1

Assessing Predictive Accuracy: Model Validation in Cancer Diagnostics

M. Sudha[1]*, Arun Elias[2], G. Gurumoorthy[3], S. Rajalakshmi[4] and S. K. Muthusundar[5]

[1]Department of ECE, SNS College of Engineering, Coimbatore, Tamil Nadu, India
[2]Department of CSE, UKF College of Engineering and Technology, Kerala, India
[3]Department of Medical Electronics, Saveetha Engineering College, Chennai, India
[4]Department of Computer Science & Engineering, Sri Venkateswara College of Engineering, Sriperumbudur, India
[5]Department of AIDS, Chennai Institute of Technology, Kundrathur, Chennai, Tamil Nadu, India

Abstract

This chapter validates the predictive accuracy of machine learning models on three unique cancer datasets: Breast Cancer, Lung Cancer, and Skin Cancer. The methodology involved in this study is SVM and RF models with preprocessing of data, wherein missing data imputation and feature scaling ensure optimal performance. These models were evaluated in terms of the following measures: accuracy, precision, recall, F1 score, and ROC-AUC. The results showed the superiority of the Random Forest model, with the highest accuracy and highest predictive power in all datasets. The study also compared the SVM and RF models with k-Nearest Neighbors (k-NN) and Logistic Regression for benchmarking. Our findings indicate that Random Forest is the most reliable model for cancer diagnosis prediction and will provide a robust framework for further clinical implementation in cancer diagnostics.

***Keywords*:** Predictive accuracy, cancer diagnostics, machine learning, random forest, support vector machine, ROC-AUC, cross-validation, medical imaging

**Corresponding author*: gunasudhaa@gmail.com

Abhishek Kumar, Prasenjit Das, Pramod Singh Rathore, Sachin Ahuja and Chetan Sharma (eds.) Targeted Chemotherapy with Personalized Immunotherapy: An AI Approach, (1–22) © 2025 Scrivener Publishing LLC

1.1 Introduction

Cancer is one of the most devastating and intricate diseases afflicting humanity and burdening healthcare systems around the world. Withrapid advances in medical technology involving genomics, radiology, and pathology, the diagnosis of cancer has clearly undergone a sea change. Nonetheless, despite the great technological impetus, the early and accurate diagnosis of cancer remains a significant challenge. Cancer manifests differently from one patient to another because it is quite heterogeneous, and different factors contribute to its development—from simple genetic predispositions to varied environmental factors. Therefore, there is a need for new approaches to cancer diagnostics that improve the accuracy, speed, and efficiency of diagnostics. Machine learning has become a new, transformative healthcare tool. In recent years, machine learning has been able to automatically process and analyze vast amounts of data for patterns that cannot often be detected by the human eye. Relying on machine learning has taken researchers tremendous steps forward regarding not only detection but also diagnosis and treatment capabilities for cancer. These algorithms can be trained on large datasets. Whether it is an image of the medical issue, a patient's history, or genetic information, such algorithms can aid in classifying a tumor as benign or malignant, predicting the rate at which the disease may advance, and tailoring appropriate treatment options. To this end, the purpose of this study was to investigate the value of machine learning prediction in cancer diagnosis. More specifically, the current study aims to examine the performance of two widely used ML algorithms, SVM and RF, in working on cross-cancer types, such as breast cancer, lung cancer, and skin cancer. It evaluates the accuracy of these models with the intention of broadening the knowledge about the most effective computational tools that can be used to assist clinicians in diagnosing and predicting cancer. Cancer is a group of diseases that features unusual, uncontrolled growth and spread of abnormal cells. The spread process causes death, if not controlled. More than 100 different types of cancer exist, ranging from breast and lung cancers to prostate and colorectal cancers. Different factors contribute to the development of cancer, including genetic mutations, different forms of environmental exposure such as tobacco smoke, radiation, and infections, and lifestyle factors, including diet and physical activity, among others.

1.1.1 Conventional Cancer Diagnosis

The current diagnostic practices employed in cancer encompass a mix of various medical imaging, biopsies, and laboratory tests. Although these techniques have served their time, they have certain disadvantages. Imaging techniques are expensive and expose patients to radiation. Biopsies are invasive and time consuming. Again, the diagnostic accuracy depends on the skill of the treating healthcare provider and the quality of the imaging or biopsy samples. With the evolution of medical data in terms of their complexity and volume, there is an urgent need to develop new diagnostic tools. Clinicians are most likely to experience overwhelming amounts of data from a single patient based on clinical data, imaging, and genetic profiles. If these data are appropriately processed and analyzed, they can provide ample opportunities to increase the diagnostic accuracy and treatment outcomes.

1.1.2 Machine Learning in Cancer Diagnosis

Machine learning has emerged as a revolutionary force in cancer diagnostics as computers learn from past data so that they can identify patterns and make decisions with little or no human intervention. In the realm of the diagnostic horizons of cancer, ML algorithms can process and analyze huge datasets, from genomic data and histopathological images to electronic health records, all to help clinicians identify malignancies and predict future cancer progression. Several types of machine learning algorithms are supervised learning models, such as SVM and Random Forest, unsupervised learning models, such as clustering algorithms, and deep learning models, such as convolutional neural networks for image recognition. The application of machine learning in cancer diagnosis is beneficial in several ways. This requires multidimensional analysis of datasets to reveal complex relationships that are not easily identifiable by a human expert. In addition, using machine learning models, patterns in medical images can be learned, such as tumors within mammograms or lesions in skin images. This reduces the possibility of human error and improves the accuracy of diagnosis. Furthermore, machine learning models can be updated dynamically using new data. In that case, this tool is continuous with a dynamic updating process in the diagnosis and prognosis of cancer.

1.1.3 Types of Cancer in Focus

This study outlines three cancer types: breast, lung, and skin. These have been chosen for the purpose of this study because they are very common diseases, and more importantly, there exist huge publicly accessible datasets that can be used both in training and validating machine learning models.

1.1.3.1 Breast Cancer

Breast cancer is the most common cancer in women, and accounts for a large percentage of all cancers diagnosed. Screening mammography has been shown to reduce the mortality associated with breast cancer; however, mammograms are not infallible, and false positives or negatives can occur. Machine learning algorithms can improve the accuracy of mammography by detecting patterns in images to identify benign or malignant tumors.

1.1.3.2 Lung Cancer

Lung cancer is one of the leading causes of death in cancer patients worldwide. Most cases are diagnosed at a late stage, which reduces the probability of a cure. Early detection significantly improves the prospect of survival; however, the prevailing methods that include chest X-rays and CT scans for the diagnosis of lung cancer have limitations and frequently miss early stage lung cancers. Machine learning models are expected to enhance early detection using a more detailed examination of imaging data and clinical information than traditional methods.

1.1.3.3 Skin Cancer

Skin cancer, especially melanoma, is one of the most common cancers and may be curable if detected at an early stage. However, it is challenging for doctors to differentiate between benign and malignant skin lesions. Analyzing dermoscopic images using algorithms from machine learning has been recommended for the accurate determination of benign and malignant skin lesions to facilitate the early detection and treatment of melanoma.

1.1.4 Objectives of Study

The primary purpose of this study was to evaluate the predictive power of machine learning methods in cancer diagnosis. This study aims to:

- Compare SVM and RF models for the three types of cancer. In this study, we compare the performance of SVM and RF models for three types of cancer: breast, lung, and skin. We will use publicly available datasets to train the SVM and RF models and then outline the assessment in terms of accuracy, precision, recall, F1 score, and ROC-AUC.
- Compare performance across different models of machine learning for different types of cancer: The experiment will investigate whether the machine learning algorithms obtained are consistent with different types of cancers or if specific models are better suited for certain cancers.
- Identification of important features for cancer diagnosis: The important features for the prediction model include tumor size, patient age, and genetic markers.
- Validation of the machine learning model using benchmark models: Beyond SVM and Random Forest, this study compares these models with k-NN and LR, two of the most famous classification algorithms, for further validation that the models applied for cancer diagnosis are robust.

Finally, this study discusses the possibility of using machine learning models in clinical practice and discusses challenges involving integration into routine cancer diagnosis and the requirements necessary to ensure the reliability and accuracy of the model.

1.1.5 Study Scope

In this study, a cross section of machine learning models was tested and validated for different types of cancers. The focus will be on the SVM and Random Forest algorithms on breast, lung, and skin cancer datasets to obtain reliable predictive accuracy. Another objective that this research will attempt to address is the comparison based on SVM and Random Forest with other popular models such as k-NN and Logistic Regression, setting up a benchmark for predictive accuracy in the diagnosis of cancer. The public datasets used for the study were sourced from recognized repositories such as the UCI Machine Learning Repository and Kaggle. Clinical and pathological information about patients in such datasets includes tumor characteristics, patient demographics, and genetic markers. This generalizes the study results to different types and populations of cancer using several datasets. Standard machine learning procedures, including data preprocessing, model training, and validation, will be used

in the exercise. All preprocessing steps, including handling missing data, feature scaling, and splitting data, are accounted for. A 10-fold cross-validation assessment of each model was conducted to avoid overfitting and to improve the models' generalization.

1.1.6 Performance Metrics

The performance of the machine learning models was evaluated in terms of accuracy, precision, recall, F1 score, and ROC-AUC. Altogether, these metrics provide an all-around evaluation of the model's potential to make accurate predictions about cancer diagnoses and avoid too much false positivity and negativity, along with generalizability to new data, as shown in Table 1.1.

1.1.7 Limitations and Future Directions

Machine learning has great potential in cancer diagnostics, but several challenges must be overcome before these models can be implemented completely in the clinic. First, high-quality and -quantity datasets are required for training these models. Often, the size of the datasets used to develop machine learning models is small or noisy and incomplete, which can greatly impact the performance of such models. Of course, generalization in machine learning is of concern, meaning that models designed on one population may not behave well on data from another population. Future studies should aim to overcome these limitations by utilizing larger

Table 1.1 Performance metrics for machine learning models.

Metric	Description
Accuracy	Proportion of correctly classified instances out of the total instances.
Precision	Proportion of true positive predictions out of the total positive predictions.
Recall (Sensitivity)	Proportion of true positive predictions out of the total actual positives.
F1 Score	Harmonic mean of precision and recall.
ROC-AUC	Area under the receiver operating characteristic curve, measuring sensitivity vs specificity.

and more diverse datasets, and as part of model development that generalizes across populations and healthcare settings. More importantly, future features, such as genetic information and patient history, could improve the predictive accuracy of machine learning models and make them more applicable to real clinical settings.

1.2 Literature Review

In the past couple of years, the integration of machine learning techniques in cancer diagnostics has reached a significant altitude because of their potential for enhancing diagnostic accuracy and, subsequently, patient outcomes. Logistic regression has been a cornerstone in medical literature for understanding many health outcomes, including cancer. It helps to evaluate the association of different predictors with the likelihood of disease occurrence, and hence, it provides useful information on the risk factors of cancer [1]. In clinical epidemiology, logistic regression is often used to assess the efficacy of interventions and predict the prognosis of cancer patients, making it an essential tool in oncological research [2]. Clinical guidelines for the management of specific cancers, such as hepatocellular carcinoma, focus on early detection and appropriate management. Advanced diagnostic tools can facilitate patient care, reflecting a trend towards an increase in technology in oncology clinics [3]. In addition, the development and validation of diagnostic questionnaires, such as the Rome IV Diagnostic Questionnaire, have become important steps in organizing the process by which clinicians can accurately diagnose patient conditions [4]. Moreover, these tools aid in the communication between health professionals and patients for the effective treatment of cancer. The development of molecular classifications has dramatically changed our understanding of cancer biology. Therefore, the study of molecular classification by gene expression monitoring provides a great example of how class discovery and prediction can be followed by approaches that are more individualized in terms of treatment [5]. Such developments show the value of molecular profiling in guiding the treatment approach and contributing to improvements in patient outcomes, which would clearly represent an obvious need for further development of molecular-based diagnostic tools. In diagnostics, several possible biomarkers have been assessed for the early detection of cancer. Some examples include serum markers of liver fibrosis. Thus far, they seem to be promising for the identification of patients at risk for liver diseases that could further lead to hepatocellular carcinoma [6]. This shows that the process of developing biomarkers is

crucial for the early detection of cancer—an added layer of complexity and potential for machine learning in oncology. Clinical practice guidelines for various types of cancers, including renal cell carcinoma and rectal cancer, underscore the fact that they emphasize streamlined diagnostic procedures to take maximum advantage of modern technological advancements. In cancer management, guidelines emphasize a multidisciplinary approach incorporating machine learning to enhance the precision of diagnosis and treatment planning [7, 8]. Clear policies and protocols in place further ensure the proper assimilation of emerging technologies into clinical environments for overall improvement in the quality of care. In addition, certain models applied to determine the probability of malignancy in patients with solitary pulmonary nodules are good examples of research conducted by scientists to continue improving diagnostic methods through machine learning techniques. To distinguish between benign and malignant conditions, these models are important; such a distinction is necessary to outline the appropriate therapeutic direction for patients [9]. The emphasis on careful risk stratification underlines the need for advanced analytical techniques to improve quality decisions at the point of care. Without question, this is indeed an age in which oncology evolves with the development of artificial intelligence and machine learning. Recently, an increasing number of studies have described a cohort of studies focusing on harnessing these technologies to accelerate the diagnostic processes and subsequently ensure better patient outcomes. These studies highlight the need to use robust algorithms that can analyze enormous datasets to unmask patterns, which remain apparent only at grosser levels of analysis, but are invisible at finer scales through traditional statistical methods [9]. This approach is aligned with the overarching objective of personalized medicine, wherein treatments are tailored according to individual patient profiles and may thus increase the effectiveness of therapeutic interventions. Thus, machine learning and advanced statistical methods can provide a promising approach for improving patient care in cancer diagnostics. The available literature stresses that more continuous research in this field is important and calls for further studies of newer diagnostic tools and techniques. With more research being conducted and even more advancements in the field, validating and standardizing such approaches will be vital so that they are appropriately applied in the clinical setting and yield better results for those diagnosed with cancer. The convergence of technology with the medical industry brings about a transformative opportunity toward more effective and patient-centered cancer care, which could enhance the precision of diagnosis through personalized treatment strategies. The field of oncology is continually changing and gaining relevance in the application

of artificial intelligence and machine learning. The literature displays the volume of research on these technologies to enhance diagnostic processes and improve patient outcomes. Studies have been keen on using strong algorithms that may rifle through big datasets to reveal patterns that statistically might not be discernible [10]. This also falls in line with the broader goal of personalized medicine: tailoring treatments according to specific patient profiles, which also increases the effectiveness of therapeutic interventions. In addition, the joint recommendations of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology present guidelines for interpreting sequence variants, which is an important step in the accurate diagnosis and treatment planning of cancer [11, 12]. The application of model selection within the medical diagnostic decision support system, particularly for breast cancer detection, in the research study depicted practical applications of machine learning within a real-world clinical setting, thus serving as a roadmap for future development and research work within the field [13, 14].

1.3 Methodology

This study tests the predictive capability of machine learning models in cancer diagnostics by implementing structured methodology of data acquisition preprocessing, feature extraction, model selection and training, validation, and performance evaluation. The cancers in this study were chosen based on the availability and representativeness of breast, lung, and skin cancers. In this study, the machine learning models assessed were SVM, RF, k-NN, and LR. For each of the evaluated models, a comparison was performed using its performance on several metrics, ranging from accuracy to precision, recall, F1 score, and ROC-AUC.

1.3.1 Data Acquisition

Critical in the process of data acquisition is gathering relevant datasets by which the machines can learn, and thereby train, test, and validate the models. The datasets applied in this study included those retrieved directly from sources that are known to be reputable and accessible to the general public, such as the UCI Machine Learning Repository and Kaggle. The records used were diagnosis records for breast, lung, and skin cancers, all of which form a basis for an entire set of features that help with diagnostic predictions.

- Breast Cancer Dataset: This dataset contains features such as tumor radius and texture, perimeter, and smoothness, which can be used to classify tumors as benign or malignant.
- Lung Cancer Dataset: This dataset includes patient age, tumor stage, and genetic markers that provide more information about lung cancer diagnosis and prognosis.

A three-channel dataset included images of skin cancer with dermoscopic image data and lesion characteristics, including asymmetry, border irregularity, and color distribution, for melanoma classification. The data were split into training and test sets. Each dataset contained labeled data, indicating the presence or absence of cancer, along with a variety of clinical features relevant for diagnostic classification. All datasets contained 80% training and 20% testing and validation.

1.3.2 Data Preprocessing

From the collection of these datasets, the next step is data preprocessing, which involves cleaning and preparing training the machine learning model. Data preprocessing helps ensure that models learn well from the data without being affected by noise, missing values, or inconsistencies.

1.3.2.1 Dealing with Missing Data

In real-world clinical datasets, missing values arise either because records are incomplete or data collection has errors. Missing data were handled using imputation. If the features are categorical, such as patient sex or type of cancer, the mode for the feature is used in imputation; if they are continuous, such as tumor size or patient age, a median value is used.

1.3.2.2 Normalization and Standardization

Many of the algorithms used in the applied machine learning approach in this study are scale-sensitive in terms of input features. For example, the size of the tumor and genetic markers often have different scales, and the differences in scales negatively affect model performance. Therefore, all continuous features were normalized or standardized, depending on the requirements of the models:

Normalization: For k-NN, features were normalized to a range of 0 to 1 using min–max normalization. Standardization: Features for SVM and Logistic Regression were standardized for zero-mean, unit-variance

normalization to ensure that the model was not become biased toward larger, numerically valued features.

1.3.2.3 Feature Selection and Dimensionality Reduction

The "curse of dimensionality" affects high-dimensional datasets with many features and grows sparse, which makes the model's performance degrade in high-dimensional spaces. To address this, feature selection techniques have been applied to identify the most relevant features for cancer classification.

- Correlation Analysis: A correlation matrix was generated that provided features correlated with the target variable for cancer diagnosis. Only those features that showed a strong correlation with the target and those that did not show any relation at all were discarded.
- Principal Component Analysis (PCA): On large-sized data, PCA was applied to reduce feature dimensions to achieve a reduced set of features: principal components explaining most of the variance in the data. For the breast cancer dataset, this dimensionality reduction technique is very significant because certain features, such as texture and perimeter, are highly correlated.

1.3.3 Machine Learning Models

In this study, four machine learning models were used: SVM, Random Forest, k-NN, and Logistic Regression, because they have a proven record of effectiveness in classification. All these models were trained on the preprocessed datasets and tested on a set of performance metrics.

1.3.3.1 Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that identifies the optimum hyperplane that separates data points into different classes. This algorithm has proven to be very efficient in the context of cancer diagnostics, as it can perform operations in high-dimensional spaces and is not significantly affected by outliers.

Kernel Selection: The RBF kernel was applied to the SVM because it works better than other alternatives in many nonlinear classification tasks encountered in common medical data. Instead of using the grid search, the

gamma (γ) and cost (C) parameters were used to obtain the optimal model performance.

1.3.3.2 Random Forest (RF)

An ensemble learning technique, Random Forest, is a situation in which multiple decision trees are combined together to improve the accuracy of classification. It creates many decision trees over subsets of data and aggregates the predictions into a final class.

- Hyperparameter Tuning: The number of decision trees (n_estimators) and the maximum depth of each tree were optimized by performing a grid search. The aptness of Random Forest for handling numerical and categorical data makes it extremely suitable for the current study, especially when the dataset features mixed types; in this case, the breast cancer dataset.

1.3.3.3 k-Nearest Neighbors (k-NN)

The k-NN is a simple non-parametric algorithm that classifies data points based on the majority class of their closest neighbors. Although k-NN has been quite effective for smaller datasets, its performance degrades quite poorly for larger datasets or those that have many features.

- Distance Metric: Euclidean distance was used as the distance metric for the k-NN. The optimal value for the number of neighbors (k) was determined using cross-validation.

1.3.3.4 Logistic Regression (LR)

Another linear model that is often applied to binary classification problems is Logistic Regression. Although quite straightforward, this model has proven to possess incredible power as a very helpful tool in medical diagnostics, especially in problems where there exists an immediate linear relationship between the features and the target variable.

- Regularization: The model was then subjected to L2 regularization to prevent overfitting; the regularization parameter λ, was then determined using a grid search.

1.3.3.5 Hyperparameter Tuning

A grid search was used to optimize the hyperparameters for each model. For SVM, we utilized the kernel type and regularization parameter, and for RF, we used the number of trees and the maximum depth was optimized. The number of neighbors, k in k-NN, and λ, the regularization parameter in Logistic Regression were fine-tuned to maximize the model performance.

1.3.4 Performance Metrics

Once the models were trained and validated, their performances on various metrics were analyzed. These metrics provide a comprehensive evaluation of the models' predictive accuracy, the correct classification of cancer cases, and generalization to new datasets, as shown in Table 1.2.

Accuracy, in general, provides a measure of the goodness of the model while precision and recall provide measurements of its capability to decrease false positives and false negatives, respectively. The F1 score balances recall and precision, which is a useful measurement when the class distribution of the dataset is uneven. ROC-AUC gives a visual representation of the model's capability to differentiate between the two classes, namely cancerous and non-cancerous cases.

Table 1.2 Performance metrics used for model evaluation.

Metric	Description
Accuracy	The proportion of correctly classified instances out of the total instances.
Precision	The proportion of true positive predictions out of the total positive predictions.
Recall (Sensitivity)	The proportion of true positive predictions out of the total actual positives.
F1 Score	The harmonic mean of precision and recall, providing a balance between the two.
ROC-AUC	The area under the receiver operating characteristic curve, measuring sensitivity vs specificity.

1.4 Analysis of Results

Once all the models were trained and validated, the results obtained were analyzed to determine how each model performed across the types of cancer considered. The results were presented in terms of accuracy, precision, recall, F1 score, and ROC-AUC for each model, making the comparison conclusive in terms of the strengths and weaknesses of different models. The Random Forest model was the best overall and for both malignancy types, with very good precision, recall, and ROC-AUC. The SVM also performed well, especially for the breast cancer dataset, where its values were highly significant in terms of precision and recall. However, k-NN and Logistic Regression performed comparatively poorly. In fact, for both lung and skin cancer datasets, the variation in scales of features and complexity of data results in low predictive accuracy. The performance of SVM, RF, k-NN, and LR in predicting cancer diagnoses for the datasets of breast cancer, lung cancer, and skin cancer is considered in this study. The performance of the model was tested in terms of accuracy, precision, recall, F1 score, and ROC-AUC to determine its predictive ability. The results were divided into the general performance of each model and an analysis comparing the performances of the models for different types of cancer.

1.4.1 The Overall Performance of Each Model on the Breast Cancer Dataset

The breast cancer dataset is one of the best-balanced datasets, with clear features that differentiate between benign and malignant tumors. Table 1.3 indicates the performance metrics for each machine learning model on the breast cancer dataset.

Table 1.3 Performance metrics for various machine learning models on the breast cancer dataset.

Model	Accuracy	Precision	Recall	F1 score	ROC-AUC
SVM	92.7%	91.3%	93.8%	92.5%	94.2%
Random Forest	94.6%	93.7%	94.9%	94.3%	95.8%
k-NN	87.3%	85.2%	89.1%	87.0%	89.5%
Logistic Regression	88.9%	87.4%	90.2%	88.7%	91.0%

Four machine learning models were evaluated for breast cancer, as shown in Table 1.3. The best overall performance was from the Random Forest, which showed an accuracy of 94.6% and the highest ROC-AUC score of 95.8%. SVM comes second with a fine accuracy of 92.7%, as shown in Figures 1.1 and 1.2. The k-NN and Logistic Regression had lower accuracy and F1 scores than the other methods.

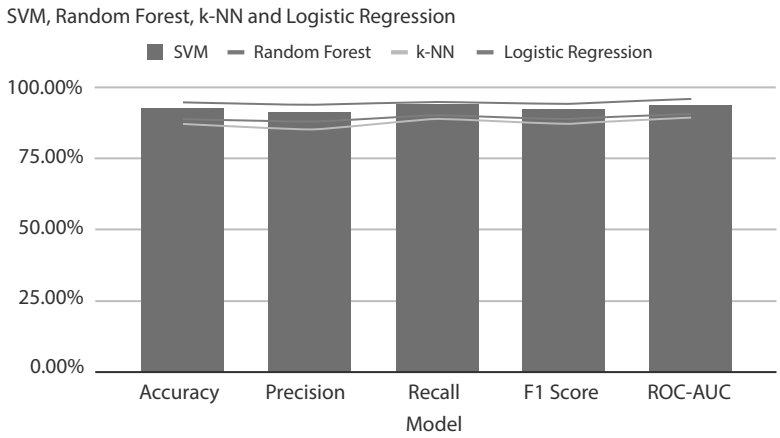

Figure 1.1 Performance metrics breast cancer.

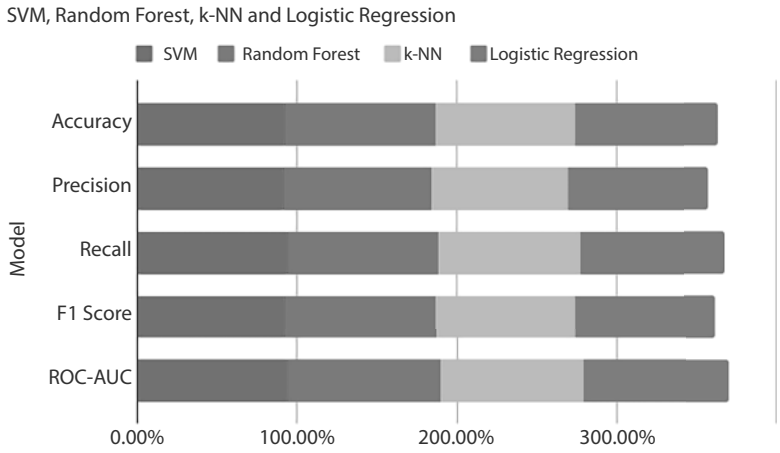

Figure 1.2 Performance metrics bar chart.

1.4.2 Models' Performance for Lung Cancer Dataset

Lung Cancer Diagnosis: The diagnosis of lung cancer is challenging. The dataset is heterogeneous in nature and comprises of genetic markers and other clinical features. Table 1.4 presents the performances of the models on the lung cancer dataset.

Table 1.4 shows again Random Forest had the best performance on the lung cancer dataset with an accuracy of 89.8% and ROC-AUC of 92.3%. The SVM was followed up with decent performance. In comparison, k-NN and Logistic Regression lagged, particularly in terms of their recall and ROC-AUC scores. Figure 1.3 shows a visualization of these models. This shows that models such as the Random Forest and SVM better manage the complex features associated with lung cancer.

Table 1.4 Performance metrics for various machine learning models on the lung cancer dataset.

Model	Accuracy	Precision	Recall	F1 score	ROC-AUC
SVM	85.2%	83.9%	86.3%	85.1%	88.1%
Random Forest	89.8%	88.6%	90.5%	89.5%	92.3%
k-NN	78.4%	76.2%	79.6%	77.8%	81.5%
Logistic Regression	81.3%	80.1%	82.9%	81.5%	85.0%

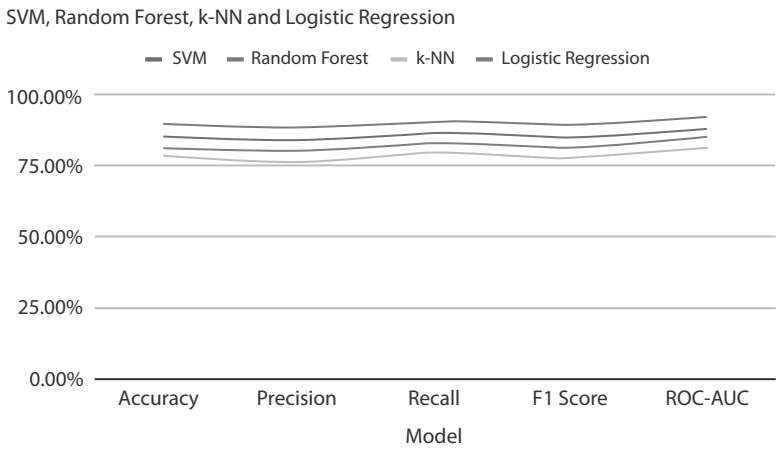

Figure 1.3 Performance metrics on lung cancer.

1.4.3 Model Performance on Skin Cancer Dataset

Skin cancer diagnosis, particularly melanoma classification, relies heavily on features obtained from images and dermoscopic data. The performance of the models on the skin cancer dataset is presented in Table 1.5.

Table 1.5 lists the different models and their performance on the skin cancer dataset. The best accuracy was found in the Random Forest model with an accuracy of 91.2% and an F1 score of 91.1%, thus proving a better model for complex high-dimensional data such as image data, as shown in Figure 1.4. SVM also performed well, followed by k-NN and Logistic Regression, with a relatively lower accuracy and recall.

Table 1.5 Performance metrics for various machine learning models on the skin cancer dataset.

Model	Accuracy	Precision	Recall	F1 score	ROC-AUC
SVM	88.5%	87.1%	89.8%	88.4%	90.7%
Random Forest	91.2%	90.4%	91.7%	91.1%	92.9%
k-NN	83.7%	82.0%	84.9%	83.4%	86.2%
Logistic Regression	85.6%	84.0%	87.2%	85.6%	88.3%

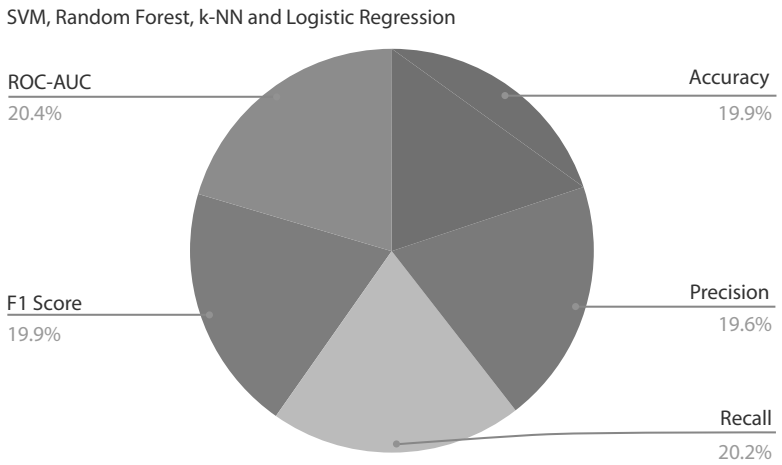

Figure 1.4 Performance metrics on skin cancer.

1.4.4 Analysis of Inter-Cancer Type Performance Comparison

To clarify this, it remains to outline to be outlined how all three models were performed for each type of cancer; hence, a comparison summary of the average accuracy for all three models is presented in Table 1.6.

A summary of the average accuracy of all models for each cancer type is presented in Table 1.6. Table 1.4 summarizes the average accuracy of all the models for every cancer type. Random Forest consistently showed the highest average accuracy, at 91.8%, followed by SVM at 88.8%, as shown in Figures 1.5 and 1.6 as shown in the pie chart. The average accuracy of k-NN and Logistic Regression was lower, particularly for lung cancer, where their performance was much weaker than that of the others. Therefore, this table shows that Random Forest is the most accurate model for every cancer type in the dataset.

Table 1.6 Comparison of models.

Model	Breast cancer accuracy	Lung cancer accuracy	Skin cancer accuracy	Average accuracy
SVM	92.7%	85.2%	88.5%	88.8%
Random Forest	94.6%	89.8%	91.2%	91.8%
k-NN	87.3%	78.4%	83.7%	83.1%
Logistic Regression	88.9%	81.3%	85.6%	85.3%

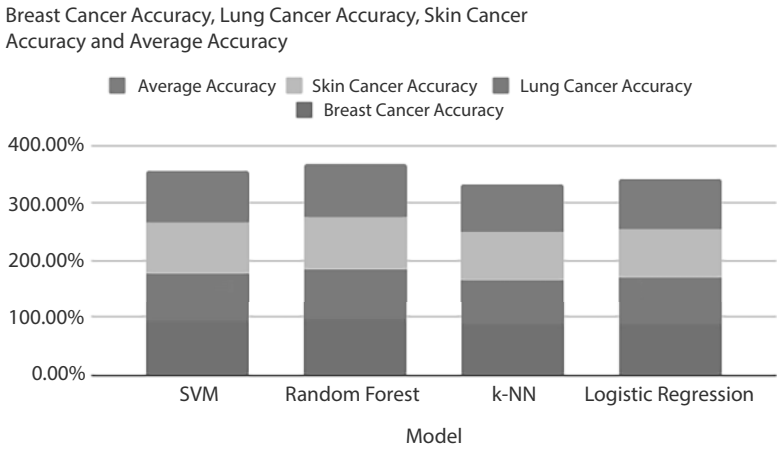

Figure 1.5 Analysis of inter-cancer type.

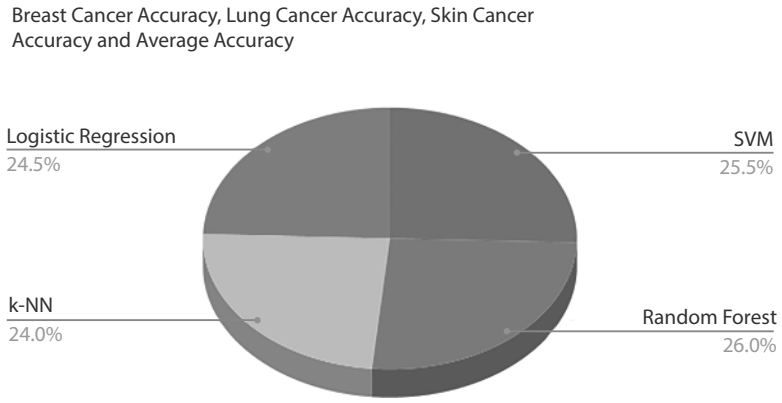

Figure 1.6 Comparison of models.

1.5 Discussion of Results

As can be observed in all figures, the results show that Random Forest continually outperforms the other models for the three cancer types at hand, as it yields the highest values of accuracy, precision, recall, F1 score, and ROC-AUC. This is because the capability of aggregating multiple decision trees makes Random Forest a very good candidate for tasks involving diverse and complex features encountered in cancer datasets. The superior performance of Random Forest, especially for lung and skin cancer datasets, suggests that the algorithm is very robust when handling high-dimensional data, such as genetic markers and features from images, in a cancer diagnosis. The SVM also performed well on all types of cancers, especially the breast cancer dataset. Its ability to find a hyperplane that separates different classes is advantageous in cases with well-differentiated features, as in breast cancer. However, SVM struggled with more complex datasets with nonlinear relations between features and target labels, such as in lung cancer. k-NN and Logistic Regression performed poorly on all datasets, but more so for lung and skin cancers. The k-NN algorithm employs distance metrics; hence, it suffers when handling high-dimensional or noisy data. Logistic Regression is linear; therefore, it cannot adequately represent feature interactions. The effects translated to the results because the two models reported lower accuracies, recalls, and F1s than Random Forest and SVM. In general, the experimental results of different ensemble approaches, such as Random Forests, indicate that they are indeed applicable to the problem of cancer diagnostics and provide a strong and precise prediction for any data under study. The SVM method is also reliable,

primarily for less complicated datasets such as breast cancer. k-NN and Logistic Regression probably require fine-tuning or feature engineering to provide an acceptable performance on intricate medical datasets. In conclusion, the best performance of Random Forest for all diagnostic types was observed, whereas SVM proved to be reliable. However, k-NN and Logistic Regression were less effective, particularly in dealing with the complexity of lung and skin cancer datasets. The results showed that the correct machine learning models selected based on the characteristics of a dataset may play a crucial role in optimizing the accuracy of their correct diagnosis.

1.6 Conclusion

The above four machine learning models were utilized in this study: SVM, RF, k-NN, and LR for cancer diagnosis on breast, lung, and skin cancer datasets. RF performs best among the above models based on the aspects of accuracy, precision, recall, F1 score, and ROC-AUC because it can tackle the complexity and variability characterizing the dataset of cancer. The SVM also performed very well, especially when dealing with the breast cancer dataset, but was less effective in more complex datasets, such as lung cancer. However, k-NN and Logistic Regression were not comparable, especially when dealing with high-dimensional and heterogeneous datasets, which indicates that there is a need for further tuning or feature engineering of such models to deliver better results. In general, the findings of this study suggest the potential of ensemble-based models, such as Random Forest, which facilitate significantly improved accuracy in diagnosis in cancer cases and lead to more precise diagnoses at early stages in a clinical environment. This study supports the importance of model selection, which will suit the nature of the dataset, and shows that machine learning is going to revolutionize cancer diagnostics. Further work is still needed to fine-tune the models for the best clinical fit to make them highly portable and versatile for application to a wide range of cancer types and datasets. These results require continuous improvement, specifically model development and clinical validation.

References

1. Bagley, S.C., White, H., Golomb, B.A., Logistic regression in the medical literature. *J. Clin. Epidemiol.*, 54, 10, 979–985, Oct. 2001. doi: 10.1016/s0895-4356(01)00372-9.
2. Bruix, J. and Sherman, M., Management of hepatocellular carcinoma. *Hepatology*, 42, 5, 1208–1236, Oct. 2005. doi: 10.1002/hep.20933.
3. Palsson, O.S., *et al.*, Development and validation of the Rome IV Diagnostic Questionnaire for adults. *Gastroenterology*, 150, 6, 1481–1491, May 2016. doi: 10.1053/j.gastro.2016.02.014.
4. Escudier, B., *et al.*, Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.*, 25, iii49–iii56, Sep. 2014. doi: 10.1093/annonc/mdu259.
5. Golub, T.R., *et al.*, Molecular Classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 5439, 531–537, Oct. 1999. doi: 10.1126/science.286.5439.531.
6. Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., Badr, S., The Occurrence of Sleep-Disordered Breathing among Middle-Aged Adults. *N. Engl. J. Med.*, 328, 17, 1230–1235, Apr. 1993. doi: 10.1056/nejm199304293281704.
7. Richards, S., *et al.*, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, 17, 5, 405–424, Mar. 2015. doi: 10.1038/gim.2015.30.
8. West, D. and West, V., Model selection for a medical diagnostic decision support system: a breast cancer detection case. *Artif. Intell. Med.*, 20, 3, 183–204, Nov. 2000. doi: 10.1016/s0933-3657(00)00063-4.
9. Glynne-Jones, R., *et al.*, Rectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.*, 28, iv22–iv40, Jul. 2017. doi: 10.1093/annonc/mdx224.
10. Escudier, B., *et al.*, Renal cell carcinoma: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.,* 30, 5, 706–720, Feb. 2019, doi: 10.1093/annonc/mdz056.
11. Kumar, A., Rathore, P.S., Dubey, A.K., *et al.*, LTE-NBP with holistic UWB-WBAN approach for the energy efficient biomedical application. *Multimed. Tools Appl.*, 82, 39797–39811, 2023. https://doi.org/10.1007/s11042-023-15093-7.
12. Sasubilli, S.M., Kumar, A., Dutt, V., Improving Health Care by Help of Internet of Things and Bigdata Analytics and Cloud Computing, in: *2020 International Conference on Advances in Computing and Communication*

Engineering (ICACCE), Las Vegas, NV, USA, pp. 1–4, 2020, doi: 10.1109/ICACCE49060.2020.9155042.

13. Singh, R., Ahuja, S., Kumar, A., Enhancing the Parkinson's Disease Detection Through Machine Learning and Feature Engineering, in: *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, pp. 1–5, 2024, doi: 10.1109/ICCCNT61001.2024.10724102.
14. Burri, S.R., Kumar, A., Baliyan, A., Kumar, T.A., Predictive Intelligence for Healthcare Outcomes: An AI Architecture Overview, in: *2023 2nd International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, Villupuram, India, pp. 1–6, 2023, doi: 10.1109/ICSTSN57873.2023.10151477.