVA. For examining the performance, the proposed work takes the data as of the Wisconsin BC Database (WBCD). It encompasses '2' datasets, the 1st dataset includes recurrent along with non-recurrent class with 198 patient data together with 34 features, the 2nd dataset includes benign together with malignant class, which has 571 patient data with 32 features. The '2' datasets are united, and then, eradicated the redundant features. Therefore, the features in the '10' types for every cell nucleus are radius, perimeter, area, concave points, smoothness, texture, compactness, concavity, symmetry, along with fractal dimension. For every group, '3' pointers are calculated: mean value, standard error, as well as the maximal value.

4.1 Performance Analysis

Here, the proposed TWDLNN's performance is contrasted with the existing Deep Learning Neural Networks (DLNN), Artificial Neural Networks (ANN), SVM, as well as RF-centered on accuracy along with execution time.

Table 1: Performance analysis based on accuracy metric

Techniques	Accuracy (%)
Proposed TWDLNN	97%
DLNN	89%
ANN	84%

SVM	80%
RF	76%

Discussion: Table 1 contrasted the proposed TWDLNN's performance with the DLNN, ANN, SVM, as well as RF- centred on the accuracy metric. The accuracy is examined centred on the number of Data Counts (DC), and the average of the entire DC accuracy is envisioned in a table form. The TWDLNN attains high accuracy, namely 97%, however, the DLNN, ANN, SVM, along with RF have a low accuracy than the proposed one, i.e., 89%, 84%, 80%, together with 76%, correspondingly. Therefore, it deduces that the TWDLNN attains enhanced performance contrasted with the existing research method, which is illustratively signified in Figure 3.

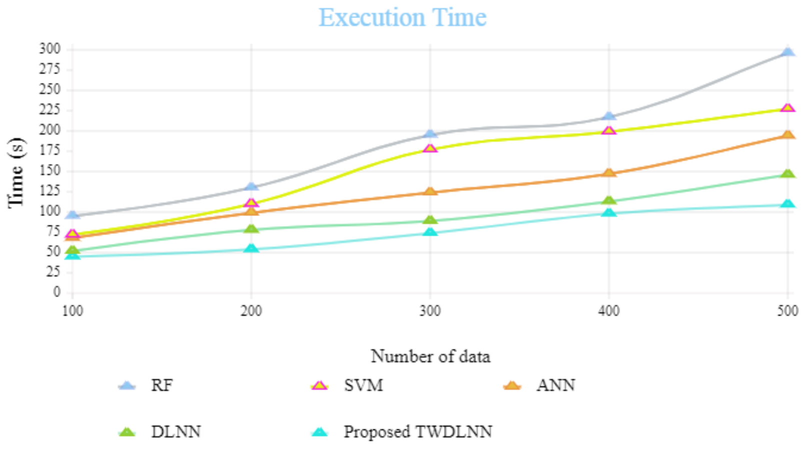


Figure 3: Comparison graph based on an accuracy metric

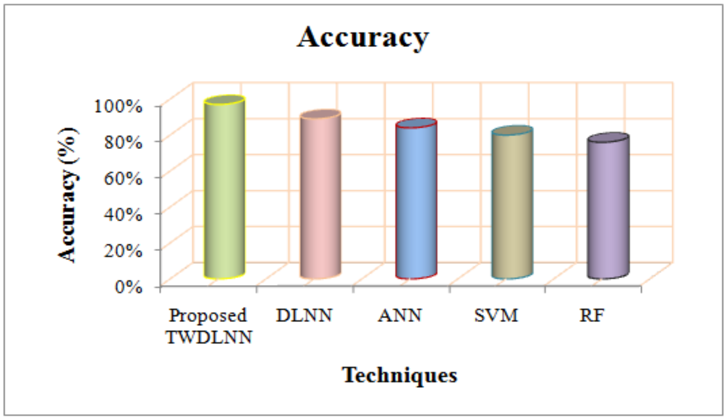


Figure 4: Execution time analysis

Discussion: Figure 4 examines the TWDLNN’s performance with that of the DLNN, ANN, SVM, along with RF-centred on the execution time. The execution time is varied centred upon the number of DC. The DC starts as of 100 and stops with 500. While the DC is 500, the TWDLNN has 95s time for implementing the process, however, the DLNN, ANN, SVM, along with RF consumes 130s, 195s, 217s, as well as 296s for performing the process. TWDLNN takes lesser time than the former methods, which illustrates that the proposed scheme has a high speed than the preceding algorithm. Likewise, centred on the remaining DC, the proposed work takes less time than the preceding methods. It deduces that the TWDLNN attains improved speed than the preceding methods.

5. CONCLUSION

Here, a novel TWDLNN and mining centred BC prediction system is presented in a BD. The proposed system comprises i) preprocessing, ii) FS, iii) ARM, as well as iv) classification phases. In pre-processing, the HDFS map-reduce function removes the redundant data and FS, and the LFCSO is employed for the classification. The TWDLNN classifies the data as a normal or cancer patient. For the performance analysis, the data is taken as of the WBCD dataset. In performance analysis, the proposed TWDLNN is contrasted with the DLNN, ANN, SVM, and RF-centred on accuracy and execution time. Centred on these metrics, the proposed system attained its objective. The TWDLNN has 97% accuracy, and for 500 DC, it takes 95s time. Therefore, it inferred that the proposed attain superior performance to the prevailing techniques. In the future, the proposed work can well be extended by means of analyzing the hazard of BC.

REFERENCE

1. Tsai, C.-F., Lin, W.-C., & Ke, S.-W, “Big data mining with parallel computing: A comparison of distributed and MapReduce methodologies”, Journal of Systems and Software, vol. 122, pp. 83-92, 2016.

2. Tee, I. C., & Gazala, A. H, “A novel breast cancer prediction system”, International Symposium on Innovations in Intelligent Systems and Applications, pp. 621-625, 2011.

3. Kamel, S. R., YaghoubZadeh, R., & Kheirabadi, M, “Improving the performance of support-vector machine by selecting the best features by Gray Wolf algorithm to increase the accuracy of diagnosis of breast cancer”, Journal of Big Data, vol. 6, no. 1, 2019.

4. Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., & Bhardwaj, A, “Unbalanced breast cancer data classification using novel fitness functions in genetic programming”, Expert Systems with Applications, pp. 112866, 2019.

5. Hajiabadi, H., Babaiyan, V., Zabihzadeh, D., & Hajiabadi, M, “Combination of loss functions for robust breast cancer prediction”, Computers & Electrical Engineering, vol. 84, pp. 106624, 2020.

6. Priyanka Gupta, and L. Shalini, “Analysis of machine learning techniques for breast cancer prediction”, International Journal Of Engineering And Computer Science, vol. 7, no. 05, pp. 23891-23895, 2018.

7. Hajiabadi, H., Babaiyan, V., Zabihzadeh, D., & Hajiabadi, M, “Combination of loss functions for robust breast cancer prediction”, Computers & Electrical Engineering, vol. 84, pp. 106624, 2020.

8. Asri, H., Mousannif, H., Moatassime, H. A., & Noel, T, “Using machine learning algorithms for breast cancer risk prediction and diagnosis”, Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.

9. Abdar, M, “CWV-BANN-SVM ensemble learning classifier for early diagnosis of breast cancer”, Measurement, vol. 169, pp. 557-570, 2019.

10. Abdar, M., Zomorodi-Moghadam, M., Zhou, X., Gururajan, R., Tao, X., Barua, P. D., & Gururajan, R, “A new nested ensemble technique for automated diagnosis of breast cancer”, Pattern Recognition Letters, 2018.

11. Jeyasingh, Suganthi, and Malathi Veluchamy, “Modified bat algorithm for feature selection with the wisconsin diagnosis breast cancer (WDBC) dataset”, Asian Pacific journal of cancer prevention: APJCP, vol. 18, no. 5, pp. 1257, 2017.

12. Kong, H., Lai, Z., Wang, X., & Liu, F, “Breast cancer discriminant feature analysis for diagnosis via jointly sparse learning”, Neurocomputing, vol. 177, pp. 198–205, 2016.

13. Zheng, B., Yoon, S. W., & Lam, S. S, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms”, Expert Systems with Applications, vol. 41, no. 4, pp. 1476–1482, 2014.