OUTLIER SELECTION METHOD FOR CLASSIFICATION OF BREAST CANCER

Tintu P B[1], S Manju Priya[2]*

ABSTRACT

One of the deadliest diseases, cancer is believed to be the second biggest cause of death for women internationally. The patient's life may be saved by early cancer identification. Breast tissues give rise to cancerous tumors that spread to other regions of the body and ultimately cause death. Cancer in women cases are anticipated to rise to a total of 2.3 million or higher in 2023, according the World Health Organization. Cancer may impact both males and females. A prompt and accurate diagnosis contributes to a higher patient survival rate. It is challenging to support medical professionals in developing a regimen that might prolong the life of valetudinarians because of the need for precise prophecy to identify symptoms. Rapid detection must be combined with effective cancer therapy, which frequently calls for the highest level of specialized cancer care. Methods based on machine learning can be applied to increase the diagnosis' speed and accuracy. If the precision is flawless, the model will function with greater efficiency and increase breast cancer diagnosis. Two independent machine learning algorithms were employed to perform outlier analysis on the Wisconsin Diagnostic Breast Cancer Dataset in order to increase the accuracy of breast cancer detection.

Keywords: outliers, breast cancer, accuracy and machine learning

I. INTRODUCTION

When a cell grows improperly, cancer enters the human body and can permeate or spread throughout the body [1]. Breast cancer is thought to be the second-leading cause of death for women and has a high mortality rate [2]. According to a five-year study, breast cancer is the most common type of cancer and is becoming more common in emerging countries [3]. With almost 2.3 million cases reported annually, breast cancer affects more people than any other type of cancer, according to the WHO. Breast cancer is the primary or secondary cause of cancer-related deaths among women in 95% of the world's countries. As per estimates, twenty million additional cases of cancer were reported around the world, and ten million deaths due to cancer occurred. In accordance with estimates, there will be 297,790 additional cases of breast cancer that are invasive in women and 2,800 new cases in males in 2023[4]. Fast cancer diagnosis is essential in order to combine effective cancer treatment, which typically needs some degree of professional cancer care. Breast cancer can be categorized into two types: aggressive and harmless. The disease is characterized by abnormal cell replication in the breast [5]. This classification can be used to describe any aberrant growth of tissues, tumors, or additional developments. Although non-cancerous cells in benign conditions are typically harmless, malignant breast cancer provides a major threat to a patient's survival. and are more probable to survive [6]. if a tumor that is malignant grows and invades adjacent breast tissue, it can propagate to nearby lymph nodes or other parts of the body. Cancer can be treated and even cured if the cancer is malignant and discovered early on [7].

Age, obesity, drinking alcohol, and smoking constitute a few of the biological risk factors for breast cancer. The identification of breast cancer is a challenging task. Early diagnosis can be achieved through a wide range of methods such as ultrasound, true cut biopsy, fine needle aspirin examination, mammograms, and so on [8,9]. Many times, professionals are unable to correctly diagnose the illness and this test might not provide an appropriate diagnosis. To

[1,2]Department of Computer Science,
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
* Corresponding Author

ascertain which factor has the biggest influence, medical professionals are still performing this type of diagnostic. However, algorithms based on Machine learning [10], the use of deep learning [11], [12], or bioinspired computation [13] techniques have been used in a number of medical diagnoses in recent years. For the purpose of identifying breast cancer, the use of machine learning has become the subject of multiple studies [14]. The study has continuously concentrated on increasing prediction accuracy to allow precise diagnosis regardless of the features of the dataset. However, machine learning (ML) modalities that have been shown on multiple breast cancer dataset still aren't able to provide reliable and consistent diagnosis results unless they are improved using particular data mining techniques [15].

II. REVIEW OF LITERATURE

The evaluation and evaluation of algorithms that use machine learning as well as the techniques used to improve their performance were detailed in the most recent study using the WDBC dataset [16]. [17] described an approach using grouped data and noise removal of the WDBC dataset before subjecting it to any ML algorithms. The WDBC dataset [18] underwent preprocessing to improve the precision of the classification of breast cancer as an illness. Gain ratio was used to identify the features, and six techniques were combined with the 10-fold cross-validation approach to model the features. In a similar vein, [19] focused on integrating an algorithm for machine learning with a number of approaches for selecting attributes; the outcomes determined which approach was more successful. The features that were selected were PCA, recursive features removal, linear discriminant analysis, and correlation-based feature selection. In order to improve the method for classification on the WDBC dataset, alternative cross-validation level and divide training dataset percentages were investigated in the comparison study carried out by [20]. When the learning batch was 85.5%, which resulted in 99% accuracy, improvement was shown. It is reasonable to argue that this improvement resulted from overfitting of the training set. Similar work was conducted by [21] using the WBDC, WDBC datasets, and that study proposed a fuzzy technique for improving ML. The Light gradient-boost Approach, AdaBoost , and Extreme gradients boost feature selection strategies were all examined using the Naive Bayes algorithm[22]. [23] evaluated the feature selection techniques of recursive feature elimination, univariate selection, and correlation-based feature selection once more. The Random Forest was used to these feature selection strategies, and the results were evaluated using the WDBC dataset. As far we are aware, no study has been done to assess the existence of outlier on the Wdbc data set, hence the problem of multi-linearity across the Wdbc dataset still exists, corresponding to an examination of the scientific literature for the cutting-edge approaches used.

III. PROPOSED METHOD

An outlier is a statistic or observation that deviates from a distribution's regular pattern. A tiny percentage of information that is noticeably unusual or outside of the general trend is known as outliers. The resulting skewness affects the distribution's parameters such as mean and standard deviation. As shown in Figure 1, this analysis reveals the existence of misfits in the dataset. As a result, outliers were found and eliminated from the pertinent attributes.

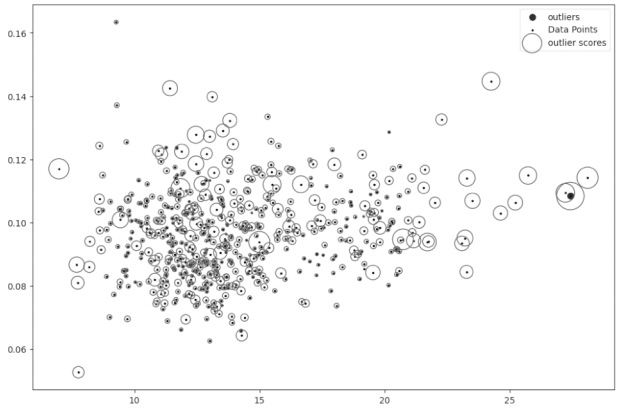

Figure 1: point plots the existence of outliers

3.1 Hybrid Approach to Outlier Selection

The HOBSM (Hybrid Outlier Behavior Selection Method) method was used to scale the features in order to remove outliers from the WDBC dataset. When features

include marginal outliers, the robust scaling in this method comes very handy. The equations can be expressed as

$$Xinew = (Xi\text{-}Xmedian) / IQR \ldots\ldots\ldots (1)$$

Where IQR means Inter Quartile Range. The non-outlier data was sent to the isolation forest after analysis, and the average absolute variance was calculated. Pearson's Correlation was used to reduce the dimension. Consider an attribute set F1 and a dataset D1.

$$\text{where} \quad F1 = \{x1, x2, x3. \ldots, xn\} \ldots\ldots\ldots\ldots (2)$$

Using a filter technique, the most predictive characteristics were identified. This page figures out and shows how each attribute's correlation with the others. According to this relationship standards, which is set at 0.5 features in the used for training dataset with a correlation coefficient of less than or equal to 0.5 are disregarded. Conversely, other characteristics with a higher threshold are selected. The analysis that follows figure 2 was constructed using seven features that showed a strong correlation with the projected attribute diagnosis.

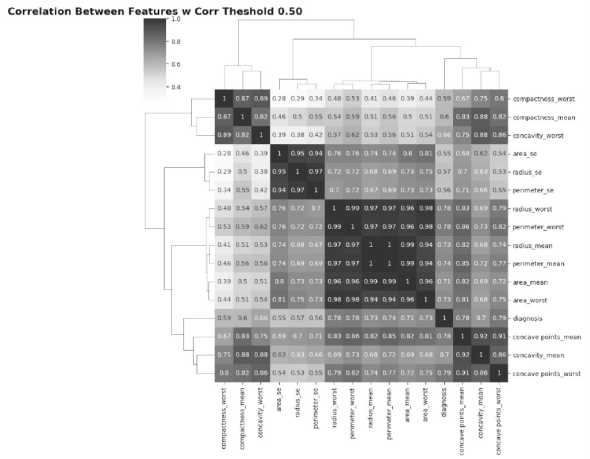

Figure 2: Highly Correlated Features

The data are split to avoid overfitting the model. Reduced dimension is transfer information into a space with fewer dimensions by eliminating unnecessary variance, which identifies the subspace in which the facts are located [24]. Some instances of dimensional reduction techniques are feature extracting and feature selection [25]. The process of extracting features from a dataset and removing redundant, superfluous, or less significant dimensions' information is known as feature extraction. Using feature selection, as much redundant and superfluous data as is practically possible can be identified and eliminated while still creating strong learning models. Consequently, feature selection lowers processing and computational expenses while simultaneously enhancing the model built using the selected data [26], [27].

The selection of features method has been used with healthcare data in several previous researches. [28]. The efficiency of such systems typically did not meet forecasts, even though certain earlier works dealing about the dataset utilized for this investigation are pertinent. The primary reason why certain systems perform poorly is because they are unable to identify the most important and closely related attributes. The dataset used for this investigation consisted of 80% training data and 20% test data. Two classifiers—SVM and Random Forest where chosen.

IV. CLASSIFIERS AT WORK

4.1 SVM Algorithm

The processes in the SVM classification approach are as follows: projecting the data given as input points into an N-dimensional vector space; choosing the right hyper-plane to maximize the variance between the two classes [29]. The selection of parameters such as the kernel itself C, and gamma has a major impact on the SVM's performance. A low-dimensional space is converted onto a multiple one that facilitates categorization via a function known as a kernel.

The nonlinearity is controlled by the kernel. The kernel's coefficients were changed for this experiment [30].

4.2 Random Forest

A Random Forest is a collection of different randomly functioning decision trees that operate independently of one another. The data is bootstrapped to create the trees. When a

random forest model is used to categories an input vector, the class with the highest number of scores generates the model's forecast.[31] A random forest is a single tree that transmits a single score based on the total amount of supplied prediction values. [32]. A binary tree is repeatedly divided into similar nodes to create Random Forest. Through inheritance, the parent node affects the similarity of the child node [33].

V. EVALUATING PERFORMANCE

The 2 by 2 confusion matrix has been used to evaluate the metrics efficiency and prediction model accuracy, as shown in Table 1. The percentage or opportunity of all correctly predicted events is represented by accuracy. [34].

Table 1: Measure of performance

Classification Algorithms	Performance Measures	IF	HOBSM
SVM	Accuracy	0.96	0.96
	Recall	0.96	0.97
	Precision	0.96	0.97
	F1 measure	0.96	0.97
Random Forest	Accuracy	0.97	0.976
	Recall	0.97	0.976
	Precision	0.97	0.976
	F1 measure	0.97	0.976

VI. RESULT, ANALYSIS, AND CONCLUSION

Different machine learning techniques were used to test the proposed methods. To assess each method and offer a performance statistic, a 2 x 2 confusion matrix was created. The suggested models were evaluated using "accuracy" as the performance metric.

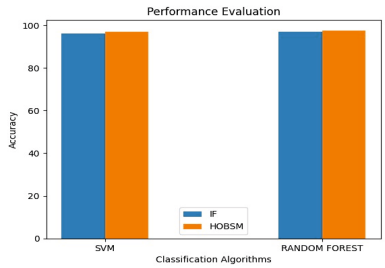

Figure 3: Performance evaluation of Isolation Forest and HOBSM method

Out of two classifier SVM, Random Forest. Random Forest has similar accuracy 97.6%.

VII. CONCLUSION

Research largely focuses on improving machine learning models to increase the accuracy of forecasting the future outcome of cancer of the breast disease. The results show that using different methods of classification in conjunction with HOBSM techniques for outlier detection may provide beneficial instruments for inference in this situation. To forecast on additional variables, classification systems need to perform better on different feature selection approaches.

REFERENCES

1. M. R. Bohemian, H. R. Marateb, M. Mansourian, M. A. Mananas & F. Mokarian, “A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning,” Computational and Structural Biotechnology Journal 15 (2017) 75.

2. S. Amin, H. S. Ewunonu, E. Oguntebi & I. Liman, “Breast cancer mor- tality in a resource-poor country: a 10-year experience in a tertiary institution,” Sahel Medical Journal 20 (2017) 9.

3. M.W. Huang, C.W. Chen, W.C. Lin, S.W. Ke & C.F. Tsai, “SVM and SVM ensembles in breast cancer prediction,” PLoS ONE 12 (2017) 161501.

4. CDC, “What is breast cancer?” (2021).

5. R. J. Oskouei, N. M. Kor & S. A. Maleki, “Data mining and medical world: breast cancers’ diagnosis, treatment, prognosis and challenges,” American Journal of Cancer Research 7 (2017) 610.

6. L. A. Aaltonen, R. Salovaara, P. Kristo, F. Canzian, A. Hemminki, P. Pel- toma¨ki, R. B. Chadwick, H.

Ka¨a¨ria¨inen, M. Eskelinen, H. Ja¨rvinen, J. P. Mecklin, & A. De la Chapelle, "Incidence of hereditary nonpolyposis col- orectal cancer and the feasibility of molecular screening for the disease," New England Journal of Medicine 338 (1998) 1481.

7. Khamparia, S. Bharati, P. Podder, D. Gupta, A. Khanna, T. K. Phung & D. N. H. Thanh, "Diagnosis of breast cancer based on modern mam- mography using hybrid transfer learning," Multidimensional Systems and Signal Processing 32 (2021) 747.

8. H. Kurihara, C. Shimizu, Y. Miyakita, M. Yoshida, A. Hamada, Y. Kanayama, K. Yonemori, J. Hashimoto, H. Tani, M. Kodaira, M. Yunokawa, H. Yamamoto, Y. Watanabe, Y. Fujiwara & K. Tamura, "Molecular imaging using PET for breast cancer," The Japanese Breast Cancer Society 23 (2016) 24.

9. T. Nagashima, M. Suzuki, H. Yagata, H. Hashimoto, T. Shishikura, N. Imanaka, T. Ueda & M. Miyazaki, "Dynamic-enhanced MRI predicts metastatic potential of invasive ductal breast cancer," Breast Cancer 9 (2002) 226.

10. C. S. Park, S. H. Kim, N. Y. Jung, J. J. Choi, B. J. Kang & H. S. Jung, "Interobserver variability of ultrasound elastography and the ultrasound BI-RADS lexicon of breast lesions," Breast Cancer 22 (2015) 153.

11. S. I. Ayon & M. Islam, "Diabetes prediction: a deep learning approach," International Journal of Information Engineering and Electronic Business 11 (2019) 2.

12. Z. Islam, M. Islam & A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (covid-19) using x-ray images," Informatics in Medicine Unlocked 20 (2020) 100412.

13. K. Hasan, M. Islam & M. M. A. Hashem, "Mathematical model devel- opment to detect breast cancer using multigene genetic programming," International Conference on Informatics, Electronics and Vision (2016) 574.

14. C. Shah & A. G. Jivani, "Comparison of data mining classification algo- rithms for breast cancer prediction," Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (2013) 1.

15. D. A. Omondiagbe, S. Veeramani & A. S. Sidhu, "Machine learning clas- sification techniques for breast cancer diagnosis," IOP Conference Series:Materials Science and Engineering 495 (2019) 012033.

16. Derangula, S. Edara & P. K. Karri, "Feature selection of breast cancer data using gradient boosting techniques of machine learning," Clinical Medicine 7 (2020) 17.

17. S. Raj, S. Singh, A. Kumar, S. Sarkar & C. Pradhan, "Feature selection and random forest classification for breast cancer disease," Data Analytics in Bioinformatics (2021) 191.

18. T. H. Cheng, C. P. Wei & V. S. Tseng, "Feature selection for medical data mining: comparisons of expert judgment and automatic approaches," 19th IEEE Symposium on Computer-Based Medical Systems (2006) 165.

19. M. M. Islam, Md. R. Haque, H. Iqbal, Md. M. Hasan, M. Hasan & M.N. Kabir, "Breast cancer prediction: A comparative study using machine learning techniques," SN Computer Science 1 (2020) 290.

20. N. Khuriwal & N. Mishra, "Breast cancer diagnosis using deep learning algorithm," International Conference on Advances in Computing, Com- munication Control and Networking (2018) 98.

21. F. A. Muhammet, “A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications,” Healthcare 8 (2020) 111.

22. N. F. Idris & M. A. Ismail, “Breast cancer disease classification us- ing Fuzzy-ID3 algorithm with FUZZYDBD method: automatic fuzzy database definition,” PeerJ Computer Science 7 (2021) 427.

23. R. Harikumar & C. Sannasi, “Effective classification framework for breast tumors using optimized multi-kernel SVM with controlled skewness,” In- ternational Journal of Aquatic Science 12 (2021) 1604.

24. S. N. Ghazavi & T. W. Liao, “Medical data mining by fuzzy modeling with selected features,” Artificial Intelligence in Medicine 43 (2008) 195.

25. S. M. Vieira, J. M. C. Sousa & U. Kaymak, “Fuzzy criteria for feature selection,” Fuzzy Sets and Systems 189189 (2012)1.

26. S. B. Sakri, N. B. Abdul Rashid & Z. Muhammad Zain, “Particle swarm optimization feature selection for breast cancer recurrence prediction,” IEEE Access 6 (2018) 29637.

27. E. E. Bron, M. Smits, W. J. Niessen & S. Klein, “Feature selection based on the SVM weight vector for classification of dementia,” IEEE Journal of Biomedical and Health Informatics 19 (2015) 1617.

28. W. Yue, Z. Wang, H. Chen, A. Payne, and X. Liu, ‘Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis’, Designs, vol. 2, no. 2, p. 13, May 2018.

29. M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, ‘Breast cancer classification using machine learning’, in 2018 Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT), Istanbul, 2018, pp. 1–4.

30. Abreu, Pedro Henrique’s, et al. "Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review." ACM Computing Surveys (CSUR) 49.3 (2016): 52.

31. Y. Freund, R. Schapiro, ‘A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting’, 1995.

32. B. Dai, R.-C. Chen, S.-Z. Zhu and W.-W. Zhang, ‘Using Random Forest Algorithm for Breast Cancer Diagnosis’, in 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 449–452.

33. B. Dai, R.-C. Chen, S.-Z. Zhu and W.-W. Zhang, ‘Using Random Forest Algorithm for Breast Cancer Diagnosis’, in 2018 International Symposium on Computer, Consumer and Control (IS3C), Taichung, Taiwan, 2018, pp. 449–452.

34. M. Kumari, V. Singh & P. Ahlawat, “Automated decision support system for breast cancer prediction,” International Journal on Emerging Tech- nologies 11 (2020) 193.