Volume 10, Issue 4, April – 2025 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr929

Predicting Cancer Outcomes: A Comparative Study of ML Models

Yuvraj Singh[1]; Swati[2]; Dhirender Pratap Singh[3]; Tanuj[4]; Yash Pratap Singh[5]; Parth Singh[6]

[2] Engineer

[1,3,4,5,6] UIE CSE, Chandigarh University, Mohali, Punjab, India

Publication Date: 2025/04/30

Abstract: Prognostic accuracy in cancer is vital for timely diagnosis and effective treatment planning. This study evaluates the performance of three machine learning techniques—Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT)—in forecasting cancer progression using clinical and histopathological data. Results demonstrate that SVM surpasses KNN and DT in predictive precision, establishing its robustness in prognostic modeling. The research highlights how machine learning can support clinicians with data-driven decision-making tools to improve patient care. Future directions may involve advanced deep learning models and optimized feature selection to enhance predictive capabilities further.

Keywords*: Cancer Survival Prediction, ML Algorithms, SVM Classifier, KNN Algorithm, Decision Tree Model.*

How to Cite: Yuvraj Singh; Swati; Dhirender Pratap Singh; Tanuj; Yash Pratap Singh; Parth Singh (2025). Predicting Cancer Outcomes: A Comparative Study of ML Models. *International Journal of Innovative Science and Research Technology*, 10(4), 1884-1888. Https://Doi.Org/10.38124/Ijisrt/25apr929

I. INTRODUCTION

Malignancies continue to pose a major globalhealth burden, representing one of the leading causes of mortality across popula tions. Timely identification and precise prognosis estimation play a vital role in enhancing therapeutic efficacy and extending patient lifespans. Traditional diagnostic approaches, including histopathological examinations and radiological imaging, typically demand specialized interpretation, rendering them labor-intensive and susceptible to subjective variability. The advent of machine learning techniques has introduced transformative potential in healthcare diagnostics, enabling sophisticated analysis of multidimensional patient data to uncover hidden predictive patterns.

This study evaluates three predictive models—a maximum- margin classifier, instance-based learner, and hierarchical decision model—for forecasting cancer progression using clin ical data. These algorithms effectively capture intricate, nonlinear relationships in patient records to improve outcome prediction accuracy. Comparative analysis revealed SVM's superior predictive performance, establishing it as the optimalchoice among the evaluated classifiers for clinical prognosis applications..

This study aims to systematically assess the predictive capability of various machine learning (ML) models in forecasting cancer prognosis. By leveraging quantitative performance metrics, we conduct a comparative analysis to determine the most reliable algorithm for clinical decision support.

The findings of this study provide healthcare practitioners with actionable, data-informed insights to optimize clinical decision- making, thereby improving patient care and therapeutic outcomes. Future research could achieve greater predictive accuracy by refining model architectures and implementing advanced feature engineering techniques.

II. LITERATURE SURVEY

In contemporary oncology research, machine learning has emerged as a critical component for prognostic prediction, demonstrating remarkable capability in analyzing complex biomedical datasets with precision. While conventional diagnostic approaches - including clinical evaluations, imaging studies, and histopathological examinations - remain valuable, they often involve substantia l time commitments and may be influenced by interpretive variability. ML techniques provide an alternative approach by automating prognosis predictions and enhancing diagnostic accuracy, thereby assisting healthcare professionals in making informed decisions.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr929

➢ *Machine Learning for Cancer Outcome Prediction*

The maximum-margin classifiers (SVM) algorithm has gained prominence in medicalinformatics due to its exceptional capability to process high-dimensional feature spaces and discern complex data patterns. Extensive research validates SVM's robust predictive performance across multiple oncology domains, particularly in breast carcinoma, pulmonary malignancies, and prostate cancer progression modeling. Researchers have noted that SVM delivers high accuracy, particularly when applied to well-structured datasets. However, its effectiveness depends on the selection of kernel functionsand hyperparameter tuning, which can influence its predictive capability. Maximum-margin classifiers show particular strength with imbalanced medical data, generating robust decision boundaries that maintain accuracy across unequal class distributions - a critical advantage in cancer prognosis studies where certain outcomes are naturally rare. identifies the most effective prognostic predictor through empirical validation on clinical datasets.

- *Data Collection*

The dataset used in this study consists of medical records, including patient demographics, clinical test results, and The tumor characteristics.

- *K-Nearest Neighbors (KNN) in Cancer Diagnosis*

The instance-based learning method (k-NN) serves as an intuitive yet effective approach for medical classification, particularly in oncology outcomes prediction. This algorithm classifies patients by analyzing similarity measures between current cases and historical records in the feature space. While demonstrating utility in cancer studies, its performance depends critically on optimal neighbor selection (k-value) and exhibits sensitivity to data dimensionality and sample size. Current literature notes particular challenges with: (1) feature noise in clinical variables, and (2) computational scalability with expanding datasets - factors that may constrain its application in comprehensive prognostic systems.

- *Decision Tree (DT) in Cancer Prognosis*

Decision Tree (DT) models are widely used in medical research due to their interpretability and ease of use. They offer a transparent Clinical decision pathways, allowing health- care professionals to understand the logic behind predictions. Studies shows that DTs are efficient in identifying key factors influencing cancer prognosis. However, they tend to overfit training data, reducing their ability to generalize to unseen cases. Pruning and ensemble methods (e.g., Random Forest, Gradient Boosting) can improve Decision Tree (DT) reliability in medical diagnostics. Although DTs offer interpretable classification, their predictive accuracy typically lags behind advanced models like SVM.

- *Advancements in ML-Based Cancer Prognosis*

Recent advances in ML/AI have enabled deep learning techniques like (ANNs) and CNNs to optimize cancer prognosis accuracy, though they demand substantial data and computing power. Despite these innovations, traditional algorithms (SVM, KNN, DT) maintain clin ical relevance due to their interpretability, computational efficiency, and actionable insights for medical decision-making.

This research extends prior work by evaluating and contrasting the effectiveness of Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees (DT) in forecasting cancer outcomes. By evaluating these models using standard metrics, this research aims to determine the most suitable ML algorithm for assisting healthcare professionals in making precise and data -driven decisions regarding cancer prognosis.

III. METHODOLOGY

This study employs a systematic machine learning pipeline for cancer prognosis prediction, comprising: (1) data acquisition and preprocessing, (2) feature selection, (3) model implementation (SVM, KNN, DT), and (4) performance evaluation. The comparative analysis of these algorithms data is sourced from publicly available medical repositories such as the UCI Machine Learning Repository or Kaggle. It contains labeled instances indicating whether a patient's condition is likely to progress or remain stable.

➢ *Data preprocessing*

The dataset undergoes rigorous preprocessing to ensure reliability and consistency:

- *Missing values were handled through statistical imputation, replacing them with the mean, median, or mode as appropriate. Feature Scaling:*

Numerical attributes are standardized or normalized using techniques like Min -Max scaling to maintain uniformity across features.

- *Encoding Non-numerical features:*

Categorical da ta such as tumor types were converted into numericalvalues using encoding methods, with one-hot encoding applied to nominal variables and label encoding used for ordinal categories.

- *Outlier Detection and Removal:*

Statistical methods such as The interquartile range(IQR) are used to identify and eliminate anomalies that could affect model performance.

➢ *Feature selection*

To enhance model efficiency and reduce computational complexity, feature selection methods are applied. Three principal feature selection approaches were employed: correlation was computed to measure the strength and direction of linear relationships, PCA for orthogonal transformation of feature space, and RFE to iteratively select optimal feature subsets. This multi- method strategy ensures robust identification of clinically significant predictors while eliminating noise and redundancy in the training data.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr929

➢ *Model implementation*

The three ML models—SVM, KNN, and DT—are implemented using Python's Scikit-Learn lib rary. The dataset was partitioned into training and testing sets following conventional splits (70-80% for training, 20-30% for validation). Model optimization was achieved through systematic hyperparameter tuning using either exhaustive (Grid Search) or stochastic (Random Search) exploration of the parameter space.

- Support Vector Machine (SVM): Utilizes different kernel functions (linear, RBF, polynomial) to optimize classification performance.
- K-Nearest Neighbors (KNN): Classifies new instances by analyzing the closest 'k' data points. The best 'k' value is determined through cross-validation.
- Decision Tree (DT): Constructs a tree-based structure to classify data points based on key features. Pruning techniques are applied to prevent overfitting.

➢ *Model Evaluation*

Each model is evaluated using standard classification metrics to determine its effectiveness:

- *Accuracy:* Quantifies the proportion of correct predictions (both positive and negative) among all cases, representing overall model performance.
- *Precision:* Measures the model's exactness in positive predictions (true positives/[true + false positives]).
- *Recall:* Evaluates modelsensitivity to identify all actual 3. positives (true positives/[true positives + false negatives]).
- *F1-Score:* Represents the harmonic mean of precision and recall (2 × [precision×recall]/[precision+recall]), providing a balanced assessment of modelperformance for imbalanced datasets.
- *ROC-AUC Score:* Evaluates the classifier's discriminative ability by measuring the area under the Receiver Operating Characteristic curve, quantifying how well the model distinguishes between positive and negative classes across all classification thresholds.

➢ *Performance Comparison and Analysis*

A comparative analysis revealed SVM's superior accuracy for cancer prognosis prediction, positioning it as the optimalchoice. Each model's diagnostic applicability was assessed through its strengths and limitations.

By following this st ructured methodology, this study ensures that the ML models are trained, tested, and evaluated effectively to enhance data-driven decision-making in cancer prognosis prediction.

IV. IMPLEMENTATION OF MODULE

The implementation process for cancer prognosis prediction is divided into several key modules, each contributing to different stages of the machine learningpipeline. These modules include data preprocessing, feature selection, model training, performance evaluation, and result analysis. The entire implementation is carried out in Python, utilizing libraries such as Scikit-Learn, Pandas, NumPy, and Matplotlib for efficient data processing and model development.

➢ *Data Preprocessing*

This module ensures that the dataset is properly prepared before training the models. The preprocessing steps include:

- Loading: The dataset is imported using Pandas and structured for analysis.
- Handling Missing Values: Missing data is addressed using statistical techniques like mean, median, or mode imputation.
- Feature Scaling: Numerical attributes are normalized using Min-Max Scaling or Standardization to ensure uniformity.
- Encoding Categorical Features: Non-numeric variables are transformed into numerical values using Label Encoding or One-Hot Encoding.
- Outlier Detection and Removal: Unusual data points are identified and removed using statistical methods like the Interquartile Range (IQR).

➢ *Feature Selection*

To improve model performance and reduce unnecessary computations, only the most relevant features are selected using:

- Correlation Analysis: Identifies and removes features with high redundancy.
- Principal Component Analysis (PCA)
- Performs Recursive Feature Elimination (RFE): Employs an iterative backward selection process that:

✓ Trains the model on all features
✓ Eliminates the least important feature(s)
✓ Repeats until optimal feature subset is identified

➢ *Model Training*

Three supervised learning algorithms were evaluated: Support Vector Machines (SVM) for maxima l margin classification, K- Nearest Neighbors (KNN) for instance-based learning, and Decision Trees (DT) for hierarchical feature partitioning. The experimental protocol followed these key steps:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr929

- *Data Partitioning:*

✓ Stratified 70-30 train-test split to maintain class distribution
✓ Alternative 80-20 partitioning for sensitivity analysis

- *Model Development:*

✓ Baseline training with default parameters
✓ Systematic hyperparameter optimization using:

- Grid Search with exhaustive parameter space exploration
- Randomized Search for efficient sampling of hyperparameters

- *Performance Validation:*

✓ 10-foldcross-validation toensure robust generalization
✓ Stratified sampling in each fold to preserve class ratios

➢ *Model Evaluation*

The trained models are assessed using standard evaluation metrics, including:

- Accuracy: Overall prediction correctness (TP+TN)/Total
- Recall (TP/TP+FN) & Precision (TP/TP+FP): Class-specific performance
- F1-Score: Harmonic mean (2×P×R/P+R) for balanced evaluation
- ROC-AUC: Class discrimination ability across thresholds

➢ *Performance Analysis and Comparison*

This module compares the performance of the three models to determine the most effective approach for cancer orthogonal transformation to convert correlated features into uncorrelated principal components demonstrate the diagnostic trade off between true positive rates and false positive rates.

V. RESULT

This study comparatively assessed three machine learning algorithms - Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT) - using accuracy, precision, and recall metrics. The experimental findings demonstrate SVM's superior performance, exhibiting the highest predictive accuracy amongthe evaluated models. Table 1 presents a comprehensive summary of each model's performance metrics:

Table 1 Performance Metrics of Ml Models

Model	Accuracy (%)	Precision (%)	Recall (%)
SVM	97.84	98.88	98.83
KNN	93.54	96.51	96.51
Decision Tree	95.69	98.83	98.83

VI. CONCLUSION

This study investigates ML techniques for cancer outcome prediction, evaluating three distinct algorithms: Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees (DT). The methodological framework incorporated data preprocessing, feature selection, model optimization, and comprehensive performance assessment to determine the optimal prognostic model.

➢ *Comparative analysis revealed SVM's superior predictive accuracy, establishing it as the most effective algorithm for this clinical application. This advantage stems from SVM's:*

- Robust handling of high-dimensional clinical data
- Effective separation of non-linear class boundaries
- Optimal generalization capability with limited samples

While KNN and Decision Tree also performed reasonably well, their results were slightly less accurate due to their sensitivity to variations in data and noise.

The findings emphasize the potential of machine learning in the medical field, demonstrating that predictive models can assist healthcare professionals in assessing cancer prognosis at an early stage. Further improvements, such as using larger datasets, incorporating deep learning techniques, or integrating additional medical parameters, could enhance prognosis prediction. The results are visualized using Matplotlib and Seaborn through:

- Confusion Matrix: The confusion matrix tabula test model predictions versus actual outcomes across four classification categories.
- ROC Curve Receiver Operating Characteristic (ROC) curves linearly

This study demonstrates machine learning's potential to enhance cancer prognosis through data-driven modeling. SVM outperformed other algorithms, proving particularly effective for clinical prediction tasks. The results highlight how ML can improve diagnostic accuracy and support evidence-based oncology decisions.

FUTURE SCOPE

While our results confirm SVM's superior performance in cancer outcome prediction, significant opportunities exist to enhance prognostic modeling through:

➢ *Deep Learning Architectures:*

Deploying convolutional (CNN) and recurrent (RNN) neural networks may enable more sophisticated analysis of complex clinical patterns in large oncology datasets, potentia

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr929

lly increasing prediction precision.

- *Multimodal Data Integration*

Combining genomic data, medical images, and longitudinal electronic health records could provide a more holistic patient profile, improving prognostic accuracy through comprehensive data synthesis.

- *Enhancing Feature Selection Methods*

Applying more advanced feature selection techniques, such as Genetic Algorithms or Ensemble-Based Feature Selection, can help eliminate irrelevant attributes and improve model efficiency while maintaining high accuracy.

- *Developing Personalized Prognosis Models*

Customizing prediction models based on individual patient factors such as genetics, lifestyle, and medical history can en- hance the reliability of predictions, leading to better treatment recommendations.

- *Cloud-Based and AI-Integrated Healthcare Systems*

By deploying the model on cloud-based AI healthcare systems, it can achieve remote access, scalable performance, and smooth compatibility with current medical infrastructure, enhancing its practical utility in clinical settings.By implementing these future improvements, machine learning-based cancer prognosis prediction can becomean even more valuable tool in the healthcare sector, aiding in early diagnosis, personalized treatment planning, and improved patient outcomes.

REFERENCES

[1]. Kourou, A., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). An overview of how machine learning contributes to predicting and determining cancer outcomes. *Computational and Structural Biotechnology Journal*, 13, 8–17.J.

[2]. Cruz, J. A., & Wishart, D. S. (2007). The role of machine learning in forecasting and analyzing cancer prognosis. *Cancer Informatics*, 2, 59–77.

[3]. Piryani, S. S., & Saxena, S. (2020). Comparative evaluation of machine learning models for breast cancer forecasting. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(8), 371–377.

[4]. Patel, N. R., Patel, R., & Patel, B. G. (2016). Analysis and comparison of classifiers used for breast cancer prognosis. *International Journal of Computer Applications*, 145(2), 34–39.

[5]. Chaurasia, A., & Pal, S. (2014). A fresh perspective on breast cancer identification through data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineering*, 2(1), 2456–2471.

[6]. Shinde, G., & Kalbande, D. S. (2019). Classification and detection of lung cancer via machine learning methods. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* , 8(6), 349–354.

[7]. Asri, B., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Implementing machine learning techniques for assessing breast cancer risk and diagnosis. *Procedia Computer Science*, 83, 1064–1069.

[8]. Chien, H. J., Feizi-Derakhshi, M. R., & Chien, D. K. Y. (2021). Review of machine learning-based models for cancer diagnosis. *Cancers*, 13(6), 1386.

[9]. Al-Shaer, A. M., & Al-Khasawneh, R. (2019). A study on machine learning techniques for predicting breast cancer. *Applied Computing and Informatics*, 15(1), 1–13.

[10]. Yu, W., Jiang, S., & Zou, Y. (2020). Evaluating different machine learning approaches for breast cancer outcome prediction. *IEEE Transactions on Medical Imaging* , 39(3), 860–869.

[11]. Lee, S. H., Kim, C. S., & Kim, M. (2020). Assessing the effectiveness of machine learning algorithms in breast cancer prediction. *Scientific Reports*, 10(1), 1–10.

[12]. Krawczyk, B. (2016). Challenges and future trends in learning from

[13]. data with class imbalance. *Progress in Artificial Intelligence*, 5(4), 221–232.

[14]. Ding, J., Zhang, A., & Yang, W. C. (2018). Application of support vector machines in cancer classification tasks. *BMC Bioinformatics*, 19(1), 1–15.

[15]. Azar, M. V., & El-Sappagh, T. A. (2020). Enhancing cancer prognosis accuracy through a hybrid feature selection technique. *Healthcare*, 8(3), 246..