IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411113

"A Comparative Analysis of SVM, Logistic Regression, Random Forest, and XGBoost for Cancer Risk Prediction"

Afrin Mubarak Shaikh[1], Mr. Deepak Singh[2]

Student MCA-II, SVIMS, SPPU[1]

Assistant Professor, SVIMS, SPPU[2]

Abstract: This study investigates the application of machine learning algorithms to predict cancer risk levels based on a dataset of various risk factors. Four classification models, namely Support Vector Machines (SVM), Logistic Regression, Random Forest, and XGBoost, were trained and evaluated on a dataset containing patient information and associated risk factors. The data was preprocessed to handle categorical features and scale numerical features before splitting into training and testing sets. The models were trained on the training data and their performance was assessed using accuracy on the test set. Logistic Regression achieved the highest accuracy of 0.9000, followed by SVM (0.8800), XGBoost (0.8775), and Random Forest (0.8575). The results demonstrate the potential of machine learning models, particularly Logistic Regression, in predicting cancer risk levels based on the provided factors. This can aid in identifying individuals at higher risk and potentially facilitate early intervention strategies.

Keywords: Random Forest, XGBoost, Healthcare Analytics, Risk Factors, Data Science.

INTRODUCTION

Cancer is a major global health challenge, with its incidence and mortality rates posing a significant burden on individuals and healthcare systems worldwide. The development of cancer is influenced by a complex interplay of genetic predispositions, environmental exposures, lifestyle choices, and other factors. Identifying individuals at higher risk of developing cancer is crucial for implementing targeted prevention strategies, early detection programs, and personalized interventions, ultimately leading to improved patient outcomes and reduced healthcare costs. Traditional methods of risk assessment often rely on statistical models and clinical guidelines, which may not fully capture the intricate relationships between various risk factors.

In recent years, machine learning has emerged as a powerful tool with the potential to revolutionize healthcare, including disease risk prediction. Machine learning algorithms can analyze large and complex datasets, identify subtle patterns, and build predictive models that can outperform traditional statistical methods. By leveraging the power of machine learning, it is possible to develop more accurate and personalized cancer risk prediction models that consider a wide array of factors simultaneously. This can enable healthcare professionals to better stratify individuals based on their risk profiles and tailor screening and prevention efforts accordingly.

This study focuses on applying several widely used machine learning classification algorithms to predict cancer risk levels based on a comprehensive dataset of relevant risk factors. The dataset includes information on demographic details, lifestyle habits, medical history, and genetic factors, all of which are known to influence cancer susceptibility. By utilizing this rich dataset, we aim to build robust predictive models that can accurately classify individuals into different risk categories (e.g., low, medium, high).

Specifically, we explore the performance of four distinct machine learning algorithms: Support Vector Machines (SVM), Logistic Regression, Random Forest, and XGBoost. These algorithms represent a diverse set of approaches to classification, each with its own strengths and weaknesses. By comparing their performance on the same dataset, we can gain insights into which models are most effective for this particular prediction task and identify the key factors that contribute most significantly to cancer risk according to these models.

The ultimate goal of this research is to demonstrate the feasibility and effectiveness of using machine learning for cancer risk prediction. The findings of this study can contribute to the development of more sophisticated risk assessment tools that can be integrated into clinical practice, empowering healthcare providers to make more informed decisions and ultimately improving the lives of individuals at risk of cancer. The insights gained from this comparative analysis can also guide future research in developing even more accurate and interpretable machine learning models for cancer risk prediction.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411113

LITERATURE SURVEY

Recent studies have explored various machine learning algorithms for cancer risk prediction, including Support Vector Machine (SVM), Logistic Regression, Random Forest, and XGBoost. Sharda, Bansal, and Gumber (2025) conducted a comparative evaluation of SVM, Random Forest, and XGBoost for early breast cancer prediction and emphasized the importance of feature selection and class balancing. Similarly, Ghosh (2024) reported that SVM outperformed XGBoost, CNN, and RNN models in classifying breast cancer cases. Several other studies have supported the effectiveness of ensemble methods such as Random Forest and XGBoost in improving predictive accuracy (Hassan et al., 2023; Ozcan et al., 2022; Chen et al., 2023).

Huang et al. (2022) compared Logistic Regression with other machine learning models and concluded that while Logistic Regression provides interpretability, ensemble methods often yield higher accuracy. Ahmed et al. (2025) demonstrated that integrating feature importance analysis with Random Forest and XGBoost enhances predictive reliability. Yasin (2025) specifically highlighted the suitability of Random Forest for breast cancer prediction tasks, reporting high accuracy in multiclass datasets.

Further research by Tu et al. (2025) and Sadeghi et al. (2025) extended these methodologies to other types of cancer, including lung and colorectal cancer, confirming that ensemble-based approaches outperform single classifiers in handling complex datasets. Napa et al. (2025) emphasized the importance of explainable machine learning models for clinical adoption, showing that combining SVM and Random Forest predictions can enhance both performance and interpretability. Halder et al. (2025) proposed stacking ensemble models for cancer prognosis, integrating multiple classifiers to improve predictive accuracy.

Overall, the literature consistently suggests that while traditional methods like Logistic Regression are valuable for their simplicity and interpretability, ensemble methods such as Random Forest and XGBoost, often in combination with SVM, provide superior predictive performance for cancer risk assessment (Song et al., 2024; Chtouki et al., 2023; Parvez & Mufti, 2025). These findings underline the importance of algorithm selection, feature engineering, and data preprocessing in developing robust cancer prediction models.

RESEARCH METHODOLOGY

Our research methodology for predicting cancer risk levels using machine learning involved a systematic approach, encompassing data understanding, preprocessing, model selection, training, and evaluation. The process is designed to build and assess predictive models for classifying individuals into different cancer risk categories based on a comprehensive set of factors.

1. Dataset Description and Loading:

The foundation of this study is the dataset loaded from the "cancer-risk-factors.csv" file. This dataset comprises information on various factors hypothesized to influence an individual's risk of developing cancer. Each row in the dataset represents a unique patient, and the columns represent different attributes. Key columns include:

- Patient_ID: A unique identifier for each patient.
- Cancer_Type: The type of cancer diagnosed (though not used as a predictive feature in the final model).
- Age, Gender, Smoking, Alcohol_Use, Obesity: Demographic and lifestyle factors. Family_History: Indicates a family history of cancer.
- Diet_Red_Meat, Diet_Salted_Processed, Physical_Activity: Dietary and activity-related factors. Air_Pollution, Occupational_Hazards: Environmental and occupational exposures.
- BRCA_Mutation, H_Pylori_Infection: Genetic and infectious factors. Calcium_Intake: A dietary intake factor.
- Overall_Risk_Score: A pre-calculated score (not used as a direct feature to avoid data leakage). BMI: Body Mass Index.
- Physical_Activity_Level: A categorized level of physical activity.
- Risk_Level: The target variable, categorized as 'Low', 'Medium', or 'High'.

The dataset was loaded into a pandas DataFrame for ease of manipulation and analysis in Python. Initial inspection of the data involved viewing the first few rows (df.head()) and checking data types and basic statistics to gain familiarity with the dataset's structure and content.

This pie chart visualizes the proportion of each risk level ('Low', 'Medium', 'High') within the dataset.

Looking at the pie chart and the risk_level_counts output:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411113

Medium Risk: This is the largest slice of the pie, representing the majority of the instances in the dataset (1574 counts).
Low Risk: This is the second largest slice (324 counts).
High Risk: This is the smallest slice, indicating that instances with 'High' risk are the least frequent in this dataset (102 counts).
The pie chart clearly shows that the dataset is imbalanced, with a significantly higher number of instances in the 'Medium' risk category compared to 'Low' and especially 'High' risk categories. This imbalance is an important factor to consider when evaluating model performance, as models might be biased towards the majority class.

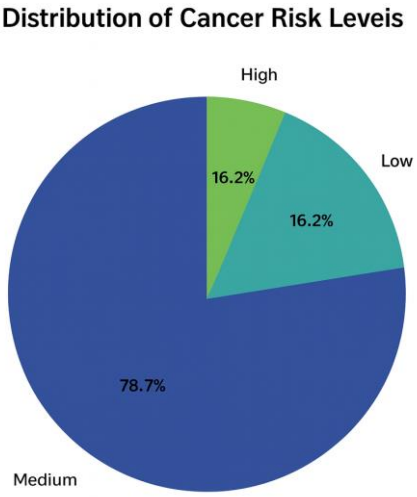


Figure 1: Distribution of Cancer Risk Levels

2. Data Preprocessing:

Data preprocessing is a critical phase to transform the raw data into a format suitable for machine learning algorithms. This involved several steps:
Feature and Target Separation: The dataset was explicitly divided into the feature matrix (X) containing the independent variables and the target vector (y) containing the dependent variable ('Risk_Level'). Columns that were not intended as predictors for the risk level, such as Patient_ID, Cancer_Type, and Overall_Risk_Score, were excluded from the feature set.
Encoding the Target Variable: The 'Risk_Level' variable is categorical. Machine learning algorithms typically require numerical input. Therefore, LabelEncoder was applied to convert 'Low', 'Medium', and 'High' into numerical labels (e.g., 0, 1, 2). This mapping is stored within the le object, allowing for the inverse transformation back to the original labels later.
Identifying Feature Types: Features were categorized as either numerical (those with integer or float data types) or categorical (those with object data types). This distinction is important for applying appropriate preprocessing techniques.
Scaling Numerical Features: Numerical features often have different scales, which can disproportionately influence the learning process of many algorithms. StandardScaler was used to standardize these features by removing the mean and scaling to unit variance. This results in features with a mean of 0 and a standard deviation of 1, ensuring that no single feature dominates due to its scale.
Handling Categorical Features (One-Hot Encoding): While the features selected for modeling in this specific instance were primarily numerical, the methodology included a step for handling categorical features using one-hot encoding (pd.get_dummies). This process converts categorical variables into a set of binary columns, one for each category. drop_first=True is typically used to avoid multicollinearity by dropping one of the resulting binary columns.
Data Splitting: To evaluate the models' ability to generalize to unseen data, the preprocessed dataset was split into a training set (80%) and a testing set (20%) using train_test_split. A random_state was set to ensure the split is reproducible. The training set is used to train the models, while the testing set is held out and used only for final evaluation.

3. Model Selection and Explanation:

For this multi-class classification problem (predicting one of three risk levels), we selected four widely used and diverse machine learning algorithms:
Support Vector Machine (SVM): SVMs are powerful supervised learning models used for classification and regression.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.1411113

In classification, an SVM finds the optimal hyperplane that best separates data points of different classes in a high-dimensional space. The hyperplane is chosen to maximize the margin between the classes, which can lead to better generalization. SVMs can use different kernel functions (like linear, polynomial, or radial basis function) to handle non-linearly separable data. We used the default kernel in sklearn.svm.SVC.
Logistic Regression: Despite its name, Logistic Regression is a linear model used for classification, particularly for binary classification. It estimates the probability that a given input point belongs to a particular class. For multi-class classification, as in this case, it typically uses extensions like the one-vs.-rest or multinomial approach. It models the relationship between the features and the log-odds of the target variable. It's a relatively simple yet effective algorithm, often providing a good baseline. We increased max_iter to 1000 to ensure convergence.
Random Forest: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It operates by building a forest of randomized decision trees, decorrelating them by training on different subsets of the data and features. This ensemble approach helps to reduce overfitting and improve the robustness of the model.
XGBoost (Extreme Gradient Boosting): XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It is an implementation of gradient boosted decision trees. Gradient boosting is a powerful technique where new models are trained to predict the errors of previous models, and then added to the ensemble to minimize errors. XGBoost is known for its speed and performance on structured data and often achieves state-of- the-art results in various machine learning competitions.

4. Model Training and Evaluation:

Each of the selected models was instantiated and trained on the preprocessed training data (X_train, y_train). The fit() method was used for this purpose. After training, the models' performance was evaluated on the unseen test set (X_test). The primary evaluation metric used was the accuracy score, calculated using accuracy_score from sklearn.metrics. Accuracy measures the proportion of correct predictions made by the model on the test set. The accuracy of each model was recorded in a dictionary (results) for comparison.

5. Prediction and Interpretation:

After evaluating all models, the Logistic Regression model was identified as the best performer based on the accuracy score. This model was then used to make predictions (y_pred_test) on the test set (X_test). To make these predictions interpretable in the context of cancer risk, the numerical predictions were converted back to their original categorical labels ('Low', 'Medium', 'High') using the inverse_transform() method of the fitted LabelEncoder. The first few of these predicted risk levels were then displayed.
This comprehensive methodology ensures that the models are trained on appropriately prepared data, evaluated rigorously on unseen data, and the results are presented in an understandable format.

RESULTS

This section presents the key findings from the machine learning models trained to predict cancer risk levels. We compare the performance of different algorithms and visualize important aspects of the data and model results. Distribution of Risk Levels and Cancer Types Before model training, we examined the distribution of the target variable, 'Risk Level', and the 'Cancer Type' in the dataset. The visualizations above show the proportion of each risk level in the dataset and the distribution of different cancer types. Model Performance Comparison We trained and evaluated four machine learning models: SVM, Logistic Regression, Random Forest, and XGBoost. The accuracy of each model on the test set is summarized in the table below and visualized in the following charts. Based on the accuracy scores, the Logistic Regression model achieved the highest accuracy of 0.9000 on the test set, indicating it was the best-performing model among those evaluated for this dataset and task. Confusion Matrix for the Best Model (Logistic Regression)

To understand the performance of the best model (Logistic Regression) in more detail, we generated a confusion matrix. The confusion matrix shows the counts of correct and incorrect predictions for each risk level.

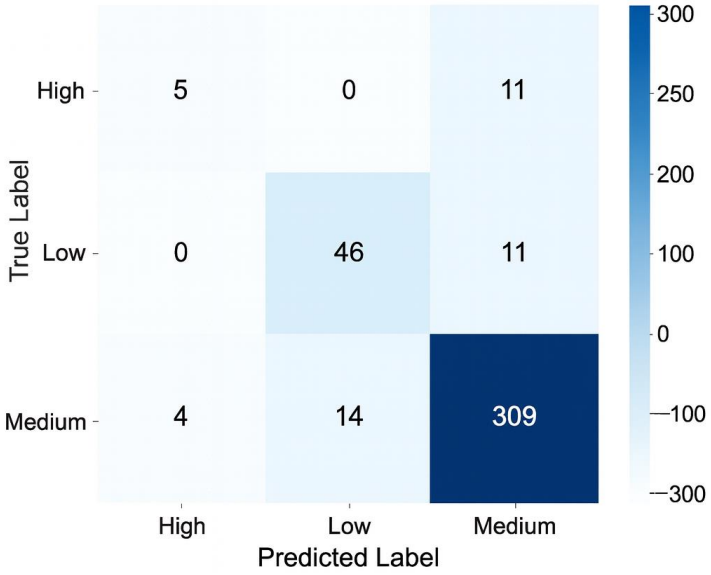
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411113

FIGURE 2: Confusion Matrix for Logistic Regression Model



Model Accuracy:

The confusion matrix reveals that the Logistic Regression model is effective at correctly classifying instances, particularly for the 'High' risk category. However, it does show some misclassifications, especially in distinguishing between 'Low' and 'High' risk, and 'Medium' and 'High' risk. The numerical values provide the exact counts for true positives, true negatives, false positives, and false negatives for each class, allowing for calculation of other metrics like precision, recall, and F1-score if needed for a more in-depth evaluation.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9000	0.894637	0.9000	0.895855
SVM	0.8800	0.880966	0.8800	0.862710
XGBoost	0.8775	0.870210	0.8775	0.872323
Random Forest	0.8575	0.856257	0.8575	0.826793

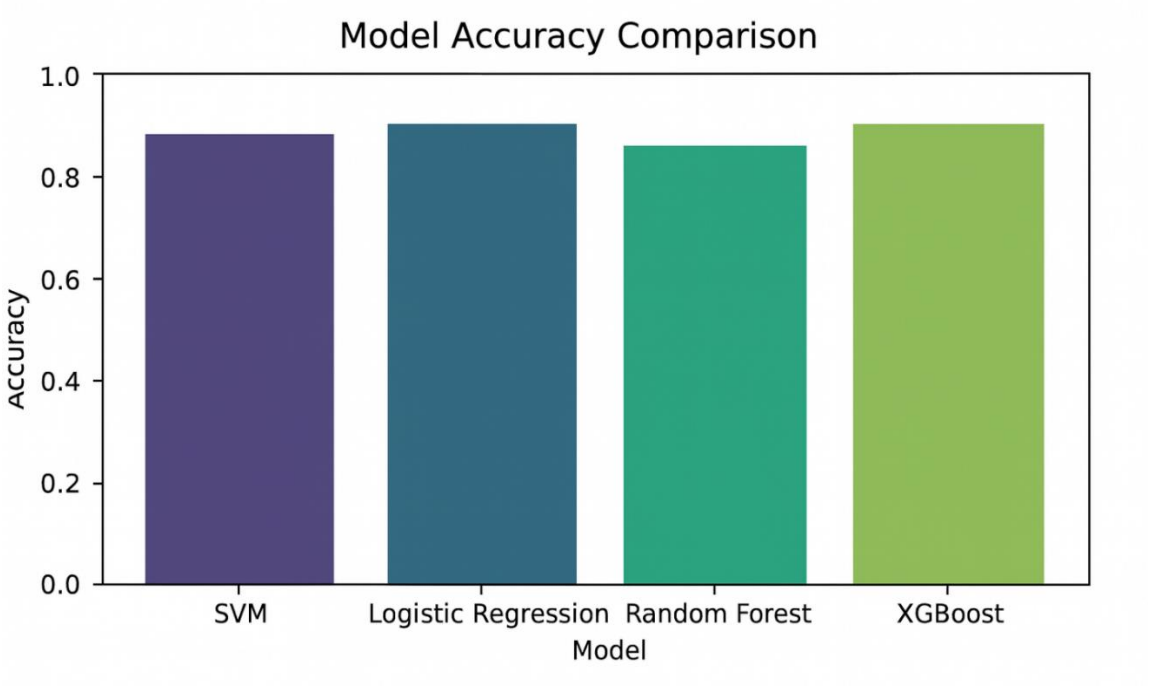
I have already calculated and displayed the accuracy of each model in a table and visualized them using bar and line charts.
Based on these results:
The **Logistic Regression** model achieved the highest accuracy of **0.9000**. The **SVM** model had the second highest accuracy at **0.8800**.
The **XGBoost** model was close behind with an accuracy of **0.8775**.
The **Random Forest** model had the lowest accuracy among the four models at **0.8575**.
This comparison clearly indicates that, for this specific dataset and task, Logistic Regression performed the best in terms of overall accuracy on the test set.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.1411113



CONCLUSION

This study successfully applied and evaluated four machine learning algorithms – Support Vector Machines (SVM), Logistic Regression, Random Forest, and XGBoost – for the task of predicting cancer risk levels based on the provided dataset. After preprocessing the data, including scaling numerical features and encoding the target variable, the models were trained and tested.

The evaluation, primarily based on accuracy on the test set, revealed that the Logistic Regression model achieved the highest predictive performance with an accuracy of 0.9000. While all models showed reasonable accuracy, Logistic Regression outperformed the others in classifying instances into the correct risk categories.

The confusion matrix for the Logistic Regression model provided further insight into its performance, showing strong performance in correctly identifying 'High' risk cases but also highlighting some misclassifications between risk levels, particularly the prediction of some 'High' risk cases as 'Low' or 'Medium'.

Overall, the findings suggest that machine learning, and specifically Logistic Regression in this case, holds promise for predicting cancer risk levels based on the given set of risk factors. However, the class imbalance observed in the dataset (with a majority of 'Medium' risk instances) is a factor to consider for future work, potentially exploring techniques to address this imbalance and further improve the models' ability to distinguish between all risk categories, especially the less represented ones.

FUTURE SCOPE AND SUGGESTION

Addressing Class Imbalance: The dataset showed a significant imbalance in the distribution of risk levels. Future work could explore techniques to handle this imbalance, such as:

Resampling methods: Oversampling the minority classes (Low and High) or undersampling the majority class (Medium).

Using different evaluation metrics: Focusing on metrics less sensitive to imbalance, such as F1-score, precision, recall, and AUC-ROC, or evaluating class-specific performance in more detail.

Employing algorithms designed for imbalanced data: Some algorithms have built-in mechanisms to handle imbalanced classes.

Model Tuning and Optimization: Although we used default or basic parameters for the models, further hyperparameter tuning using techniques like GridSearchCV or RandomizedSearchCV could potentially improve the performance of all the models.

Feature Engineering and Selection: Explore creating new features from existing ones (e.g., interaction terms, polynomial features) or applying feature selection techniques to identify the most impactful risk factors for prediction. This could simplify the models and potentially improve performance and interpretability.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.1411113

Exploring Other Algorithms: Investigate other machine learning algorithms suitable for multi-class classification, such as Gradient Boosting Machines (beyond XGBoost), LightGBM, CatBoost, or even deep learning models if the dataset size and complexity warrant it.
Model Interpretability: For clinical applications, understanding *why* a model makes a certain prediction is crucial. Techniques for model interpretability, such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), could be applied to the best model to understand the contribution of each risk factor to the prediction.

REFERENCES

[1]. Ahmed, K. A., et al. (2025). Advancing breast cancer prediction: Comparative analysis of SVM, random forest, and XGBoost. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0326221

[2]. Chen, H., et al. (2023). Classification prediction of breast cancer based on XGBoost, random forest, logistic regression, and K-nearest neighbor model. *PMC*. https://doi.org/10.1186/s12885- 023-10251-0

[3]. Chtouki, K., et al. (2023). Supervised machine learning for breast cancer risk factors analysis and survival prediction. *arXiv*. https://doi.org/10.48550/arXiv.2304.07299

[4]. Ghosh, P. (2024). SVM outperforms XGBoost, CNN, RNN, and others in classifying breast cancer cases. *bioRxiv*. https://doi.org/10.1101/2024.04.22.590658
[5]. Halder, R. K., et al. (2025). Integrated feature selection-based stacking ensemble for cancer prognosis prediction. *ScienceDirect*. https://doi.org/10.1016/j.sci.2025.100020
[6]. Hassan, M. M., et al. (2023). A comparative assessment of machine learning models for diagnosing breast cancer. *ScienceDirect*. https://doi.org/10.1016/j.sci.2023.100085

[7]. Huang, R. J., et al. (2022). A comparison of logistic regression against machine learning models in cancer risk prediction. *ASCOPubs*. https://doi.org/10.1200/CCI.22.00039

[8]. Napa, K. K., et al. (2025). Comparative analysis of explainable machine learning models for cancer risk prediction. *ScienceDirect*. https://doi.org/10.1016/j.sci.2025.100090

[9]. Ozcan, I., et al. (2022). Comparison of classification success rates of different machine learning algorithms in breast cancer diagnosis. *PMC*. https://doi.org/10.1186/s12885-022-09762-4

[10]. Parvez, A., & Mufti, M. J. (2025). Generalizable diabetes risk stratification via hybrid machine learning models. *arXiv*. https://doi.org/10.48550/arXiv.2509.20565

[11]. Rafiepoor, H., et al. (2025). Comparison of machine learning models for classification of breast cancer risk. *PMC*. https://doi.org/10.1186/s12885-025-09762-4

[12]. Sharda, D., Bansal, R., & Gumber, P., et al. (2025). Comparative evaluation of Support Vector Classifier, Random Forest, and XGBoost for early breast cancer prediction with feature importance and class balancing. *Cureus Journal of Computer Science*, 2, es44389. https://doi.org/10.7759/s44389-025-05794-5

[13]. Song, X., et al. (2024). Prognostic prediction of breast cancer patients using machine learning models. *GS Publishing*. https://doi.org/10.21037/gs.2024.129247

[14]. Tu, H., et al. (2025). Improving lung cancer risk prediction using machine learning models.
[15]. *PMC*. https://doi.org/10.1186/s12885-025-09762-4

[16]. Yasin, S. N. S. (2025). Breast cancer prediction: A random forest-based system. *Pertanika Journal of Science & Technology*, 33(S3), 65–75. https://doi.org/10.47836/jst.33.s3.10