ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206

Volume 8, Issue 11, November 2025

© 2025 IJMRSET | Volume 8, Issue 11, November 2025| DOI:10.15680/IJMRSET.2025.0811028

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Predictive Healthcare Analytics System

C Pavan Adithya, G Sai Pavan Kumar, G Ravi Chandra, P Manoj Kumar, G Veera Sai Nadha Reddy, B Pavan Kumar

UG Students, Dept. of ECE, Jain University, Bangalore, Karnataka, India

ABSTRACT: Cancer remains one of the leading causes of mortality worldwide, with cases expected to rise significantly by 2040. To address the limitations of traditional diagnostic methods, this work proposes an AI-based multi-cancer prediction system capable of detecting breast, lung, and prostate cancer using machine learning algorithms. The system employs preprocessing, feature filtration, and three classification models—AdaBoost-Logistic Regression, Gaussian Naïve Bayes, and Gradient Boosting—to enhance prediction accuracy. Experimental results show high performance, with accuracies of 99.12%, 96%, and 98.02% respectively across the three cancers. The proposed system offers a scalable, reliable, and efficient solution to support early diagnosis and clinical decision-making, highlighting the growing impact of AI in medical prediction.

KEYWORDS: Artificial Intelligence, Cancer Prediction, Machine Learning, AdaBoost-Logistic Regression, Gaussian Naïve Bayes, Gradient Boosting.

I. INTRODUCTION

The Disease Prediction System Using AI focuses on addressing the growing global burden of cancer by developing an automated, accurate, and reliable multi-cancer prediction model. With cancer cases expected to rise significantly by 2040, early diagnosis has become essential for improving survival rates and reducing pressure on healthcare systems. Traditional diagnostic methods such as biopsies, radiology, and laboratory tests are often costly, time-consuming, and dependent on expert interpretation, which may introduce delays or inaccuracies. To overcome these challenges, this project applies advanced machine learning techniques to analyze structured medical data and detect three major cancers—breast cancer, lung cancer, and prostate cancer. The system uses preprocessing and correlation-based feature selection to refine input data, improving the reliability and precision of predictions.

In this work, three supervised machine learning models—**AdaBoost-Logistic Regression, Gaussian Naïve Bayes, and Gradient Boosting**—are developed and tested for their effectiveness in cancer prediction. AdaBoost-Logistic achieved **99.12% accuracy** for breast cancer detection, Gaussian Naïve Bayes delivered **96% accuracy** for lung cancer prediction, and Gradient Boosting reached **98.02% accuracy** in prostate cancer classification. Comparative analysis showed that ensemble-based methods outperform traditional classifiers like SVM and Random Forest in terms of precision, robustness, and error minimization. Overall, the project successfully demonstrates the potential of AI-driven diagnostic tools to support clinicians, enable early cancer screening, reduce diagnostic delays, and provide a scalable system that can be integrated into hospitals, telemedicine platforms, and electronic health record (EHR) systems. This paper highlights how machine learning can enhance medical decision-making and contribute to more accurate and accessible healthcare solutions.

II. SYSTEM MODEL AND ASSUMPTIONS

The proposed Disease Prediction System is designed to analyze structured medical datasets and classify cancer types using supervised machine learning techniques. It considers a dataset consisting of N patient records, where each record contains multiple clinical attributes such as age, tumor measurements, medical history parameters, and diagnostic indicators. It is assumed that the dataset consists of M independent features that contribute to the classification of breast, lung, and prostate cancer. Each feature is preprocessed through normalization, missing-value handling, and high-correlation filtration to ensure clean and balanced input for model training.

The system operates by selecting the most relevant features from the dataset before applying classifiers. This selection acts as a "path selection" mechanism, enabling the model to focus only on attributes that significantly influence cancer prediction. The communication among different components of the system—data preprocessing, feature extraction, and

© 2025 IJMRSET | Volume 8, Issue 11, November 2025| DOI:10.15680/IJMRSET.2025.0811028

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

model classification—is sequential, where the output of one module becomes the input to the next. Similar to frequency hopping in communication networks, the system iteratively evaluates multiple models (AdaBoost-Logistic Regression, Gaussian Naïve Bayes, and Gradient Boosting) and adapts to the best-performing classifier for each cancer type based on accuracy and error metrics.

It is assumed that the dataset used for training and testing is reliable, labeled, and representative of real-world clinical cases. The system also assumes sufficient computational resources to train ensemble models and perform iterative optimization. During prediction, each model receives processed input features and generates classification results, which act as diagnostic signals indicating whether a patient is at risk of breast, lung, or prostate cancer. To ensure robustness, it is assumed that the system can handle moderate noise in input data and that all preprocessing steps are completed within feasible computation time. Furthermore, the system assumes that predictions are used as supportive diagnostic suggestions for clinicians rather than definitive medical decisions.

III. METHODOLOGY

The proposed Disease Prediction System follows a structured pipeline that processes patient medical data and classifies it into breast, lung, or prostate cancer categories using machine learning models. The system begins with data preprocessing, where missing values are handled, features are normalized, and high-correlation filtration is applied to remove redundant attributes. This filtration ensures that only the most influential clinical parameters are considered, improving model efficiency and accuracy. Once the essential features are identified, the system determines the most suitable classifier for each cancer type by analyzing performance metrics such as accuracy, precision, and error rates. This selection mechanism functions similarly to optimal path selection in communication systems, enabling the system to choose the most efficient predictive “route” for each cancer dataset.

After model selection, the system applies the chosen algorithms—AdaBoost-Logistic Regression for breast cancer, Gaussian Naïve Bayes for lung cancer, and Gradient Boosting for prostate cancer—to generate prediction outputs. Each classifier receives the processed input features and computes a probability-based decision score that determines whether the patient record indicates a cancer risk. The algorithms iteratively optimize themselves to reduce residual errors and improve generalization, ensuring robust performance even in the presence of noisy or imperfect data. By integrating preprocessing, feature selection, and supervised learning into a unified framework, the system provides a reliable and scalable methodology capable of supporting early cancer detection and aiding clinicians in faster decision-making.

IV. IMPLEMENTATION

The implementation of the Disease Prediction System is organized as a modular pipeline that converts raw clinical records into validated cancer predictions. The system uses Python (pandas, NumPy, scikit-learn) as the core stack and was developed and tested on a standard desktop environment (Intel i5 / 16GB RAM; GPU optional). Implementation begins with dataset ingestion and verification: labeled datasets for breast, lung, and prostate cancer are loaded, schema-checked, and merged where applicable. Preprocessing steps include missing-value imputation (mean/median or model-based as appropriate), outlier detection and removal, normalization/standardization of continuous features, categorical encoding (one-hot or ordinal), and classbalance handling (SMOTE or class-weighting when necessary). A high-correlation feature filter is applied to drop strongly collinear attributes, retaining a compact, high-signal feature set for each cancer model. Model training follows a supervised-learning workflow with reproducibility controls (fixed random seeds, stratified splits). Each cancer type is associated with the classifier selected by comparative analysis:

1. **AdaBoost + Logistic Regression (Breast cancer):** AdaBoost wraps logistic regression as the base estimator; weak learners are trained iteratively with sample reweighting. Hyperparameters tuned include number of estimators, learning rate, and regularization strength for logistic regression. Early stopping and cross-validation folds are used to detect and reduce overfitting (noting that very high training accuracy can indicate overfitting risk).
2. **Gaussian Naïve Bayes (Lung cancer):** Implemented for its efficiency with continuous features under Gaussian assumptions. Variance smoothing and feature-wise checks are applied to ensure stable probability estimates. No heavy hyperparameter tuning is required; evaluation focuses on calibration and sensitivity/specificity trade-offs.
3. **Gradient Boosting (Prostate cancer):** Implemented using a tree-based gradient boosting library (scikit-learn’s HistGradientBoostingClassifier or XGBoost/LightGBM where available). Key hyperparameters tuned include number of trees, learning rate, max depth, and subsampling ratios.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Early stopping on a validation fold and regularization (shrinkage, column/row subsample) are applied to control complexity.

Training and validation use stratified k-fold cross-validation with hold-out test sets. Evaluation metrics recorded include accuracy, precision, recall (sensitivity), specificity, F1-score, ROC-AUC, and confusion matrices. Models are calibrated (Platt scaling or isotonic regression) when probability outputs are used for clinical thresholds. The reported performances in experiments—AdaBoost-Logistic ≈ **99.12%** (breast), Gaussian NB ≈ **96%** (lung), and Gradient Boosting ≈ **98.02%** (prostate)—are obtained from test-set evaluations after hyperparameter tuning and cross-validation.

V. RESULT AND DISCUSSION

The performance of the Disease Prediction System was evaluated using separate datasets for breast, lung, and prostate cancer. Each machine learning model was trained using an 80:20 train–test split, followed by stratified k-fold cross-validation to ensure stable and unbiased evaluation. The results obtained demonstrate that the selected classifiers effectively handle the variability and complexity present in the medical datasets, providing high predictive accuracy across all three cancer types. The evaluation metrics used include accuracy, precision, recall, F1-score, and confusion matrices, enabling a comprehensive performance comparison between models.

For breast cancer prediction, the AdaBoost-Logistic Regression model achieved an accuracy of **99.12%**, outperforming traditional classifiers due to its iterative error-correcting mechanism and ability to strengthen weak learners. The Gaussian Naïve Bayes model used for lung cancer achieved **96% accuracy**, benefiting from its probabilistic nature and strong performance on high-dimensional clinical data. The Gradient Boosting model, applied for prostate cancer prediction, produced **98.02% accuracy**, proving effective in capturing nonlinear relationships and minimizing residual errors. Across all experiments, ensemble-based methods (AdaBoost and Gradient Boosting) consistently outperformed baseline algorithms such as SVM and Random Forest, demonstrating higher robustness, lower test error, and superior precision.

These results indicate that the proposed system can reliably classify cancer types with high accuracy, proving its suitability for early diagnostic support. The findings also show that preprocessing and feature correlation filtration significantly improved prediction quality by reducing noise and enhancing model generalization. While the system performs exceptionally well in controlled datasets, further improvements could be achieved through larger clinical datasets, real-world validation, and integration of deep learning models. Overall, the results confirm the effectiveness of the AI-driven approach, highlighting its potential as a supportive tool for clinicians in early cancer detection and medical decision-making.

VI.CONCLUSION

The Disease Prediction System Using AI was developed with the primary objective of enabling early and accurate detection of three major cancer types—breast, lung, and prostate cancer—through advanced machine learning techniques. With cancer rates rising globally, the need for automated, reliable, and costeffective diagnostic support has become more urgent than ever. This project demonstrates how AI can effectively process structured medical data, identify meaningful patterns, and support clinicians by providing rapid and highly accurate predictions. By applying preprocessing, feature selection, and carefully chosen classifiers, the system successfully minimizes diagnostic delays and enhances decision-making in medical environments.

The experimental results show that the selected models—AdaBoost-Logistic Regression, Gaussian Naïve Bayes, and Gradient Boosting—perform exceptionally well, achieving accuracies of **99.12%**, **96%**, and **98.02%** respectively. These findings confirm the strength of ensemble-based machine learning approaches in handling complex medical datasets. Beyond high accuracy, the system's modular architecture makes it scalable and adaptable for future extensions, such as integrating additional diseases, larger datasets, or deep learning models. While further real-world validation is needed before clinical deployment, this work highlights the growing potential of AI-driven predictive systems in transforming healthcare. Ultimately, the project demonstrates that machine learning can play a significant role in early diagnosis, improved patient outcomes, and the future of intelligent medical decision support.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

[1] Breast Cancer Prediction Based on Multiple Machine Learning Algorithms, Sheng Zhou, Chujiao Hu, 2024
[2] Comparative Analysis of Machine Learning Models for Prostate Cancer Prediction, Saul BeltozarClemente, Enrique Diaz-Vega, 2023
[3] Artificial intelligence based medical decision support system for early and accurate breast cancer prediction, Law Kumar Singh, Munish Khanna,2023
[4] Early Prediction of Lung Cancer Using Gaussian Naive Bayes Classification Algorithm, M. Vedaraj, C.S. Anita, 2023
[5] A Systematic Review of Artifcial Intelligence Techniques in Cancer Prediction and Diagnosis, Yogesh Kumar,Surbi Gupta,2022
[6] Cancer Prediction using Machine Learning, Ganta Sruthi, Neha Sharma, 2022
[7] Prediction of Cancer Disease using Machine learning Approach, F.J. Shaik, D.S. Rao,2022
[8] A Systematic Review of Applications of Machine Learning in Cancer Prediction and Diagnosis, Aman Sharma, Rinkle Rani, 2021
[9] Lung Cancer Incidence Prediction Using Machine Learning Algorithms, Kubra Tuncal, Boran Sekeroglu, 2020
[10] Machine Learning Based Approaches for Cancer Prediction, Ajay Kumar, Rama Sushil, 2019

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com