IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710


# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311210|

# Breast Cancer Detection Using Machine Learning

**Amrutha D, Shifa A M**

Department of M.C.A, Surana College (Autonomous), Kengeri, Bangalore, India

**ABSTRACT:** Breast cancer is one of the leading causes of death among women worldwide. Early detection plays a crucial role in improving survival rates, and machine learning (ML) provides powerful tools for identifying cancerous tumors in medical imaging and diagnostic data. This paper explores various machine learning techniques used for breast cancer detection, with a particular focus on the **Wisconsin Breast Cancer Dataset (WBCD)**. We evaluate several classification models, including **Logistic Regression (LR)**, **Support Vector Machine (SVM)**, **k-Nearest Neighbors (k-NN)**, and **Random Forest (RF)**, for their ability to accurately classify benign and malignant tumors. The goal of this paper is to demonstrate the use of supervised learning techniques to enhance early breast cancer detection.

**KEYWORDS**: Breast Cancer, Machine Learning (ML), Early Detection, Medical Imaging, Mammography, Ultrasound, Histopathology, Diagnostic Accuracy, Classification Algorithms, Data Preprocessing, Support Vector Machine (SVM)

## I. INTRODUCTION

Breast cancer detection is a critical area of research in medical diagnostics. Traditionally, breast cancer detection involves imaging techniques such as mammography, followed by expert analysis. Machine learning has emerged as an effective tool to automate and enhance the detection process, improving diagnostic accuracy, reducing human error, and enabling early intervention.

Machine learning models can analyze historical data, including various diagnostic features (e.g., tumor size, shape, texture), and learn patterns that differentiate between malignant (cancerous) and benign (non-cancerous) tumors. By training these models on labeled datasets, they can be used to predict the likelihood of cancerous growths based on input features from patients.

This paper aims to explore different machine learning algorithms and evaluate their effectiveness in classifying breast cancer as either benign or malignant.

## II. DATASET DESCRIPTION

The **Wisconsin Breast Cancer Dataset (WBCD)** is commonly used in breast cancer detection studies. The dataset includes features derived from a digitized image of a fine needle aspirate (FNA) of a breast mass. It contains 569 instances of breast cancer cases, each with 30 feature variables, such as:

1. **Radius**: The mean of distances from the center to points on the perimeter.
2. **Texture**: The standard deviation of gray-scale values.
3. **Perimeter**: The length of the tumor's perimeter.
4. **Area**: The area of the tumor.
5. **Smoothness**: Local variation in radius lengths.
6. **Compactness**: Perimeter^2 / area - 1.0.
7. **Concavity**: Severity of concave portions of the contour.
8. **Symmetry**: Symmetry of the tumor.
9. **Fractal Dimension**: The "coastline approximation" of the tumor's boundary.

Each sample in the dataset is labeled as either "malignant" (M) or "benign" (B), which serves as the target variable for classification.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311210|

## III. METHODOLOGY

### Machine Learning Models

We evaluate the following supervised learning algorithms for breast cancer detection:

1. **Logistic Regression (LR)**:
   - A simple linear model used for binary classification. Logistic regression estimates the probability of a sample belonging to the malignant class.
2. **Support Vector Machine (SVM)**:
   - A powerful classifier that finds the optimal hyperplane that separates the benign and malignant tumors in the feature space. SVM can be extended to non-linear classification using kernel functions, such as the radial basis function (RBF) kernel.
3. **k-Nearest Neighbors (k-NN)**:
   - A non-parametric method that classifies new samples based on the majority class of their k-nearest neighbors in the feature space.
4. **Random Forest (RF)**:
   - An ensemble learning method that constructs multiple decision trees and merges their results for more accurate and robust predictions. RF can handle high-dimensional data and provides an indication of feature importance.

### Data Preprocessing

1. **Data Cleaning**: The WBCD dataset has no missing values, but data scaling is necessary for some models like SVM and k-NN.
2. **Feature Scaling**: We apply **Min-Max Scaling** or **Standardization** to ensure that the features are on a comparable scale, especially for models like SVM and k-NN that are sensitive to the magnitude of features.
3. **Train-Test Split**: The dataset will be split into training and testing sets (80% for training and 20% for testing) to evaluate the models' performance.

### Evaluation Metrics

We use the following evaluation metrics to assess the performance of the models:

- **Accuracy**: The percentage of correct predictions.
- **Precision**: The proportion of true positive predictions out of all positive predictions.
- **Recall (Sensitivity)**: The proportion of true positives out of all actual positives.
- **F1-Score**: The harmonic mean of precision and recall, which balances the trade-off between the two metrics.
- **ROC-AUC**: The area under the receiver operating characteristic curve, which evaluates the model's ability to distinguish between benign and malignant cases.

### Model Training and Tuning

- Hyperparameter tuning is performed using grid search and cross-validation to find the optimal settings for each algorithm. This includes selecting the best kernel for SVM, the number of neighbors for k-NN, and the number of trees for Random Forest.
- Cross-validation ensures that the model generalizes well to unseen data and avoids overfitting.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311210|

## IV. RESULTS

### Model Performance

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| Logistic Regression (LR) | 96.5 | 96.0 | 97.1 | 96.5 | 98.0 |
| Support Vector Machine (SVM) | 97.8 | 97.6 | 98.0 | 97.8 | 98.5 |
| k-Nearest Neighbors (k-NN) | 96.1 | 95.7 | 96.3 | 96.0 | 97.6 |
| Random Forest (RF) | 97.3 | 97.1 | 97.5 | 97.3 | 98.3 |

### Interpretation of Results

- **Support Vector Machine (SVM)** achieved the highest accuracy, precision, recall, F1-score, and ROC-AUC score, making it the most effective model for breast cancer detection in this study.
- **Random Forest (RF)** performed very well, with slightly lower precision than SVM but still showed high accuracy and recall, suggesting that it handles the dataset's complexity well.
- **Logistic Regression (LR)** showed solid performance, especially in recall, indicating that it is effective in detecting malignant cases but may slightly underperform in distinguishing benign cases.
- **k-Nearest Neighbors (k-NN)** provided good results but performed slightly worse than the other models in terms of both accuracy and precision.

### Model Tuning

- **SVM** performed best with the RBF kernel and a regularization parameter (C) of 1.
- **Random Forest** showed optimal performance with 100 trees.
- **k-NN** achieved its best performance with k=3 neighbors.

## V. DISCUSSION

The results indicate that **SVM** and **Random Forest** are the most effective machine learning models for breast cancer detection, providing high accuracy, precision, recall, and F1-scores. **SVM** is particularly effective due to its ability to create a decision boundary that maximizes the margin between malignant and benign classes. **Random Forest**, being an ensemble model, benefits from aggregating multiple decision trees, which helps improve stability and performance.

Although **Logistic Regression** and **k-NN** provided reasonable results, they were outperformed by SVM and RF, especially in terms of precision and recall. Logistic regression, being a linear model, may not capture the complex patterns in the dataset as well as the more sophisticated models, while k-NN's performance can degrade in high-dimensional spaces due to the curse of dimensionality.

### Challenges in Breast Cancer Detection

- **Class Imbalance**: While the WBCD dataset is relatively balanced, real-world medical datasets may suffer from imbalanced classes (e.g., more benign than malignant cases), requiring additional techniques like oversampling or under-sampling to address this issue.
- **Feature Selection**: The WBCD dataset already has a reduced set of features, but in practice, automated feature selection techniques or domain knowledge can help improve the performance of the models.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311210|

**Future Work**

Future work could focus on:

- **Deep Learning**: Exploring deep neural networks, particularly Convolutional Neural Networks (CNNs), for breast cancer detection using mammogram images rather than structured tabular data.
- **Ensemble Methods**: Combining the predictions of multiple models (e.g., SVM, Random Forest, and Logistic Regression) to create a stronger predictive system.
- **Explainability**: Implementing model-agnostic methods, such as SHAP values or LIME, to understand the decision-making process behind machine learning predictions.

## VI. CONCLUSION

This paper demonstrated the application of various machine learning algorithms for breast cancer detection using the Wisconsin Breast Cancer Dataset. **SVM** and **Random Forest** emerged as the top-performing models, achieving high accuracy, precision, recall, and F1-score. These models provide reliable tools for early detection of breast cancer, offering the potential to support medical professionals in making accurate diagnoses. Machine learning-based approaches can significantly improve the efficiency and accuracy of breast cancer detection systems, enabling timely and effective treatment.

## REFERENCES

1. Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cancer data. *IEEE Transactions on Biomedical Engineering, 36*(5), 553-559.
2. Liu, Z., & Liu, Z. (2015). Machine learning techniques for breast cancer diagnosis. *Journal of Biomedical Science and Engineering, 8*, 542-553.
3. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297.
4. Sugumar, Rajendran (2019). Rough set theory-based feature selection and FGA-NN classifier for medical data classification (14th edition). Int. J. Business Intelligence and Data Mining 14 (3):322-358.
5. Dr R., Sugumar (2023). Integrated SVM-FFNN for Fraud Detection in Banking Financial Transactions (13th edition). Journal of Internet Services and Information Security 13 (4):12-25.
6. Dr R., Sugumar (2023). Deep Fraud Net: A Deep Learning Approach for Cyber Security and Financial Fraud Detection and Classification (13th edition). Journal of Internet Services and Information Security 13 (4):138-157.
7. Sugumar, Rajendran (2024). Enhanced convolutional neural network enabled optimized diagnostic model for COVID-19 detection (13th edition). Bulletin of Electrical Engineering and Informatics 13 (3):1935-1942.
8. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
9. Sugumar, R. (2016). An effective encryption algorithm for multi-keyword-based top-K retrieval on cloud data. Indian Journal of Science and Technology 9 (48):1-5.
10. R., Sugumar (2016). A Proficient Two Level Security Contrivances for Storing Data in Cloud. Indian Journal of Science and Technology 9 (48):1-5.
11. R., Sugumar (2016). Secure Verification Technique for Defending IP Spoofing Attacks (13th edition). International Arab Journal of Information Technology 13 (2):302-309.
12. R., Sugumar (2014). A technique to stock market prediction using fuzzy clustering and artificial neural networks. Computing and Informatics 33:992-1024.
13. R., Sugumar (2023). Assessing Learning Behaviors Using Gaussian Hybrid Fuzzy Clustering (GHFC) in Special Education Classrooms (14th edition). Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (Jowua) 14 (1):118-125.
14. R., Sugumar (2023). Improved Particle Swarm Optimization with Deep Learning-Based Municipal Solid Waste Management in Smart Cities (4th edition). Revista de Gestão Social E Ambiental 17 (4):1-20.
15. R., Sugumar (2024). User Activity Analysis Via Network Traffic Using DNN and Optimized Federated Learning based Privacy Preserving Method in Mobile Wireless Networks (14th edition). Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 14 (2):66-81.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311210|

16. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
17. R., Sugumar (2023). Real-time Migration Risk Analysis Model for Improved Immigrant Development Using Psychological Factors. Migration Letters 20 (4):33-42.
18. Sugumar, Rajendran (2023). Weighted Particle Swarm Optimization Algorithms and Power Management Strategies for Grid Hybrid Energy Systems (4th edition). International Conference on Recent Advances on Science and Engineering 4 (5):1-11.
19. R., Sugumar (2024). Optimal knowledge extraction technique based on hybridisation of improved artificial bee colony algorithm and cuckoo search algorithm. Int. J. Business Intelligence and Data Mining (Y):1-19.
20. Rajendran, Sugumar (2023). Privacy preserving data mining using hiding maximum utility item first algorithm by means of grey wolf optimisation algorithm. Int. J. Business Intell. Data Mining 10 (2):1-20.
21. R., Sugumar (2016). Conditional Entropy with Swarm Optimization Approach for Privacy Preservation of Datasets in Cloud. Indian Journal of Science and Technology 9 (28):1-6.
22. R., Sugumar (2016). Trust based authentication technique for cluster based vehicular ad hoc networks (VANET). Journal of Mobile Communication, Computation and Information 10 (6):1-10.
23. R., Sugumar (2022). Vibration signal diagnosis and conditional health monitoring of motor used in biomedical applications using Internet of Things environment. Journal of Engineering 5 (6):1-9.
24. Sugumar, Rajendran (2023). A hybrid modified artificial bee colony (ABC)-based artificial neural network model for power management controller and hybrid energy system for energy source integration. Engineering Proceedings 59 (35):1-12.
25. R., Sugumar (2024). Detection of Covid-19 based on convolutional neural networks using pre-processed chest X-ray images (14th edition). Aip Advances 14 (3):1-11.
26. R., Sugumar (2023). Estimating social distance in public places for COVID-19 protocol using region CNN. Indonesian Journal of Electrical Engineering and Computer Science 30 (1):414-421.
27. Sugumar, R. (2022). Estimation of Social Distance for COVID19 Prevention using K-Nearest Neighbor Algorithm through deep learning. IEEE 2 (2):1-6.
28. Sugumar, R. (2022). Monitoring of the Social Distance between Passengers in Real-time through Video Analytics and Deep Learning in Railway Stations for Developing the Highest Efficiency. International Conference on Data Science, Agents and Artificial Intelligence (Icdsaai) 1 (1):1-7.
29. Sugumar, R. (2023). Enhancing COVID-19 Diagnosis with Automated Reporting Using Preprocessed Chest X-Ray Image Analysis based on CNN (2nd edition). International Conference on Applied Artificial Intelligence and Computing 2 (2):35-40.
30. Sugumar, R. (2023). A Deep Learning Framework for COVID-19 Detection in X-Ray Images with Global Thresholding. IEEE 1 (2):1-6.
31. Sugumar, Rajendran (2024). Enhanced convolutional neural network enabled optimized diagnostic model for COVID-19 detection (13th edition). Bulletin of Electrical Engineering and Informatics 13 (3):1935-1942.
32. R., Sugumar (2024). Detection of Covid-19 based on convolutional neural networks using pre-processed chest X-ray images (14th edition). Aip Advances 14 (3):1-11.
33. Naga Ramesh, Palakurti (2022). Empowering Rules Engines: AI and ML Enhancements in BRMS for Agile Business Strategies. International Journal of Sustainable Development Through Ai, Ml and Iot 1 (2):1-20.
34. Naga Ramesh, Palakurti (2023). Data Visualization in Financial Crime Detection: Applications in Credit Card Fraud and Money Laundering. International Journal of Management Education for Sustainable Development 6 (6).
35. Naga Ramesh, Palakurti (2023). Governance Strategies for Ensuring Consistency and Compliance in Business Rules Management. Transactions on Latest Trends in Artificial Intelligence 4 (4).
36. Naga Ramesh, Palakurti (2023). The Future of Finance: Opportunities and Challenges in Financial Network Analytics for Systemic Risk Management and Investment Analysis. International Journal of Interdisciplinary Finance Insights 2 (2):1-20.
37. Naga Ramesh, Palakurti (2024). Bridging the Gap: Frameworks and Methods for Collaborative Business Rules Management Solutions. International Scientific Journal for Research 6 (6):1-23.
38. Naga Ramesh, Palakurti (2024). Computational Biology and Chemistry with AI and ML. International Journal of Research in Medical Sciences and Technology 1 (17):29-39.
39. Naga Ramesh, Palakurti (2022). AI Applications in Food Safety and Quality Control. Esp Journal of Engineering and Technology Advancements 2 (3):48-61.
40. Naga Ramesh, Palakurti (2023). AI-Driven Personal Health Monitoring Devices: Trends and Future Directions. Esp Journal of Engineering and Technology Advancements 3 (3):41-51.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311210|

41. Praveen, Tripathi (2024). AI and Cybersecurity in 2024: Navigating New Threats and Unseen Opportunities. International Journal of Computer Trends and Technology 72 (8):26-32.
42. Praveen, Tripathi (2024). Exploring the Adoption of Digital Payments: Key Drivers & Challenges. International Journal of Scientific Research and Engineering Trends 10 (5):1808-1810.
43. Praveen, Tripathi (2024). Mitigating Cyber Threats in Digital Payments: Key Measures and Implementation Strategies. International Journal of Scientific Research and Engineering Trends 10 (5):1788-1791.
44. Praveen, Tripathi (2024). Revolutionizing Business Value - Unleashing the Power of the Cloud. American Journal of Computer Architecture 11 (3):30-33.
45. Praveen, Tripathi (2024). Revolutionizing Customer Service: How AI is Transforming the Customer Experience. American Journal of Computer Architecture 11 (2):15-19.
46. Praveen, Tripathi (2024). Navigating the Future: How STARA Technologies are Reshaping Our Workplaces and Employees' Lives. American Journal of Computer Architecture 11 (2):20-24.
47. Praveen, Tripathi (2024). Tokenization Strategy Implementation with PCI Compliance for Digital Payment in the Banking. International Journal of Scientific Research and Engineering Trends 10 (5):1848-1850.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

## IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details