International Journal of Education and Science Research Review

E-ISSN 2348-6457

Volume-5, Issue-6 Nov-Dec- 2018

P-ISSN 2349-1817

www.ijesrr.org

Email- editor@ijesrr.org

MACHINE LEARNING TECHNIQUES FOR BREAST CANCER DETECTION

Jeetendra Kumar

Atal Bihari Vajpayee University, Bilaspur, Chhattisgarh, India

Abstract: Early detection of breast cancer is crucial as it significantly increases the chances of successful treatment and survival rates. This study gives a complete investigation of the categorization of breast cancer using three different classifiers. These classifiers include the Naive Bayes classifier, the J-48 classifier (An algorithm for a decision tree), and the Decision Stump classifier. The research makes use of the UCI Breast Cancer Wisconsin (Diagnostic) Data collection, which has been utilized by a large number of researchers previously. The objective of the study is to categorize tumours as either malignant or benign based on a collection of features collected from fine needle aspirate (FNA) pictures. The performance of these classifiers is analysed in this research in terms of accuracy, sensitivity, specificity, precision, F-measure, and error rate. According to the findings, J-48 performed better than the other classifiers on a consistent basis across a variety of evaluation parameters, which demonstrates that it is a promising option for breast cancer classification tasks.

Keywords: Cancer, Naïve Bayes, J-48, tumour, Decision Stamp.

1. Introduction

Cancer of the breast is the most prevalent form of the disease found in Indian females. It is estimated that over 14.5 lakh women are diagnosed with breast cancer each year, making it one of the most common causes of death in the country due to cancer. The diagnosis of breast cancer at an early stage is one of the most critical parts of the treatment process [1]. In addition to having frequent mammograms and clinical breast exams, regular self-examinations of the breast can be an effective way to diagnose breast cancer in its earlier stages, which in turn improves the likelihood that treatment will be successful. Cancer of the breast, often known as BC, is the most frequent form of cancer among women worldwide [2]. The term "breast cancer trends" refers to an in-depth investigation of the patterns and shifts associated with breast cancer over the course of time. These trends provide useful insights into the impact of the disease, and they are essential for evaluating the efficacy of preventative efforts and treatment methods, as well as indicating areas that require greater attention and resources. The incidence of breast cancer is a key trend that tracks key trends, including whether the number of new cases of breast cancer is increasing, decreasing, or keeping stable. Survival trends are a way of measuring how well people who have been diagnosed with breast cancer fare over lengthy periods of time. These patterns frequently indicate improvements in the early identification and treatment effectiveness of breast cancer. Changes in mortality trends demonstrate a shift in the number of fatalities caused by breast cancer; a decrease in mortality rates indicates that there have been breakthroughs in therapy. In addition, developments in screening and diagnosis, variables that put patients at risk, treatment modalities, preventative initiatives, and health inequities all contribute to our comprehension of this complicated disease. It is possible for healthcare professionals and policymakers to better adjust their methods to address the changing landscape of breast cancer and improve the outcomes for those who are impacted by this disease if they track these trends and respond to them accordingly.

Machine learning techniques have shown tremendous promise in the classification of breast cancer, aiding in early detection and personalized treatment strategies. Leveraging various data sources, such as medical images, patient records, and genomic data, machine learning models are capable of accurately categorizing breast cancer cases [3]. The process typically involves data preprocessing, feature selection, model choice, training, and rigorous evaluation. Models are often optimized through hyper-parameter tuning and cross-validation to ensure robust performance. Interpretability is a key consideration in healthcare applications, allowing medical professionals to understand and trust the model's predictions. Furthermore, deploying such models in clinical settings requires strict adherence to data privacy regulations, like HIPAA, and ongoing monitoring and updating to maintain their accuracy and relevance.

Collaboration between data scientists and healthcare experts is essential for the development of ethical, effective, and safe machine learning solutions for breast cancer classification.

2. Review of literature

A lot of research has been done to detect breast cancer using machine learning techniques.

In 2015, a research article by Karabatak [4] proposed a novel classifier for breast cancer detection is introduced, utilizing the Naïve Bayesian method. The study explored the effectiveness of this statistical approach in accurately identifying breast cancer cases. By leveraging the probabilistic nature of the Naïve Bayesian classifier, the study aims to improve diagnostic accuracy and offer a robust tool for early detection. The results indicated that the proposed classifier performed well, highlighting its potential application in medical diagnostics and its contribution to improving breast cancer detection methodologies. In 2016, Alarabeyyat et al. [5] proposed the use of the k-nearest neighbour (k-NN) machine learning algorithm for breast cancer detection. The research investigated the efficacy of the k-NN algorithm in accurately classifying and detecting breast cancer. By analyzing a dataset of medical records, the authors demonstrated that the k-NN algorithm can be an effective tool for breast cancer diagnosis, offering promising results in terms of accuracy and reliability. The study underscored the potential of machine learning techniques in enhancing medical diagnostics and improving early detection of breast cancer. In 2016, Montazeri et al. [6] investigated various machine-learning models to predict breast cancer survival. The study evaluated machine learning algorithms to identify their effectiveness in forecasting patient outcomes. By analysing a dataset of breast cancer cases, the authors demonstrated that machine learning models can provide accurate survival predictions, potentially aiding in personalized treatment planning and improving patient prognoses.

In 2017, Trister et al. [7] discussed the potential of machine learning to revolutionize breast cancer screening by enhancing the accuracy and efficiency of mammography. They highlighted that machine learning algorithms, trained on large datasets, can identify subtle patterns in mammographic images, potentially reducing false positives and false negatives. The authors emphasized the need for high-quality data, interdisciplinary collaboration, and rigorous validation before clinical implementation. They also addressed ethical and regulatory concerns, advocating for machine learning to augment radiologists' capabilities rather than replace them, ultimately aiming to improve patient outcomes through more accurate and efficient diagnostics. In 2018, Islam et al.[8] investigated the effectiveness of Support Vector Machine (SVM) and K-Nearest Neighbours (K-NN) algorithms in improving breast cancer diagnosis. The study used a publicly available breast cancer dataset, applying feature selection and data pre-processing techniques. The results showed that both SVM and K-NN achieve high accuracy in predicting breast cancer, with SVM slightly outperforming K-NN. The research highlighted the potential of these machine learning models to assist clinicians in making more informed diagnostic decisions, ultimately aiming to enhance patient outcomes through more accurate and efficient breast cancer detection. In 2018, Tahmooresi et al. [9] explored the application of various machine-learning algorithms for the early detection of breast cancer. The study evaluated the performance of different models, including Support Vector Machine (SVM), Decision Trees, and Neural Networks, using a dataset of mammographic images and patient records. The findings demonstrated that these machine learning techniques significantly improve the accuracy and efficiency of breast cancer diagnosis, with SVM and Neural Networks showing particularly promising results. The research underscored the potential of advanced computational methods in early cancer detection, which is crucial for improving patient outcomes and reducing mortality rates.

3. Methodology

For classification, we have used UCI dataset of breast cancer [10]. This dataset contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The goal is to classify tumors as either malignant (cancerous) or benign (non-cancerous). Features include various measurements of cell nuclei, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

Data Set Characteristics	-	Multivariate
Attribute Characteristics	-	Categorical
Instances	-	286
Total no. of Attributes	-	10
Missing Attribute	-	No
Noisy Attribute	-	No

For the classification of the breast cancer dataset, we have used naive bayes classifier, J-48 classifier and Decision stump classifiers.

Naïve Bayes Classifier- The Naive Bayes Classifier is a Probabilistic Machine Learning Algorithm Based on Bayes' Theorem The Naive Bayes Classifier is a probabilistic machine learning algorithm that is based on Bayes' theorem. It is referred to as "naive" since it simplifies things by assuming that all characteristics are conditionally independent given the class designation. This means that it thinks the presence or absence of one characteristic is unrelated to the presence or absence of any other characteristic. In spite of this reduction, the Naive Bayes algorithm is frequently employed for a variety of classification applications, particularly in natural language processing and text categorization. It is computationally efficient and functions well when the independence assumption is reasonably true in the dataset. This is the case when it works well.

J-48 (C4.5) Classifier- J-48, commonly referred to as C4.5, is an algorithm for classifying data based on decision trees that was invented by Ross Quinlan. When it comes to classification activities, decision trees are frequently used. The J-48 algorithm constructs a decision tree by successively subdividing the dataset according to the features that contain the most relevant information. It does this by choosing the attribute that will partition the data most effectively at each node, taking into account factors such as information gain and Gini impurity. Because decision trees may be interpreted and because they are able to process both categorical and continuous data, they are an adaptable option that can be utilized in a broad variety of contexts. In addition to this, they are able to deal with missing values and are frequently utilized in ensemble approaches.

Decision Stump Classifier- A decision stump, also known as a decision tree classifier, is the most fundamental type of a decision tree. A single decision node and two leaf nodes make up its constituent parts. This indicates that it is able to make decisions that are either binary or based on a single feature. The requirement for a basic binary choice frequently calls for the use of decision stumps, which are notable for their ease of use. They are frequently used in the capacity of weak learners in ensemble approaches such as AdaBoost, which involve combining a large number of decision stumps in order to produce a more accurate classifier. Although decision stumps are restricted in their capacity to depict intricate relationships, they can be useful for undertakings in which simplicity is prized or as constituent parts of more intricate models.

We have performed the experiment with 10 fold cross validation. In 10-fold cross-validation, your dataset is first divided into ten equal-sized parts. Then, each part of the dataset is set aside one at a time for testing, while the remaining nine parts of the dataset are utilized for training. The procedure produces ten individual evaluation outcomes, and the ultimate evaluation is produced by computing the average of these ten evaluation results. This method ensures a robust and trustworthy estimation of the model's performance, reducing the risk of over fitting while also offering a full assessment of the model's capacity to generalize to new data that has not been seen before.

4. Results

After performing classification with all three types of classifiers, following results have been obtained.

Confusion Matrix- The performance of a classification model can be evaluated with the help of a fundamental tool known as a confusion matrix, which is utilized in the fields of machine learning and statistics. It is a concise assessment of how accurately a model has classified various cases taken from a dataset. Confusion matrices are most helpful when working on binary classification problems; however, they can also be modified for use in multi-class classification. A square matrix with four entries and the following arrangement is an example of a typical confusion matrix. The number of instances that were accurately predicted as positive or as belonging to the positive class is referred to as the "true positives" (TP). True Negatives, also known as TNs, are instances that were accurately predicted to be negative or to belong to the negative class. False Positives (FP) are instances that were anticipated as positive but are, in fact, in the negative class. False positives are also abbreviated as FP. A Type I error is another name for this particular problem. False negatives, sometimes known as FN for short, are incidents that were forecasted to fall into the negative class but instead belong to the positive class. A Type II error is another name for this kind of mistake.

Naïve Bayes classifier

no-recurrence-events	recurrence-events	Classified as
168	33	No-recurrence-events
48	37	recurrence-events

J48 classifier

no-recurrence-events	recurrence-events	Classified as
193	8	No-recurrence-events
62	23	recurrence-events

Decision stump classifier

no-recurrence-events	recurrence-events	Classified as
160	41	no-recurrence-events
49	36	recurrence-events

Table 1:- Performance comparison of different classifiers

Classifier	Accuracy	Sensitvity	Specificity	Precision	F-measure	Error
Naïve Baye's Classifier	0.717	-.446	0.704	0.717	0.708	0.2888
J48 Classifier	0.7555	0.524	0.752	0.755	0.713	0.339
Decision Stump Classifier	0.685	0.466	0.677	0.685	0.681	0.226

When we compare the performance of the Naive Bayes classifier, the J48 classifier, and the Decision Stump classifier, we gain several important insights. In the first place, J48 is in first place when it comes to accuracy. It has an accuracy score of 0.7555, which indicates that it properly classified around 75.6% of the occurrences. This exceeds both the Naive Bayes classifier and the Decision Stump classifier, both of which achieved an accuracy of 0.685. The Naive Bayes classifier earned an accuracy of 0.717. J48 demonstrated a recall of 0.524, which indicates that it properly detected approximately 52.4% of genuine positive instances. This brings us to the second consideration, which is sensitivity, also known as recall. This was a significant improvement above both the Naive Bayes classifier's unusually negative sensitivity of -0.446 and the Decision Stump classifier's recall of 0.466. The latter indicates that the Naive Bayes classifier can have problems identifying true positive examples of the phenomenon being studied. When it comes to specificity, J48 fared exceptionally well, scoring 0.752, which was followed by the Naive Bayes classifier, which scored 0.704, and the Decision Stump classifier, which scored 0.677. These numbers represent how well the classifiers can distinguish between false positives and true negatives. J48 displayed a precision of 0.755, which indicates a good positive predictive value. This was measured in terms of precision. The Naive Bayes classifier came in a close second with a precision of 0.717, while the Decision Stump classifier also performed quite well, achieving a precision of 0.685. The F-measure, which evaluates a classifier based on its ability to maintain a balance between accurately detecting positive examples and limiting false positives, found that the J48 classifier had the best score possible at 0.713. This indicates that it was able to achieve the maximum level of precision possible. The F-measure for the Naive Bayes classifier was 0.708, whereas the F-measure for the Decision Stump classifier was 0.681. The Naive Bayes classifier came in second place. In the end, while looking at the error rate, the Decision Stump classifier had the lowest error rate, which indicated that it had the lowest fraction of cases that were misclassified. This was determined by its error rate of 0.226. The Naive Bayes classifier, on the other hand, has the greatest error rate of all of them, which indicates that there is potential for improvement.

5. Conclusion

Using the UCI Breast Cancer Wisconsin (Diagnostic) Data Set, we conducted this research study to investigate the efficacy of three different classifiers in the context of breast cancer classification: the Naive Bayes classifier, the J-48 classifier, and the Decision Stump classifier. According to the findings of our investigation, J-48 came out on top as the classifier with the best overall performance of the three. When contrasted with the Naive Bayes and Decision Stump classifiers, it displayed much higher levels of accuracy, sensitivity, specificity, precision, and F-measure. Because it is able to strike a healthy balance between accurately identifying malignant tumors and reducing the number of false positives, J-48 is a very tempting option for this particular endeavor. Based on these findings, it appears that decision tree-based models, such as the J-48, may be particularly well-suited for the diagnosis of breast cancer. This research provides vital insights that

References

[1] Viale, G. (2012). The current state of breast cancer classification. Annals of oncology, 23, x207-x210.

[2] Dubey, A. K., Gupta, U., & Jain, S. (2015). Breast cancer statistics and prediction methodology: a systematic review and analysis. Asian Pacific journal of cancer prevention, 16(10), 4237-4245.

[3] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning, in 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT) (2018).

[4] Karabatak, M. (2015). A new classifier for breast cancer detection based on Naïve Bayesian. Measurement, 72, 32-36.

[5] Alarabeyyat, A., & Alhanahnah, M. (2016, August). Breast cancer detection using k-nearest neighbor machine learning algorithm. In 2016 9th International Conference on Developments in eSystems Engineering (DeSE) (pp. 35-39). IEEE.

[6] Montazeri, M., Montazeri, M., Montazeri, M., & Beigzadeh, A. (2016). Machine learning models in breast cancer survival prediction. Technology and Health Care, 24(1), 31-42.

[7] Trister, A. D., Buist, D. S., & Lee, C. I. (2017). Will machine learning tip the balance in breast cancer screening?. JAMA oncology, 3(11), 1463-1464.

[8] Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017, December). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In 2017 IEEE region 10 humanitarian technology conference (R10-HTC) (pp. 226-229). IEEE.

[9] Tahmooresi, M., Afshar, A., Rad, B. B., Nowshath, K. B., & Bamiah, M. A. (2018). Early detection of breast cancer using machine learning techniques. Journal of Telecommunication, Electronic and Computer Engineering (JTEC), 10(3-2), 21-27

[10] Zwitter,Matjaz and Soklic,Milan. (1988). Breast Cancer. UCI Machine Learning Repository. https://doi.org/10.24432/C51P4M.