A MODIFIED SVM ALGORITHM TO ENHANCE THE CANCER CLASSIFICATION

RETHINA KUMAR[1], GOPINATH GANAPATHY[2], JEONG-JIN KANG[3]

[1]Research Scholar, Bharathidasan University, India.

[2]Professor, Bharathidasan University, India.

[3]Professor, Dept of Information and Communication, Dong Seoul University, Seongnam, Korea.

ABSTRACT

The survival rate of breast cancer patients has increased due to the advancements in the treatment of the disease. These include the use of newer and more effective drugs. There are various types of breast cancer, which can be treated through different methods. Currently, there are numerous studies that are focused on developing a better understanding of this disease. In order to improve the classification of breast cancer, a new machine learning algorithm is proposed. This method uses support vector machine learning to enhance the performance of existing models. The proposed model can rectify the inconsistencies in the existing breast cancer dataset and improve its performance. It can also create a high-quality Wisconsin Diagnostic Breast Cancer (WDBC) data set. The proposed model can then predict the likelihood of a patient developing breast cancer. It can also diagnose the patient based on the data collected. The researchers were able to test the proposed model against several machine learning models. They were able to achieve high accuracy levels.

Keywords: *Breast Cancer, Machine Learning, Diagnosis, Prediction, Benign, Malignant*.

1. INTRODUCTION

It can be hard to believe that cancer is real. Learning about it can induce confusion and prevent people from making an informed decision. [1]. Through deep medical knowledge, people can make informed decisions regarding their treatment [2]. This helps them make the best decisions for themselves and their families. Getting the right treatment for a particular type of cancer can greatly improve a patient's chances of survival[3]. This is one of the main reasons why it is important for researchers to study the various treatment options available for cancer. Machine learning is also a major component of the cancer research conducted by researchers [4]. It has been used to develop new algorithms and models that can help detect the presence of breast cancer. Support Vector machines are also being studied to help develop more accurate models for predicting the course of cancer [5,6].

Various machine learning tools are used to predict cancer. Some of these include regression trees, kNearest neighbors, and Naive Bayes classifiers. [7]. The Support Vector Machine is one of these tools that can predict breast cancer in terms of its type[8]. It can also predict Malignant and Benign tumors. The Support Vector Machine is able to perform better than traditional methods when it comes to detecting breast cancer [9]. However, it still needs to be validated and implemented properly. This study aims to develop a modified support vector machine that can accurately classify breast cancer. It has been trained with a population of data and can detect two classifications – Benign and Malignant [10].

2. RELATED WORKS

Omondiagbe Et al. [11]- The researchers used a radial basis kernel and a support vector machine to develop their hybrid approach. They noted that by adopting a hybrid approach, it was possible to reduce the high dimensionality of features. This could then be used to diagnose breast cancer more precisely. In this study, the authors were able to detect breast cancer at an accuracy of 98.82%. They also tested the effectiveness of the various algorithms used in the study.

Tseng Et al. [12]- In this study, the researchers utilized machine learning techniques to predict the spread of breast cancer using the features of clinic nature. They were able to do so through the use of statistical logistic regression and SVM. **Bennett** Et al [13] discussed SVM is commonly used in the diagnosis of diseases. It was able to generate a more accurate prediction of the spread of breast cancer than a decision tree.

Akay Et al. [14]- this study revealed that SVM performed remarkably well in diagnosing breast cancer. It provided a score of 98.53%, 99.02%, and 99.51% for the various training test sections. The study used the same data set and five features that were selected by a genetic algorithm. The features were then chosen according to their rank in the feature discrimination test. The SVM trained the input continuously until it was able to perform optimally. The feature selection algorithm was able to reduce the number of features and eliminate the noisy data.

3. PROPOSED METHODOLOGY

The proposed model is a modified SVM algorithm that aims to enhance the breast cancer classification. It is presented in Figure 1. The algorithm first normalizes the data attributes before generating a visual representation of the data distribution. After analyzing the data, the model visualizes the distribution by means of a Gaussian distribution. It then split the data into target and predictor variables. For the proposed model, the newly formed data was separated into two datasets(WDBC): a training and a testing one. The goal of the proposed model is to improve the performance of its few algorithm by using a standardized dataset.

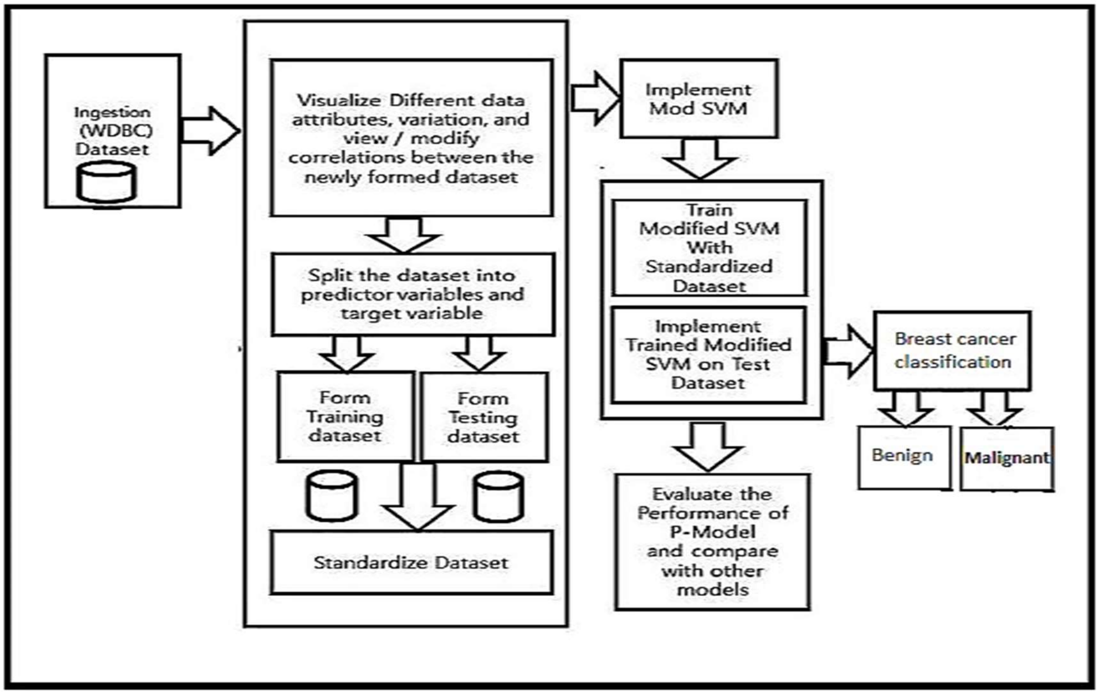

Ingestion (WDBC) Dataset
Visualize Different data attributes, variation, and view / modify correlations between the newly formed dataset
Split the dataset into predictor variables and target variable
Form Training dataset
Form Testing dataset
Standardize Dataset
Implement Mod SVM
Train Modified SVM With Standardized Dataset
Implement Trained Modified SVM on Test Dataset
Breast cancer classification
Benign
Malignant
Evaluate the Performance of P-Model and compare with other models

Figure 1: Proposed Model Architecture Frame Work.

3.1 Modified SVM for Breast Cancer

In this section, the proposed model for the SVM is described in detail. Figure 2 shows the steps of Modified SVM approach. Datasets= 'D' (d1, d2 …dn), Data points in D= n, Clusters=K, Cluster Center= c, 'X' (x1, x2, x3, …xn) is the data point. Threshold = Th.The proposed system uses a training dataset to inform the SVM about the various variables that can affect its performance. The breast cancer classification system is a type of machine learning that uses statistical learning techniques. The classification system usually divides the data into groups based on the N-dimensional hyperplane. The number of training samples that are included in the dataset is referred to as the sample class xi. The goal of support vector machines is to find a maximum margin that separates a hyperplane from its closest points in an extremely high-dimensional space.

Minimise

$$W(\alpha)\frac{1}{2}\ \sum_{J=1}^{N}\sum_{NJ=1}\ \alpha i\ \alpha jk(xi\ ,xj) - \sum{}^{N}{}_{i=1}\alpha i \qquad (1)$$

Subject to: $\forall i$: $0 \leq \alpha_i \leq C$ and $\sum^{N}_{i=1} \alpha_i y_i = 0$

The soft margin is the kf of statistical machines (SVMs). It is a group of functions that can be used to divide data into various groups. For instance, the linear kernel, the polynomial kernel, the radial basis function, and the Sigmoid kernel can be used.

The goal of this paper is to develop a set of classification models that can reduce the variance and bias issues in classification. The complexity of the models' setting parameters can greatly affect the accuracy of their classification. For instance, it is important to consider the various alternatives of kernel functions. In this paper, a radial basis function is used to set the parameters for a SVM, which then uses a t fit for its classification. The performance evaluation of classification models is also performed through the use of a matrix known as the confusion matrix. It shows the difference between positive and negative cases of breast cancer.

In SVM algorithm modification, we can specify two key parameters. One of these is the value of C and the other is the type of kernel which is used. For SVM, the default kernel is called Radial Basis Function (RBF). This algorithm will use a grid search method. In order to perform the search over a combination of C values, the algorithm will need to consider the various kernel types.

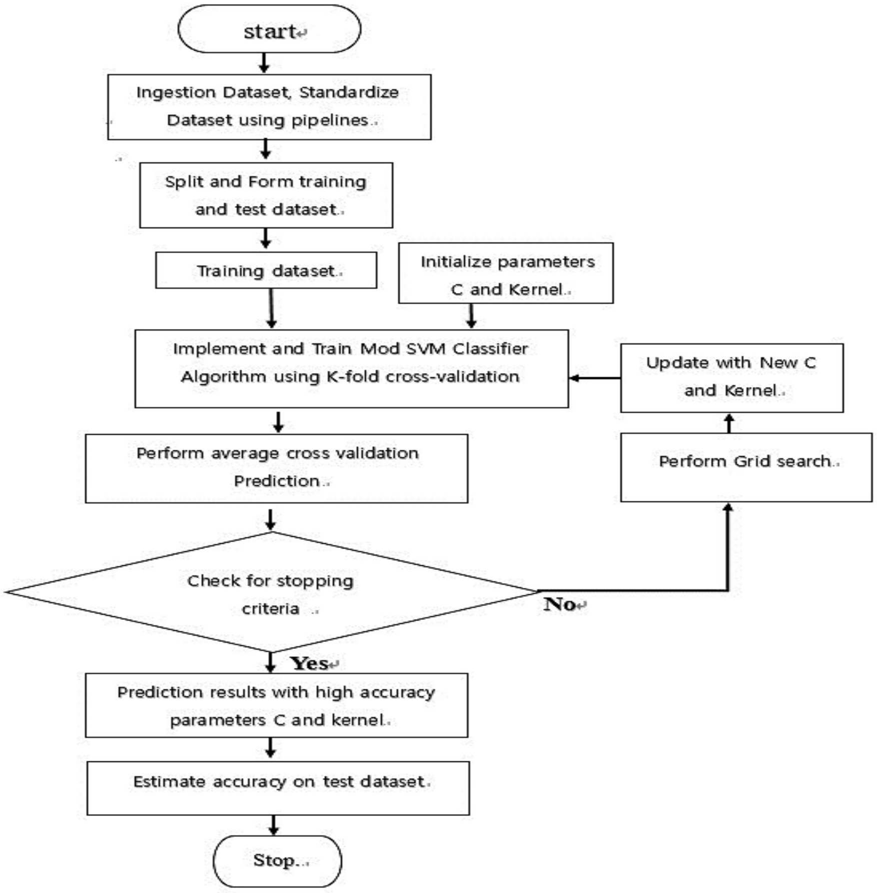

start
Ingestion Dataset, Standardize Dataset using pipelines
Split and Form training and test dataset
Training dataset
Initialize parameters C and Kernel
Implement and Train Mod SVM Classifier Algorithm using K-fold cross-validation
Update with New C and Kernel
Perform average cross validation Prediction
Perform Grid search
Check for stopping criteria
No
Yes
Prediction results with high accuracy parameters C and kernel
Estimate accuracy on test dataset
Stop

Figure 2: Modified SVM Algorithm for breast cancer prediction flow chart

4. RESULT AND DISCUSSION

The proposed model is based on a modified SVM algorithm and a breast cancer dataset. It is implemented in Python using the various functions and the WDBC dataset.

4.1. Dataset Description

The study analyzed the Wisconsin Diagnostic Breast cancer Dataset, which contains over 500 cases of breast cancer. The data included information about patients' characteristics, such as their ID and tumor distinguishing features. The data was then analyzed using a computer image. The features were then extracted and analyzed. The statistical features that were analyzed were then categorized into three categories: standard error, maximum value, and mean.

4.2. Preprocess and visualize the (WDBC) Dataset

In Python, the WDBC dataset is presented as a group with no diagnosis label. In Figure 4, the data attributes have been grouped together in a pairplot.

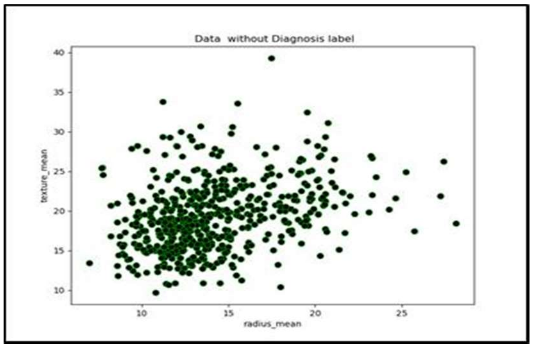


Figure 3: Dataset Without Diagnosis Label.

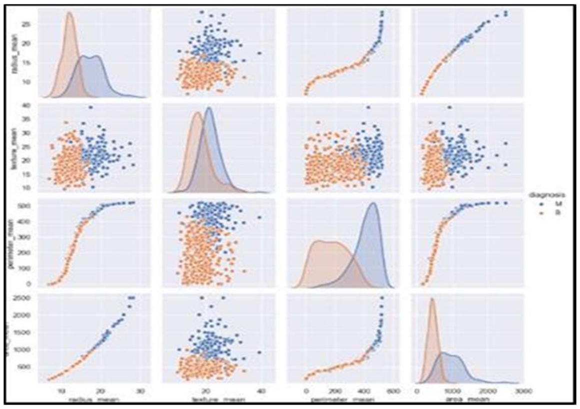


Figure 4: Data Attributes With Diagnosis Lable (M/B).

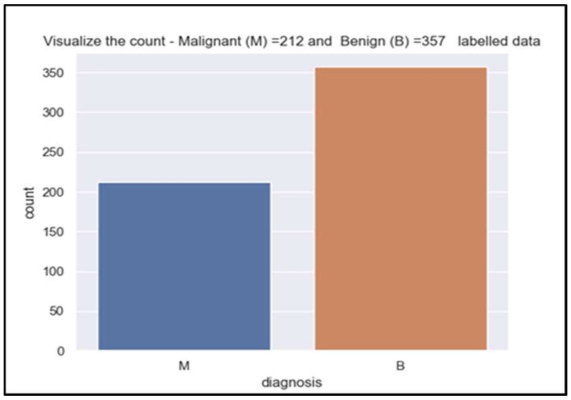


Figure 5: Diagnosis Count With Diagnosis Lable (M/B).

In the fig 5 describes the WDBC dataset has been used to visualize a group of Malignant and Benign with a total number of counts.

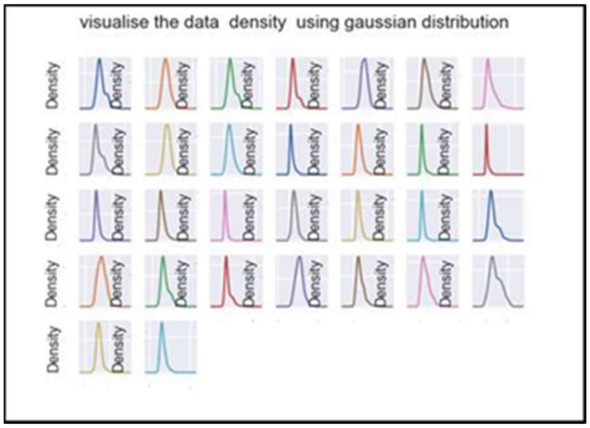


Figure 6: Data Density Using Gaussian Distribution.

Using density plots, we can see that the data distribution is linear. Likewise, the red lines indicate that the various data attributes are correlated.

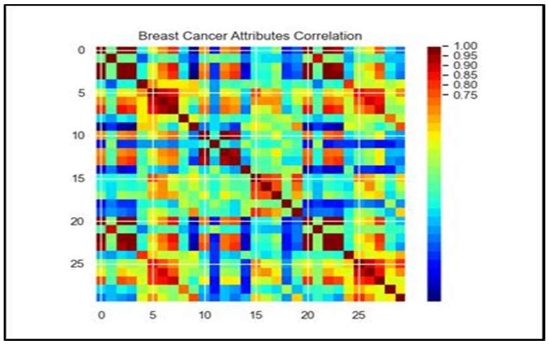
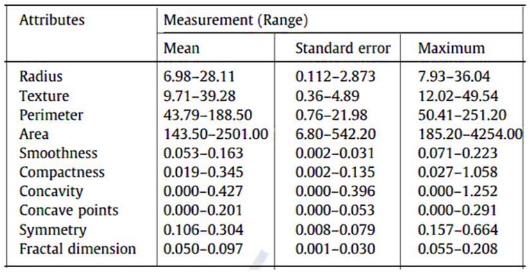


Figure 7: WDBC Data Attributes Correlation.

The data distribution can be analyzed using density plots and the general Gaussian distribution. The red line shows the correlation between the various attributes.

For our proposed modified SVM, we have obtained the best possible C value and kernel type for our requirements. The regularization parameter is used to trade off the correctness of training data against the margin of the decision function. If the C values are larger, the decision function will accept a smaller margin if it can correctly classify all training points.

Table 1: WDBC Data Attributes.



Attributes	Measurement (Range)		
	Mean	Standard error	Maximum
Radius	6.98–28.11	0.112–2.873	7.93–36.04
Texture	9.71–39.28	0.36–4.89	12.02–49.54
Perimeter	43.79–188.50	0.76–21.98	50.41–251.20
Area	143.50–2501.00	6.80–542.20	185.20–4254.00
Smoothness	0.053–0.163	0.002–0.031	0.071–0.223
Compactness	0.019–0.345	0.002–0.135	0.027–1.058
Concavity	0.000–0.427	0.000–0.396	0.000–1.252
Concave points	0.000–0.201	0.000–0.053	0.000–0.291
Symmetry	0.106–0.304	0.008–0.079	0.157–0.664
Fractal dimension	0.050–0.097	0.001–0.030	0.055–0.208

Different features of the dataset are treated as different representations in the model. In order to minimize the effects of different scales on the error function, normalization is performed before the training is started.

4.3. Measure for Performance Evaluation

The proposed modified SVM is tested on a standardized data set. It is then trained using a high quality training dataset to classify the data into two categories: benign and malignant.

Support may be defined as the number of samples of the true response that lies in each class of target values.

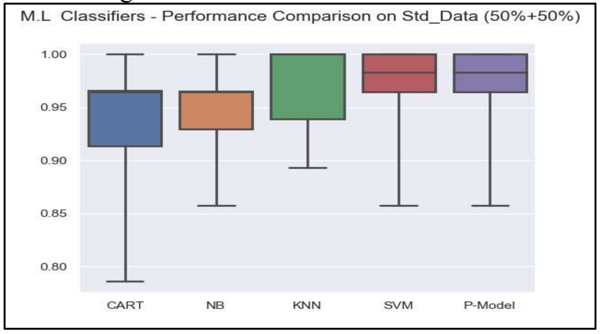


Figure 8: Comparison of Standardized Data (50% + 50%).

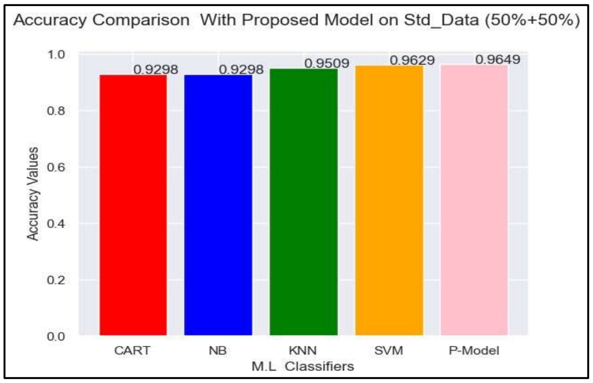


Figure 9: comparison of Accuracy

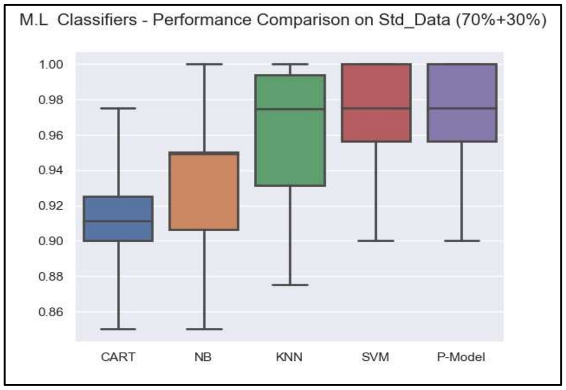


Figure 10: Performance Comparison on Standardized Data

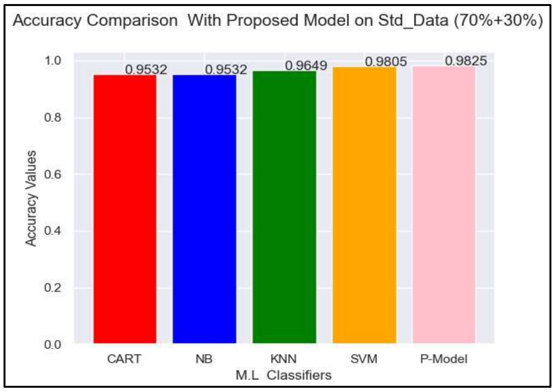


Figure 11: Accuracy Comparison with Proposed Model (70%+30%).

The graph shows the performance comparison between our modified SVM and other machine learning algorithms such as NB, KNN, CART, and KNN on a standardized dataset. The result shows that our modified SVM performed better than the other machine learning algorithms.

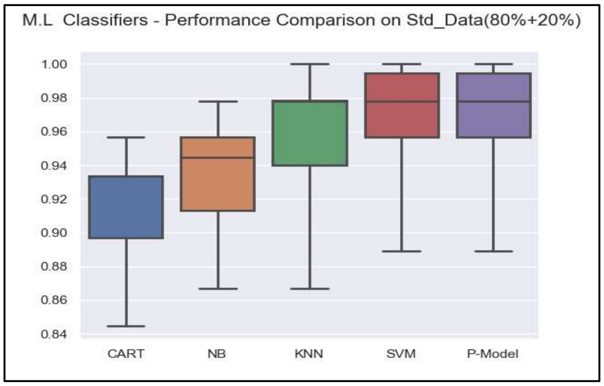


Figure 12: Performance on Standardized Data (80% + 20%).

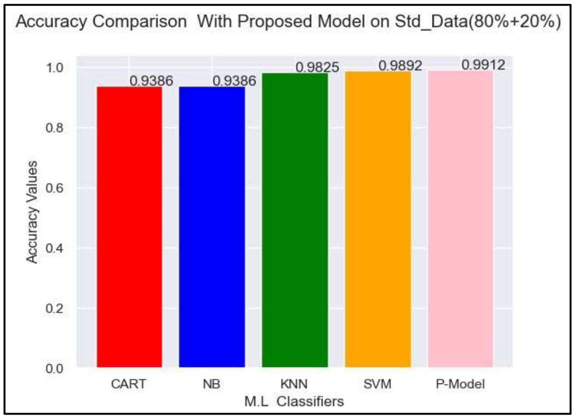


Figure 13: Accuracy Comparison with Proposed Model (80%+20%).

Figure 12 and 13 shows the comparison of our modified SVM against other machine learning algorithms. It got the best accuracy of 0.99% while achieving a linear kernel configuration. The proposed model shows stable performance and high accuracy in prediction. It is based on the confusion matrix, which divides breast cancer cases into two classes: positive and negative. The accuracy of the proposed model is better than that of other models such as KNN, SVM, and NB. It is also more accurate than the testing WDBC dataset.

5. CONCLUSION

In this paper, we present a modified SVM algorithm that can improve the prediction and classification of cancer. It has high accuracy performance. The training datasets provided by our model allow us to improve the overall performance of our algorithm. Through the use of a training dataset, our proposed model has achieved 99% accuracy when compared to the tested dataset. The goal of this study was to develop a prediction model that can perform better than those currently proposed in studies. It can classify breast cancer according to its various categories.

REFERENCES

[1] F. Liu and M. Brown, "Breast Cancer Recognition by Support Vector Machine Combined with Daubechies Wavelet Transform and Principal Component Analysis", In: Proc. of the International Conf. on ISMAC in Computational Vision and BioEngineering, Springer, pp. 1921-1930, 2018.

[2] P. Exarchos a,Michalis V. Karamouzis "Machine learning applications in cancer prognosis and prediction" Computational and Structural Biotechnology Journal 13 (2015) 8–17.

[3] L. Yang and Z. Xu, "Feature extraction by PCA and diagnosis of breast tumors using SVM with DE-based parameter tuning", International Journal of Machine Learning and Cybernetics, Vol. 10, No. 3, pp. 591-601, 2019.

[4] R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, "Classification of mammogram for early detection of breast cancer using SVM classifier and Houghtransform", Measurement, Vol. 146, pp. 800-805, 2019.

[5] V. Chaurasia, S. Pal, and B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms and Computational Technology, Vol. 12, No. 2, pp.119-126, 2018.

[6] W. L. Al-Yaseen, Z. A. Othman. "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system", Expert Systems with Applications, Vol.67, pp. 296-303, 2017.

[7] M. Kumar, A. J. Kulkarni, and S. C. Satapathy, "A Hybridized Data Clustering for Breast Cancer Prognosis and Risk Exposure Using Fuzzy C-means and Cohort Intelligence", Optimization in Machine Learning and Applications, Springer, pp. 113-126, 2020.

[8] G. F. Stark, G. R. Hart, B. J. Nartowt, and J. Deng, "Predicting breast cancer risk using personal health data and machine learning

models", Plos One, Vol. 14, No. 12, pp. 1-17, 2019.

[9] F. AlFayez, M. W. A. El-Soud, and T. Gaber, "Thermogram Breast Cancer Detection: a comparative study of two machine learning techniques", Applied Sciences, Vol. 10, No. 551, pp. 1-20, 2020.

[10] P. Ferroni, F. M. Zanzotto, S. Riondino, N. Scarpato, F. Guadagni, and M. Roselli, "Breast cancer prognosis using a machine learning approach", Cancers, Vol. 11, No. 328, pp. 1-9, 2019.

[11] **D. A. Omondiagbe,** S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis", In: Proc. of IOP Conf. Series: Materials Science and Engineering, Vol. 495, pp. 1-16, 2019.

[12] **Y. J. Tseng**, C. E. Huang, C. N. Wen, P. Y. Lai, M. H. Wu, Y. C. Sun, H. Y. Wang, and J. J. Lu, "Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies", International Journal of Medical Informatics, Vol. 128, pp. 79-86, 2019.

[13] **Bennett,** K. P., & Blue, J. A. (1998). A support vector machine approach to decision trees. In Proceedings of IEEE world congress on computational intelligence (pp. 2396–2401). Anchorage, AK: IEE.

[14] **Akay,** M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications, 36, 3240–3247.

[15] R. Turkki, D. Byckhov, M. Lundin, J. Isola, S. Nordling, P. E. Kovanen, C. Verrill, K. von Smitten, H. Joensuu, J. Lundin, and N. Linder, "Breast cancer outcome prediction with tumour tissue images and machine learning", Breast Cancer Research and Treatment, Vol. 177, pp.41-52, 2019.

[16] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning", Ultrasonics, Vol. 91, pp. 1-9, 2019.

[17] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification", Computer Science, pp. 1-16, 2008.

[18] Youness Khourdifi, Mohamed Bahaj,"Applying Best Machine Learning Algorithms for Breast Cancer Prediction and Classification", International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), 2018.

[19] Abien Fred M. Agarap,"On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset", ICMLSC , February 2–4, 2018, Phu Quoc Island, Viet Nam, 2018.

[20] Dana Bazazeh and Raed Shubair ,"Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis", 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016.

[21] Hui-Ling Chen , Bo Yang, Jie Liu , Da-You Liu,"A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis,H.-L. Chen et al. - Expert Systems with Applications 38 9014–9022, 2015

[22] Muhammad Hussain, Summrina Kanwal Wajid, Ali Elzaar, Mohammed Berbar,"A Comparison of SVM Kernel Functions for Breast Cancer Detection", Eighth International Conference Computer Graphics, Imaging and Visualization, 2014.