**Predictive Models for Breast Cancer Diagnosis using Machine Learning Algorithms**

**Abstract**

Among American women, breast cancer has the highest mortality rate. Predictive approaches based on machine learning offer methods for diagnosing breast cancer sooner. Making an assessment for models that accurately identify cancer is still difficult, though. In order to increase the accuracy of the diagnosis of breast cancer, we created four distinct prediction models and offered DET (Data Exploratory Techniques). In order to find the robust feature classification into malignant and kind classes before models, 4-layered essential DET, such as 1). feature distribution, 2). correlation, 3). Removal 4). hyperparameter optimization, remained thoroughly investigated. On the WDBC (Wisconsin Diagnostic Breast Cancer) datasets, these suggested approaches and classifiers remained put into practice. The effectiveness and training duration of each classifier were evaluated using common performance indicators, such as confusion matrices and K-fold cross-validation methods. With DET, which included Support Vector Machine, Linear Regression, K-Nearest Neighbor with the WDBC dataset, the models' diagnostic ability increased. We also assessed the accuracy of our significant findings in comparison to earlier research.

**Problem Statement**

The performance of Breast Cancer model execution time increases rapidly with the number of features. The prediction model becomes slower, affecting the diagnosis accuracy. In contrast, the model’s accuracy and time complexity are critical issues for the data analyst and physician. Recently, in Nature Cancer, a study presented an approach to classify cancer into normal and tumor tissues. Meanwhile, many studies have utilized the SVM classifier for breast cancer prediction, while a few of them used only one classifier in experiments. However, there is still a demand to explore the efficient classifier for breast cancer prediction with more effective methods such as SVM, LR, KNN and EC. This study performed four different prediction models with sufficient data mining exploratory techniques to diagnose breast cancer.

The presentation of the BC model's execution time increases quickly with more features. The accuracy of the diagnosis is impacted by the prediction model's slowdown. For the data analyst and doctor, however, the model's accuracy as well as time complexity are crucial concerns. A study published recently in Nature Cancer described a method for dividing cancerous tissues into normal and tumorous components. Meanwhile, numerous research has applied the SVM classifier for BC prediction, although a few of them used only one classifier in tests. Nonetheless, there is still a need to investigate the effective classifier for predicting BC using more efficient techniques as SVM, LR, KNN, and EC. This study tested 4-distinct prediction models with sufficient DME methodologies to diagnose BC.

**Introduction**

Breast cancer (BC) is the world’s leading cause of death in women after lung cancer, with approximately 2,261,419 new cases and 684,996 new deaths in 2020 (Sung, H., et. al., 2021). In the United States, 281,550 new cases were diagnosed with breast cancer, and 43,600 deaths were reported in the females during 2021 (Siegel, R.L., et al., 2022). Breast cancer is a type of cancer that originates from breast tissue, most generally from the internal layer of the milk conduit or the lobules that provide milk to the milk conduit. Cancer cells arise from natural cells due to modification or mutation of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). These modifications or mutations may occur spontaneously as a result of the increase in entropy, or they may be triggered by other factors. For example, electromagnetic radiation (X-rays, microwaves, ultraviolet-rays, gamma-rays, et cetera), nuclear radiation, bacteria, viruses, fungi, parasites, chemicals in the air, heat, food, water, free radicals, mechanical cell-level injury, evolution, and aging of DNA and RNA (Leão, et al., 2021). In general, benign and malignant are two classes of tumors. Although benign is not life-threatening and cancerous, it may boost the chances of breast cancer risk. In contrast, malignant is more alarming and cancerous tumors. A study performed breast cancer detection and reported 20% of women died due to malignant tumors (Subashini, T.S., et al., 2009).

These studies emphasize the diagnosis of tumors, and recently, it is a trending biomedical issue. The researchers are employing data mining (DM) and machine learning (ML) technologies for breast cancer prediction (Abdar, M., et al., 2020). Classifier-based prediction models on DM and ML can limit the diagnosis errors and enhance the efficiency of a cancer diagnosis. DM is an extensive combination of different approaches to discover hidden knowledge and information from large-scale datasets that are difficult to analyze directly. It has been broadly used in the implementation of the prediction system for various diseases, such as heart disease (Rasool, A., et al., 2020), lung cancer (McWilliam, A., et al., 2016), and thyroid cancer (Park, K.H., et al., 2021). DM and ML techniques have been embedded for diagnosing breast cancer with computer-aided systems (Park, E.Y., et al., 2021), and fuzzy genetics (Bicchierai, G., et al., 2021). The results of these studies successfully classify the features into two types of tumors by the evaluation of classifier and predicting the incoming tumor based on previous data.

In the literature, a research study proved that breast cancer prediction with machine learning classifiers in the early phases does not just increase the survival chances but can control the diffusion of cancerous cells in the body (Akay, M.F., 2009). For instance, a study used the support vector machine (SVM) based method for breast cancer diagnosis and achieved practical results in prediction (Furey, T.S., et al., 2000). Similarly, Furey et al. (2015) also employed SVM for cancer tissue classification with a linear kernel and attained a 93.4% accuracy. Later, this work was extended by Zheng et al. (2014) by delivering a K-SVM hybrid model for WDBC dataset classification and acquiring 97% accuracy (Seddik, A.F.; Shawky, D.M., 2015). Meanwhile, some other researchers worked on different classifiers, such as Seddik et al. (2015), who proposed a method based on tumor variables for a binary logistic model to diagnose breast cancer WDBC data and secure good results (Mert, A., et al., 2015). Likewise, Mert et al. used a k-nearest neighbor (KNN) classifier to predict breast cancer by designing a feature reduction method with independent component analysis. It distributed the features with reduced one feature (1C) and 30 features and computed the performance, and attained 91% accuracy (Abdar, M., et al., 2018).

Apart from these advantageous accuracies with different classifiers and methods, these studies mentioned above have not considered the data exploratory techniques, which enable the data mining techniques to be more robust to acquire efficient performance. Due to the absence of such essential techniques, various studies (Rajaguru, H., 2019) face the accuracy limitation of ML classifiers. Meanwhile, the confusion matrices misdiagnosed the malignant and benign classes in those studies due to the incorrect prediction of true negative and false negative matrices. Another defect was found in those previous studies that used criteria to assess the feature training with nonlinear classification. However, the performance of model execution time increases rapidly with the number of features (Brause, R.W., 2001). As a result, the prediction model becomes slower, affecting the diagnosis accuracy. In contrast, the model’s accuracy and time complexity are critical issues for the data analyst and physician. These problems, as mentioned above, and findings motivated us to pursue a new study for breast cancer diagnosis by proposing DM techniques with different ML models.   
In this research, four different prediction models were formulated with four machine learning algorithms (SVM, KNN, logistic regression (LR), and ensemble classifier (EC)) to deal with a massive volume of tumor features for the extraction of essential information for the diagnosis of breast cancer. The objective was to explore an accurate and efficient prediction model for tumor classification by using data mining techniques. It proposes four-layered significant data exploratory techniques (DET), including feature distribution, elimination, and constructing a hyperparameter for the practical analysis of Wisconsin Diagnostic Breast Cancer (WDBC) and Breast Cancer Coimbra Dataset (BCCD). These techniques enabled the machine learning predictive models to improve accuracy and enhance diagnostic efficiency. In the absence of these techniques, we observed some literature suffers from accuracy limitations. Although image data are more reasonable for breast cancer detection, we have not considered them in this work due to the targeted WDBC and BCCD datasets to apply the intelligent ML classifiers. It presents a framework by integrating DET and predictive models to explore the implementation method for breast cancer diagnosis. The tumor features can be presented in many details, which produces redundant information. Such features lead to tedious outcomes due to high computation times. As a result, our fundamental goal was not only to investigate the effective predictive model with attainable accuracy but also one with time complexity for the cancer diagnosis. The deliberation of time efficiency will enable our models to extract and mine vital information from a vast dataset by finding correlations and eliminating the features. The results presented satisfactory accuracy for the breast cancer diagnosis with the lowest computation time, which signifies the quality of our study as compared to others. This work will enable a data analyst to apply an intelligent machine learning model to analyze breast cancer data. Likewise, a physician would diagnose breast cancer precisely by the tumor classification.

As summarized, the following are the significant contributions of this study: (1). We investigated four prediction models (SVM, LR, KNN, and EC) with the WDBC and BCCD breast cancer datasets, which reached the next level of quality by diagnosing the tumor and classifying it into benign and malignant. (2). It proposes four-layered data exploratory techniques before implementing four ML classifiers as prediction models. These techniques enable the predictive models to acquire peak accuracies for breast cancer diagnosis. (3). We set up experiments to validate the models’ prediction and classification accuracy regarding time complexity and deliver comparative analysis with state-of-the-art studies and various evaluation matrices.

**Related Work**

Breast cancer disease causes a massive number of deaths in the world. After the traditional cancer detection methods, the latest technologies enable experts with numerous adaptive methods to discover breast cancer in women. Along with the new technologies, various data science (DS) techniques assist in cancer-based data collection and evaluation to predict this deadly disease. Machine learning algorithms have been successfully applied to cancer-based data analysis among these DS technologies. For example, research (Huang, C.L., et al., 2008) was conducted to prove that these machine learning algorithms can improve diagnostic accuracy. It turns out that a 79.97% diagnostic accuracy was achieved by an expert physician. However, 91.1% correct predictions were attained with machine learning.

In the last couple of decades, ML applications in the medical field have gradually increased. However, the data collected from the patients and evaluation by the medical expert are the essential factors for diagnosis. The machine learning classifiers have aided in minimizing human errors and delivered prompt analysis of medical data with greater depth (Polat, K.; Güne¸s, S., 2007). There are several machine learning classifiers for data modeling and prediction; in our work, we employed SVM, LR, KNN, and EC for breast cancer prediction.

SVM was a widely implemented machine learning algorithm in the diagnosis domain of breast cancer due to its highest prediction accuracy. For instance, Furey et al. (2000) presented SVM with a linear kernel for cancer tissue diagnosis and reached acceptable accuracy (Zheng, B., et al., 2014). Similarly, Polat et al. (2007) used the least square SVM for breast cancer prediction to eliminate redundant features and secured a 98.53% accuracy. It was suggested that least square SVM assisted in model training with linear equations (Prasad, Y., et al., 2010). However, his method did not deliver the feature selection process. The author (Muzammal, M.; Qu, Q.; Nasrulin, B., 2019) delivered a distributed database for multi-active features to integrate different technologies. In Prasad and Jain et al. (2010) proposed a heuristic model for feature subset to train the SVM classifier. It classifies the breast cancer data into two different classes with 91.7% accuracy. However, this accuracy can be adequately improved if the author employs the feature eradication method to get rid of the noise data.

Similarly, Zheng et al. (2014) proposed a hybrid model combining K-mean and SVM classifiers. This model objective was to diagnose the tumor features from the WDBC dataset by employing the feature selection and extraction method. A K-mean classifier was employed to identify the benign and malignant tumor patterns. The generated patterns are computed and considered as new patterns for the training of the SVM model. Then, SVM is executed for the prediction of incoming tumors. The employment of their hybrid model improved the accuracy to 97%. However, the data exploratory techniques are the fundamental tasks for the data preparation, which have not been adequately addressed to train the proposed model (Seddik, A.F.; Shawky, D.M., 2015).

Apart from the SVM, Lim and Sohn et al. (2013) performed logistic regression (LR) with optimal parameters on the Wisconsin Original Breast Cancer (WOBC) and WDBC datasets. It achieved 97.8% sufficient accuracy for the WOBC dataset and 93.8% accuracy for the WDBC dataset with optimized feature sets (Yang, L.; Xu, Z., 2019). Similarly, Seddik et al. (2015) presented a binary logistic model for the diagnosis of breast cancer data based on variables with tumor image characteristics. The proposed model classifies the WDBC data into malignant and benign and accomplished the 98% average classification accuracy. This regression model found that area, texture, concavity, and symmetry are significant WDBC features (Mert, A., et al., 2015).

A. Mert et al. (2015) delivered a feature reduction method with independent component analysis to predict breast cancer. It utilized the k-nearest neighbor (KNN) classifier to categorize the WDBC features efficiently with a reduced one feature (1C) and 30 features. It computed the performance with different matrices and attained 91% accuracy (Abdar, M., et al., 2018). Later, this study was further improved by Rajaguru et al. (2019), who tackled the breast cancer prediction challenge by implementing the KNN and decision tree (DT) machine learning algorithms to classify the WDBC features. It used a traditional principal component analysis (PCA) feature selection method for the feature categorization and found that KNN outperformed the DT (Mushtaq, Z., et al., 2020). In another study conducted by Yang and Xu et al. (2019), KNN achieved 96.4% accuracy with the same feature selection method (PCA) (Fu, Y., et al., 2020). Recently, work has involved considering KNN efficiency by the k values and many distance functions of KNN to find its effectiveness with two different breast cancer datasets. It involves the three different types of the experiment: KNN without feature selection, with linear SVM, and with Chi-square-based features. It indicated that the third technique, Chi-square-based feature selection, succeeded in accomplishing the highest accuracy on both datasets with Manhattan or Canberra distance functions (Kamyab, M., et al., 2018).

As for the fourth prediction model, named ensemble classifier (EC) with the voting technique, few studies consider this approach for breast cancer prediction. For instance, M. Abdar et al. (2020) proposed an ensemble method by vote/voting classifier to detect benign tumors from malignant breast cancer. It established a two-layer voting classifier for two or three different machine learning algorithms. The results of these voting techniques disclosed the adequate performance of the simple classification algorithm (Abdar, M., 2020). From these studies, we got the motivation to conduct experiments based on voting classifiers with different machine learning techniques. However, none of the above approaches has utilized the feature correlation and elimination for the given breast cancer dataset to the best of our knowledge. These studies conducted experiments to classify the cancer features, which is still a challenging issue. Recently, in Nature Cancer, a study presented an approach to classify cancer into normal and tumor tissues [28]. Meanwhile, many studies have utilized the SVM classifier for breast cancer prediction, while a few of them used only one classifier in experiments. However, there is still a demand to explore the efficient classifier for breast cancer prediction with more effective methods (Abdar, M., 2020) (Seddik, A.F., et al., 2015) (Mert, A., et al., 2015) (Mushtaq, Z., et al., 2020). This study performed four different prediction models with sufficient data mining exploratory techniques to diagnose breast cancer.

**Proposed Solution**

A solution is suggested in Figure 1 to address these issues. There are several important steps in this solution. These are the broad strokes of this methodology:

1. Datasets for the WDBC are downloaded from the repository for ML.
2. Carry out the core data preparation operations.
3. Sort the data in WDBC into benign and malignant categories.
4. By determining their correlation, distribute the attributes into the three categories of positive, negative, and random.
5. To get the best results, find less important characteristics and then remove such repetitive features.
6. Divide the dataset into training and testing datasets after EDA.
7. The application of the datasets to different predictive models such as SVM, LR, KNN, and EC.
8. Following the execution of the models, the classifier's prediction is made using various matrices, such as confusion matrices, to gauge the performance of the models.
9. Review the findings and compare the precision of each model to earlier research.

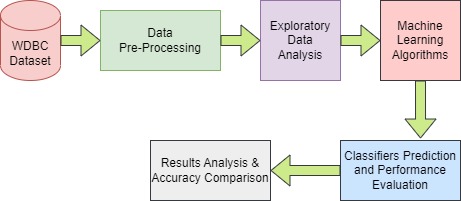


Fig 1: Proposed diagram breast cancer prediction with DET and ML classifiers

**Expected results and Evaluation**

The F1 score was primarily considered in our outcomes evaluations while analyzing our data. Largely unbalanced class distributions occurred in datasets, much as they do in classification tasks in the real world. The feature distribution results reveal some observations that significantly differ between the classes. For instance, the distribution of benign and malignant classes differed significantly in the Supplementary Note 01 concavity mean. To balance these traits, the resampling techniques of oversampling, under sampling, and cross-validation were used. The oversampling method duplicates the minority classes; however it causes a problem for ML called overfitting. The under-sampling strategy, in contrast, eliminates the majority classes that exclude potential data.

However, the cross-validation technique was suggested as a dominant technique for overcoming the imbalanced class distribution. Cross-validation tests and trains a model by using different parts of the data. To balance the benign and malignant features in the training and testing datasets. Because of the efficient use of critical values of TP, TN, FP, and FN to deal with actual and predicted classes, the cross-validation matrices, including F1 score, precision, and recall, were compared. With these proposed prediction models and techniques, we can efficiently assist the cancer domain in obtaining highly satisfying results for BC diagnosis.