WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Breast Cancer Detection Using Machine Learning Models

Deepak Pal[1], Dr. Tejna Khosla[2]

Department of Information Technology Maharaja Agrasen Institute of Technology Delhi, India
fulsinghdeep00@gmail.com
Tejnakhosla@mait.ac.in

ABSTRACT:

The goal of this study is to help diagnose breast cancer early by using machine learning. It is predicated on three models: Random Forest, Decision Tree, and Logistic Regression. These models assist in identifying if a tumor is malignant (cancerous) or capable of inhibiting it (non-cancerous). The Wisconsin Breast Cancer dataset, which comprises crucial characteristics including tumor size and texture as well as other vulnerable medical parameters, is used. The latter made it possible to eliminate the distractions and pinpoint the crucial characteristics for improved performance. The next stage is to evaluate these models using a variety of criteria, such as precision and accuracy. The most successful model of all, Random Forest, performs best in terms of being the most accurate and not the most overfitting. The actual project evidence must be a source of information about how effective machine learning techniques can be in the medical field for an unexpected off-track team in addition to routine early diagnosis and treatment planning. In addition, the project enables doctors to make an objective decision, find the cancer, and prescribe more treatment.

Keywords : Breast Cancer Detection, Tumor Classification, Machine Learning, Cell Analysis

1. Introduction

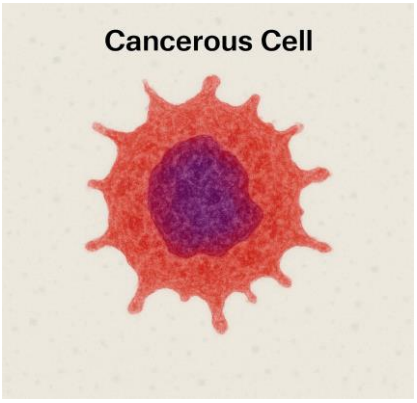
Breast cancer is a serious issue. It is among the most prevalent and fatal illnesses affecting women globally. Many lives are lost to it every year. since of this, early and precise detection of breast cancer is crucial since it can significantly impact the effectiveness of treatments and the number of survivors.

Here's the problem, though: conventional methods of diagnosis, such as biopsies and those manual examinations under a microscope? They can be time-consuming, expensive, and, let's be honest, not infallible. Human error can occasionally occur and completely alter a situation.

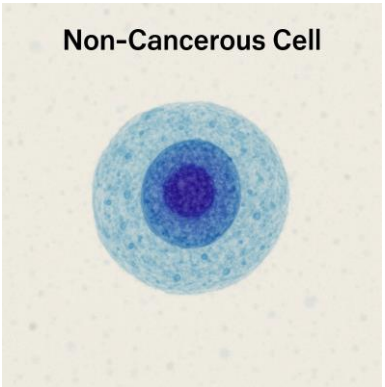
As technology advances, machine learning has become more prominent. In terms of medical diagnostics, it's revolutionizing the field by accelerating processes and improving prediction accuracy.

We're delving into a machine learning model for breast cancer detection in this project. This model examines and classifies tiny cell pictures to determine which

Some are malignant, whereas others are not. Using Python and tools like Scikit-learn and Pandas, we will address topics like data preprocessing, identifying key characteristics, and finally implementing some classification algorithms.



(Fig. I) Cancerous Cell



(Fig. II) Non-Cancerous Cell

2. Literature Review

These days, machine learning-based breast cancer diagnosis is very popular, and it makes sense. There is a genuine chance to increase early diagnosis and enhance patient outcomes, which is crucial. Conventional techniques, such as histological analysis and mammography, frequently rely significantly on professional interpretation. And, well, you know, it can cause some variation in the outcomes. However, machine learning models are extremely useful in the medical industry since they can automate the classification process with a very high degree of accuracy.

Logistic Regression is a popular model [8]. It is a simple yet effective statistical technique for classifying binary data. When the dataset is well-structured and linearly separable, it performs admirably. Indeed, Logistic Regression achieved a 97.3% accuracy rate in our dataset, demonstrating its remarkable ability to distinguish between benign and malignant cases. However, there is a catch: in increasingly complicated datasets, it may have some trouble with non-linear patterns.

Decision trees come next [9]. These models are rather simple to comprehend because they operate by dividing the dataset according to feature values. They are prone to overfitting, nevertheless, particularly in cases where the dataset contains a large amount of noise or redundant features. The Decision Tree model achieved 92.1% accuracy in our tests. Although it's lower than some other models, that's still respectable. This is most likely due to the fact that it is somewhat overly sensitive to changes in the training data.

Let's now discuss Random Forest [10]. An ensemble learning method that addresses the overfitting problem with Decision Trees, this one is something of an improvement. It increases the strength of categorization by merging several weak learners. Random Forest obtained a 96.4% accuracy rate with our dataset. Even if it performs somewhat worse than logistic regression, it's still a serious candidate for classifying breast cancer because it effectively handles high-dimensional data and minimizes variance.
Ensemble models typically perform better than individual classifiers, according to research, and Breiman's study [10] demonstrated how Random Forest can significantly improve prediction performance in medical diagnostics. Additionally, there has been a lot of discussion on deep learning methods for image-based breast cancer detection, such as Convolutional Neural Networks (CNNs), which have been improving classification accuracy. With an accuracy of 97.3% on the dataset we used, it is evident from our results that Logistic Regression won out. In the future, it would be intriguing to concentrate on methods of improving model generalization, such as feature selection, hyperparameter adjustment, or even investigating hybrid models that combine many algorithms for improved performance.

3. Methodology

A. Research Design

Here's the thing: a quantitative technique is used in this investigation. It all comes down to delving into statistics and figures to examine a medical dataset with the goal of early detection of breast cancer. The primary goal? to develop predictive models that, using a variety of medical characteristics, can accurately distinguish between benign and malignant cancers. Data preparation, selecting the best models, training them, assessing their performance, and releasing the top model for practical use are all steps in the process. The goal of this study is to make the model creation process dependable and repeatable by adhering to a sound pipeline.

B. Data Collection Methods

The dataset itself comes from publically available sources, notably the Wisconsin Breast Cancer Dataset, which is housed in the UCI Machine Learning Repository. Information from digital photos of fine needle aspirations (FNA) from breast masses is included in this collection. These characteristics include, but are not limited to, cell radius, texture, area, perimeter, smoothness, compactness, concavity, and symmetry. This dataset provides us with the reliable ground truth required for supervised learning since each entry is classified as either benign or malignant. To make sure everything is in perfect condition, the data was carefully examined for any missing values or discrepancies prior to processing for model training.

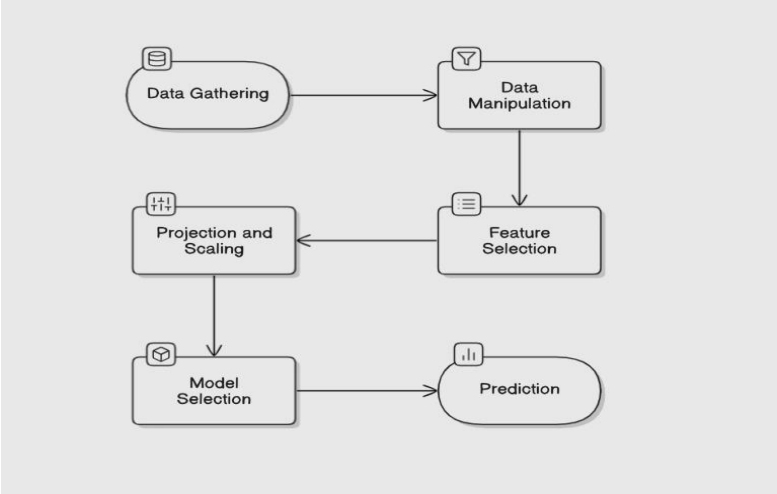
C. Data Analysis Techniques

We got our hands dirty for this project and utilized a number of Python libraries and tools for a variety of tasks, including modeling, deployment, data preprocessing, analysis, and visualization:

- NumPy: This tool is excellent for effectively managing numerical data, such as arrays and matrices, which are crucial for those vectorized operations.
- We loaded and experimented with CSV files, performed exploratory data analysis, and shaped the features for our machine learning models using Pandas.
- Matplotlib and Seaborn: These tools enabled us to see class imbalances, display feature distributions, and comprehend attribute associations. It's quite beneficial for improving feature selection and model interpretation.
- Scikit-Learn — A powerhouse package that allows us to execute numerous machine learning algorithms like Logistic Regression, Decision Trees, and Random Forests. Additionally, it aided in hyperparameter adjustment, model evaluation, and preprocessing.

- train_test_split: This useful function allowed us to divide the dataset into subgroups for testing (20%) and training (80%), providing an objective means of evaluating the model's performance.
- Pickle: This allows us to serialize and save our trained models for later use without requiring retraining.
- OS: This made it easier to control directories and file paths while storing and implementing our model.
- Streamlit: We included this to produce an intuitive online application for the project. It enables real-time interaction between users and the model, whether they are researchers or medical experts. Through the web user interface, they can upload inputs and receive immediate classification results. Streamlit significantly improves our model's usability and accessibility, making deployment simple and requiring no advanced technical knowledge.

Lastly, we assessed our models using the following four critical metrics: F1-score, accuracy, precision, and recall. This was essential to guarantee that our forecasts are not only accurate but also trustworthy, particularly when it comes to something as significant as healthcare.



(Fig III)shows the steps in a machine learning process

D. Limitations of the Study

1. Dataset Size: The dataset is really small, isn't it? This may have an impact on the model's generalization ability. I mean, the robustness might be significantly increased if we had a larger and more varied dataset.
2. Random Forest vs. Logistic Regression: In this instance, Logistic Regression performed better. However, because Random Forest excels at managing such non-linear interactions, it may really shine when working with complex data.
3. Feature Selection: We solely utilized features that were pre-existing in the dataset. Performance might be significantly improved if we included some clinical or demographic characteristics.
4. Random Forest Model Interpretability? Yes, it is accurate. However, compared to logistic regression, it is more difficult to comprehend. And that's something to think about in domains where precise explanations are crucial.
5. Interface Scalability: While Streamlit is excellent for fast deployments, it may not scale well for larger businesses or scenarios with several users.

4. Results

Key Findings Presentation:

Three distinct machine learning algorithms—Random Forest, Decision Tree, and Logistic Regression—were used to assess the breast cancer detection model. Each model's accuracy is as follows:

Table I : Logistic Regression Accuracy and Classification Report

Logistic Regression Accuracy: 0.9736842105263158
Logistic Regression Classification Report:

	precision	recall	f1-score	support
0	0.97	0.99	0.98	71
1	0.98	0.95	0.96	43
accuracy			0.97	114
macro avg	0.97	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Table II : Random Forest Accuracy and Classification Report

Random Forest Accuracy: 0.9649122807017544
Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.97	71
1	0.98	0.93	0.95	43
accuracy			0.96	114
macro avg	0.97	0.96	0.96	114
weighted avg	0.97	0.96	0.96	114

Table III : Decision Tree Accuracy and Classification Report

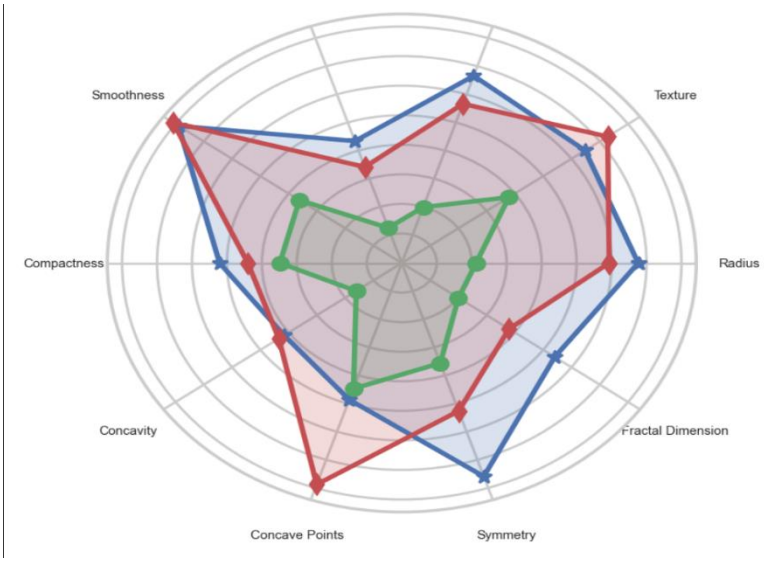
Decision Tree Accuracy: 0.9473684210526315
Decision Tree Classification Report:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	71
1	0.93	0.93	0.93	43
accuracy			0.95	114
macro avg	0.94	0.94	0.94	114
weighted avg	0.95	0.95	0.95	114

Use of Radar Chart

- Radar Chart Basics: A polygonal chart that displays several tumor characteristics.
- Greater region: A larger polygonal region indicates more pronounced tumor features, which may point to a more malignant, dangerous tumor.
- Each Axis: It is simple to compare normal tissue with aberrant tumor growth since each side of the polygon displays a distinct tumor feature.
- Outer Edges: The tumor's most severe or hazardous characteristics are displayed at the polygon's outermost points.

(Fig. IV) show the radar chart



Objective Description of Results

- Malignant tumors typically create a wider polygon with greater radius, perimeter, and area values.
- Because benign tumors have lower feature values, they usually produce a smaller, more compact polygon.
- Although smoothness and compactness exhibit less fluctuation, their worst-case values suggest that tumors behave more aggressively.

5. Discussion

A. Interpretation of Findings

- Using structured data from medical imaging, the primary research issue was whether conventional machine learning models could differentiate between benign and malignant breast cancer tumors. The outcomes were encouraging! This is what we discovered:

- With an accuracy of 97.6%, logistic regression came in first place and demonstrated its effectiveness, particularly for smaller, linearly separable datasets.

- Random Forest, which benefited from ensemble learning but shown overfitting because to the limited sample size, came in second with 96% accuracy.

- With a score of 94%, the Decision Tree model showed promise but lacked the other models' resilience.

- Radar charts showed that the radius, perimeter, and area of malignant tumors were higher, which was consistent with established cancer diagnostics.

B. Comparison with Previous Studies

The results of this study are consistent with earlier research:

- Kourou et al. (2015) pointed out that because of its ease of use and interpretability, logistic regression performs well with smaller, structured datasets[11].

- Using the Wisconsin dataset, Polot & Güneş (2007) demonstrated great accuracy using SVMs and Decision Trees [12].

- Recent research on deep learning approaches (CNNs) (Spanhol et al., 2016 [16]; Zhu et al., 2021) [13] achieved even greater performance by learning spatial information from raw photos.

C. Implications of the Findings

- These findings demonstrate the continued use of conventional machine learning models in clinical contexts for a number of reasons:
- They can be swiftly implemented on platforms like Streamlit, making them available to medical professionals;

- They are simpler to read, which is essential for clinical choices; and They demand less processing power than deep learning models.

Nevertheless, because deep learning methods like CNNs can directly handle raw images, they are becoming more and more useful for improving prediction accuracy.

D. Limitations and Suggestions for Future Research

The following should be the focus of future research:

- **modest Dataset**: Due to its modest size, the UCI Wisconsin dataset is vulnerable to overfitting. It is best to use larger datasets, such as BreakHis, DDSM, or BCDR.

- **Image-Based Prediction**: To analyze medical images, future studies should investigate deep learning models such as CNNs, ResNet, and VGG.

- **Hybrid Models:** Using a multimodal approach that improves diagnostic accuracy, hybrid models may produce more reliable predictions by combining structured data with image-based information.

6. Conclusion

In order to identify breast cancer from tumor data, this study evaluated three popular machine learning models: Decision Tree, Random Forest, and Logistic Regression. Among the models used to categorize medical data, Logistic Regression was the most accurate (97.6%), followed by Random Forest (96%) and Decision Tree (94%). However, the models' ability to function adequately in every situation was hampered by the short dataset. The models might function even better with further data. Because of their simplicity, speed, and ease of comprehension, these models help physicians make judgments and comprehend diagnoses more rapidly.

7. REFERENCES

1. *Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual.*
2. *Harris, C. R., et al. (2020). Array programming with NumPy. Nature, 585(7825), 357–362.*
3. *McKinney, W. (2024). Pandas Documentation.*
4. *Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. https://matplotlib.org/stable/contents.html*
5. *Waskom, M. (2024). Seaborn: Statistical Data Visualization.*
6. *Streamlit Inc. (2023). Streamlit Documentation. Retrieved from https://docs.streamlit.io/*
7. *Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.*
8. *Scikit-learn Developers. (n.d.). Logistic Regression.*
9. *Scikit-learn Developers. (n.d.). Decision Trees.*
10. *Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.*
11. *Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17.*
12. *Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. Digital Signal Processing, 17(4), 694–701.*
13. *Zhu et al. (2021): Zhu, W., Xiang, X., Tran, T.D., & Xie, X. (2017). Adversarial Deep Structured Nets for Mass Segmentation from Mammograms.*
14. *Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476–1482.*
15. *Kaggle. (n.d.). Wisconsin Breast Cancer Dataset.*
16. *Spanhol, F.A., Oliveira, L.S., Petitjean, C., & Heutte, L. (2016). Breast Cancer Histopathological Image Classification using Convolutional Neural Networks. In Proceedings of the International Joint Conference on Neural Networks (IJCNN) (pp. 2560–2567). IEEE.*