International Journal of Innovative Research in Engineering
Volume 7, Issue 2 (March - April 2026), PP: 274-279.
https://www.doi.org/10.59256/ijire.20260702033
www.theijire.com
ISSN No: 2582-8746

Cancer Patient Identification using Machine Learning and Clustering

Amrita Sinha [1], Sujit Kumar Chatterjee[2]
[1, 2] Department of Computer Science & Engineering, Birla Institute of Technology, Patna Campus, Bihar, India.

How to cite this paper:
Amrita Sinha [1], Sujit Kumar Chatterjee[2] "Cancer Patient Identification using Machine Learning and Clustering", JIRE-V7I2-274-279.

Abstract: *Since cancer is one of the main causes of mortality worldwide, risk assessment and early detection are crucial. A machine learning-based method for identifying cancer patients using both clustering and classification techniques is presented in this paper. For analysis, a sizable dataset comprising more than 50,000 patient records with clinical, lifestyle, and demographic characteristics was used.*
Managing missing values, encoding categorical variables, and getting the dataset ready for modeling were all part of the data preprocessing step. Patients were divided into low, medium, and high risk groups using K-Means clustering. The models' prediction power was significantly improved by using these risk groupings.
Several machine learning techniques were developed and assessed, such as Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost. The models were evaluated using ROC analysis, confusion matrix, and typical assessment measures like precision, recall, and F1-score.
Additionally, a prediction system was created that enables users to enter patient information and obtain probability estimation and cancer risk. The suggested method shows how well clustering and classification techniques can be combined for healthcare prediction and decision assistance.

Key Words: *Cancer Prediction, Classification, Clustering, Healthcare Analytics, Machine Learning, Risk Assessment, XGBoost.*

I.INTRODUCTION

One of the biggest diseases that are known to cause death in the world today and a huge burden to the healthcare systems is cancer. The prevalence of cancer cases has been on the rise according to the world health reports because of the aging population, changes in lifestyles, exposure to the environment and genetic predisposition. Prompt identification of the disease and proper risk assessment should be made to enhance survival rates and decrease the pressure on the healthcare system.

The conventional diagnostic procedures used to diagnose cancer include clinical tests, imaging techniques, and expert analysis that may be time-consuming, costly, and rely on medical expertise. In addition, such techniques are not necessarily effective in early detection of high-risk patients. This is why the demand to develop smart systems that could aid in the early prediction and decision-making process grows.

Machine learning is a relatively new concept in healthcare that has become an effective tool in the previous years to analyze large datasets and detect previously unknown patterns. Machine learning algorithms have the capability of handling large quantities of patient data, such as demographic data, medical history, lifestyle, and laboratory results to accurately foresee the outcomes of diseases. These methods aid in automation in the prediction process and minimize human error.

This paper is aimed at creating a cancer prediction model based on a hybrid of clustering and classification methods. Unsupervised learning, which is known as clustering, is employed to cluster the patients into various risk groups like low, medium and high risk depending on their characteristics. This grouping assists in comprehending the structure of the data behind the scenes and makes predictive models more useful.

After clustering, supervised machine learning models are applied to classify patients based on their risk of cancer. Prediction is done using models like Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost. These models can be used to model complex relationships among features and enhance prediction accuracy. The ensemble methods such as random forest and XGBoost are especially useful among them because they can combine a number of weak learners.

The data, employed in the current study, is comprised of more than 50 thousand patient records, which makes the model stronger and more credible. The big data assists in the improved generalization and decrease of overfitting, which results in enhanced performance on the unknown data. The dataset is prepared to modeling by various preprocessing methods like data cleaning, categorical variables encoding, and classes balancing.

Moreover, a prediction system is created which enables users to enter patient information and receive cancer risk and probability estimate. This renders this system more viable and applicable to real life health care uses.

The primary goal of the given research is to develop an effective and precise machine learning model of cancer risk prediction that could help medical workers to diagnose cancer at the earliest stage and make their choice. The proposed solution will utilize a mixture of clustering and classification methods and is expected to deliver a reliable and scalable cancer prediction solution.

II.MATERIAL AND METHODS

Study Design

This research paper is founded on a machine learning model of identifying cancer patients with the use of supervised and unsupervised learning methods. The aim is the creation of a predictive system which can be used to categorize patients according to their level of risk of cancer.

Dataset Description

The data utilized in this research is composed of about 50000 patient records with demographic, clinical, and lifestyle-related data. The dataset covers variables like age, gender, BMI, smoking habits, alcohol use, genetic defects, chronic illnesses, blood markers and other health related variables.

The variable of interest is the prediction of cancer, that is, whether a patient is at risk or not. The size of the dataset is very large and enhances robustness and generalization of the models.

Data Preprocessing

Preprocessing of data is a necessary process to ascertain the quality and usability of the data. The steps followed were:

Missing Values: Missing or null values were deleted or filled in.
Encoding: Categorical variables like gender, smoking status and type of diet were encodified using Label Encoding into numerical values.
Selection of Features: The irrelevant features were eliminated including; patient ID.
Data Scaling: Standardization was done where necessary to normalize the feature values.
Data Balancing: SMOTE (Synthetic Minority Oversampling Technique) was employed to overcome the imbalance of classes and enhance the performance of the model.

Clustering

K-Means clustering algorithm was used to cluster patients in various risk groups according to their health characteristics. The clusters were to be three in amount:
Low Risk
Medium Risk
High Risk

The clusters created were further included as a new feature referred to as Risk_Group that led to the enhancement of the performance of classification models by offering further information regarding patient grouping.

Train-Test Split

The data were separated into training and testing to assess the performance of the models. Typically, 80% of the data was used for training and 20% for testing. This makes sure that the model is trying unseen data.

Model Training
Four machine learning models were implemented:

Support Vector Machine (SVM): This is a classifier that has linear decision boundaries.
Decision Tree: It is a simple model that divides the data according to the conditions of the features.
Random Forest: This is an ensemble model, which is a combination of various decision trees to enhance accuracy.
XGBoost: It is a boosting algorithm capable of giving high performance and efficiency.
All the models were trained with the processed dataset and tested with the test data.

Evaluation Metrics

The following metrics were used to measure the performance of the models:
Accuracy: Assesses general model accuracy.
Precision: Measures the number of actually correct predicted positives.
Recall: Estimates the number of true positives that are correctly detected.
F1 Score: Precision and recall in a harmonic mean.

In addition, advanced evaluation techniques were used:
Confusion Matrix: To find out the true and false predictions.
ROC Curve: To compare the performance of classification.
Learning Curve: To examine the model performance as more data is added.
Prediction System

The trained model was used to develop a prediction system. The system enables users to enter details of the patient including age, BMI, lifestyle and medical conditions. Depending on the input, the model will be used to predict whether a patient is at risk of cancer or not and will also provide a score of probability that will indicate the level of risk.

III.RESULT

The models of the machine learning were tested on accuracy, precision, recall, and F1-score. Table 1 shows the results of various models.

	Model	Accuracy	Precision	Recall	F1 Score
3	XGBoost	0.891103	0.952827	0.822884	0.883101
2	Random Forest	0.838533	0.868744	0.797457	0.831576
0	SVM	0.735702	0.716600	0.779543	0.746748
1	Decision Tree	0.727470	0.713859	0.759029	0.735751

Table 1: Performance Comparison of Models

Based on the results, it is seen that XGBoost was the most accurate model with 89.11% outperforming all the other models. It also reported the best F1- score, which means that there is a good trade-off between precision and recall.

The forest also fared well with its ensemble characteristic since the accuracy of the random Forest was 83.85. Its performance was however a little lower than XGBoost.

The accuracy of Support Vector Machine (SVM) and Decision Tree was relatively low. This could be because of their weaknesses in working on large datasets and intricate relations between features.

Confusion Matrix Analysis

A confusion matrix is the technique to assess the classification model performance in comparison to the actual and predicted values. It gives details of true positives, true negatives, false positives and false negatives which would aid in knowing the accuracy and error of the model.

All models were used to create confusion matrices that assessed their performance in classification.

Decision Tree.

Decision Tree was found to have relatively high misclassification as compared to other models.

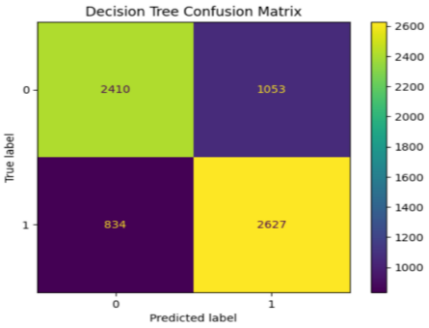

Figure 2: Confusion Matrix of Decision Tree

SVM

SVM was moderate in its performance and was characterized by high false predictions as compared to ensemble models.

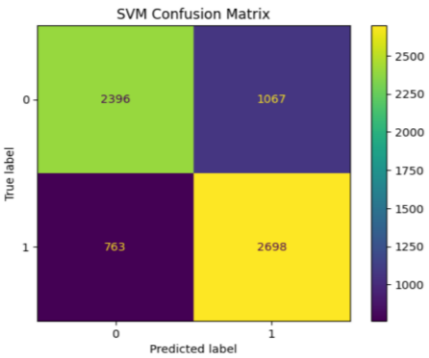

Figure 3: Confusion Matrix of SVM

Random Forest

Random Forest had better classification balance and fewer misclassifications.

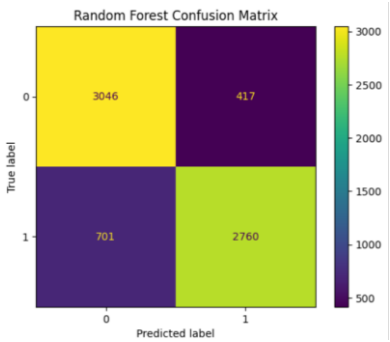

Figure 4: Confusion Matrix of Random Forest

XGBoost

With XGBoost, the most successful results were achieved with the following:

High True Positives (2848)

Low False Negatives (613)

This is very crucial in cancer prediction because it is dangerous to miss out on a patient.

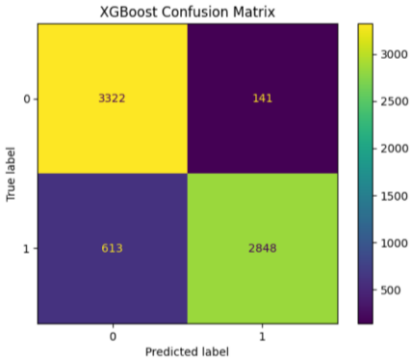

Figure 5: XGBoost Confusion Matrix.

ROC Curve Analysis

ROC (Receiver Operating Characteristic) curve is a tool to assess the performance of classification models by examining the trade-off between a True Positive Rate (TPR) and a False Positive Rate (FPR). It assists in knowing the extent to which the model is capable of differentiating between the various classes.

A model that lies nearer to the top-left corner is a pointer of better performance and a diagonal line is a representative of a random classifier.

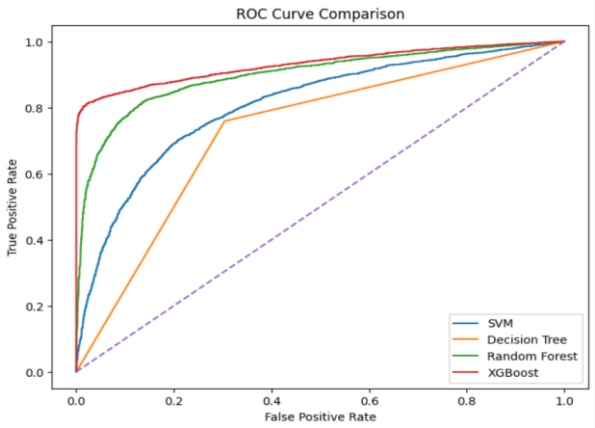

Figure 5: ROC Curve Comparison of All Models

Based on the comparison of the ROC curves:

- XGBoost demonstrates the highest performance with a curve that is nearest to the upper-left corner.
- Random Forest also exhibits good Classification capability.
- SVM is moderate.
- Decision Tree is with relatively lower performance.

This shows that the ensemble models such as XGBoost and the Random Forest are more useful in classifying cancer and non-cancer cases.

Learning Curve Analysis

The learning curve is used to assess how a machine learning model's performance increases as training data volume increases. It aids in finding problems like underfitting and overfitting.

A successful model's ability to generalize to new data is demonstrated by a modest difference between training and validation scores.

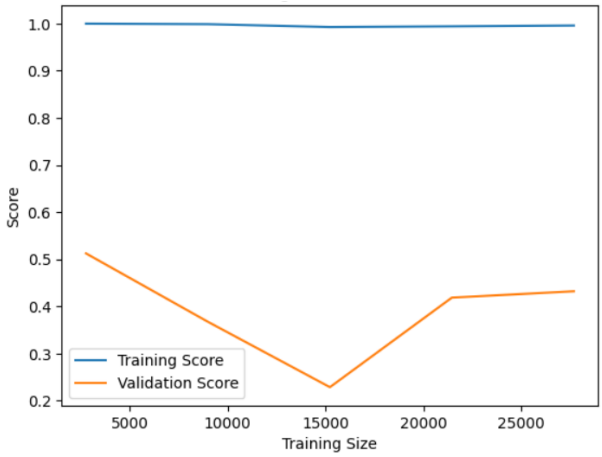

Figure 7: Learning Curve of XG Boost Model

Based on the learning curve:

- The model effectively learns from the data, as evidenced by the excellent training score.
- As the training size grows, the validation score rises.
- There is not much of a difference between training and validation scores.

This suggests that there is little overfitting or underfitting and that the model is well-balanced. It also demonstrates how performance increases with additional data.

Prediction System Output

A prediction system was designed to answer the question of cancer risk on the input of the user. The system accepts patient information like age, BMI, and medical features as input to give a classification and probability output.

Example:

Input: Patient details (age = 45, BMI = 26, etc.)
Prediction: Low Risk
Risk Percentage: 4.33%
This proves that the model can be applied to the real-world to help in the early diagnosis and decision making.

IV.DISCUSSION

The outcomes of the machine learning models indicate that ensemble techniques are more effective than conventional methods of classification. XGBoost was the best model in terms of accuracy and F1-score, which means that it demonstrated better performance in cancer prediction. This is primarily because of its boosting mechanism whereby several weak learners are pooled together to form a powerful predictive model.

Random Forest was also effective as it is able to work with complex data and lessen overfitting. Nevertheless, its performance was a bit less than XGBoost. Conversely Support Vector machine (SVM) and decision tree demonstrated relatively lower accuracy, and this could be attributed to the fact that they are less effective when it comes to large data sets and complicated features interactions.

Clustering was also made use of in this study and was used to group patients into various risk categories which enhanced the overall performance of the prediction. The extra attribute (Risk_Group) was helpful to the classification models. Based on the analysis of the confusion matrix, it was clear that the XGBoost had fewer false negatives, which could be highly important when predicting cancer, and a positive case can be vital to miss. The ROC curve also validated that XGBoost is a better classifier than other models.

The learning curve analysis indicated that the model works well as the amount of data increases and is not severely affected by overfitting or underfitting. This indicates that the model is well-generalized and reliable.

All in all, the clustering and machine learning methods were effective in predicting cancer. The system created is not merely accurate but also practical to be used in reality.

V.CONCLUSION

Using both clustering and classification techniques, a machine learning-based method for identifying cancer patients was effectively established in this work. The models' robustness and generalization were enhanced by using a sizable dataset with more than 50,000 patient records.

K-Means clustering was used to preprocess and analyze the data in order to classify patients into various risk groups. The performance of the classification models was improved by using these clusters. Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost are a few of the machine learning methods that were used and assessed.

With the highest accuracy and F1-score of all the models, XGBoost performed the best. The model's efficacy and dependability were further validated by the evaluation using the confusion matrix, ROC curve, and learning curve.

The created prediction method is valuable for real-world healthcare applications since it can provide cancer risk together with likelihood % based on patient input.

All things considered, the suggested system offers a realistic, accurate, and effective way to forecast cancer risk and can help medical professionals make early diagnoses and decisions.

References

1. Han, J., Kamber, M., & Pei, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.
2. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019.
3. [Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 2011.
4. Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the ACM SIGKDD*, 2016.
5. Breiman, L. "Random Forests." *Machine Learning Journal*, 2001.
6. Cortes, C., & Vapnik, V. "Support-Vector Networks." *Machine Learning Journal*, 1995.
7. Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.
8. World Health Organization (WHO). "Cancer Fact Sheet." Available: https://www.who.int
9. National Cancer Institute. "Cancer Statistics." Available: https://www.cancer.gov
10. Kaggle Dataset. "Cancer Prediction Dataset." Available: https://www.kaggle.com
11. Kotsiantis, S. "Supervised Machine Learning: A Review of Classification Techniques." *Informatica*, 2007.
12. Mitchell, T. M. *Machine Learning*. McGraw-Hill, 1997.
13. Dua, D., & Graff, C. "UCI Machine Learning Repository." University of California, Irvine.
14. Esteva, A., et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature*, 2017.
15. Topol, E. "High-performance medicine: the convergence of human and artificial intelligence." *Nature Medicine*, 2019.