Predictive Analysis of Cancer Patient Survival Levels Using Machine Learning Techniques

Navroop,shail

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Abstract*:* - Early diagnosis of cancer as well as detailed predictions about its severity play an essential role in providing effective treatments at the right time because cancer stands as one of the primary causes of death worldwide. This research paper conducts an extensive evaluation of cancer patient information through machine learning algorithms like Random Forest , Linear Regression , Support Vector Machines (SVM) , Decision Trees , etc to determine cancer severity levels. The study requires extensive preprocessing of data followed by exploratory data analysis and feature correlation analysis which leads to the application of classification models for extracting patient demographic and clinical patterns. The analysis used accuracy as well as precision together with recall and F1-score to measure model performance. Ensemble methods show superior predictive performance according to the study results because Random Forest stands out as an effective prediction tool for healthcare decision support systems.

Index Terms- Cancer Prediction, Machine Learning, Random Forest, Support Vector Machine, Classification, Data Analysis, Healthcare Analytics

I. INTRODUCTION

Cancer, which is one of the major causes of death worldwide, is a collection of diseases that involve the uncontrolled growth and dissemination of abnormal cells within the body. Cancer, as reported by the World Health Organization (WHO), is responsible for almost 10 million deaths every year, and it is a major public health issue. The global cancer burden is expected to increase significantly over the coming decades due to aging populations, unhealthy lifestyles, and environmental factors. Early detection and accurate prediction of cancer severity are important in lowering mortality rates, improving patient quality of life, and maximizing health resources.

The conventional cancer diagnosis has traditionally been dependent on a set of imaging, histopathology, clinical observation, and genetic testing. Though these methods have worked to a certain degree, they are not always adequate in offering timely, scalable, and personalized evaluations to all patients. Most developing nations still do not have proper infrastructure and trained experts for early screening of cancer. Consequently, the majority of cases are diagnosed in advanced stages where treatment becomes ineffective and limited.

Over the past few years, the emergence of data-centric technologies has transformed the healthcare sector. With the widespread use of Electronic Health Records (EHRs), wearable health devices, and biomedical sensors, huge amounts of structured and unstructured health data are being created. If effectively harnessed, this data can unlock previously concealed patterns and correlations, allowing for more accurate disease detection, prognosis, and treatment. But the enormity and complexity of medical data necessitate smart ways to interpret and analyze them efficiently.

Machine Learning (ML), a branch of Artificial Intelligence (AI), provides robust methods for automatic pattern recognition, prediction, and decision-making. In contrast to conventional rule-based systems, ML models can learn from data and refine their performance over time. Supervised learning algorithms, most notably, have exhibited unprecedented potential in classification and regression applications in the medical field. When used in the context of cancer prediction, ML models can aid clinicians in predicting the possibility of cancer emergence, suggesting possible biomarkers, and predicting disease stages or severity grades based on patient-specific variables.

This research study investigates the use of ML algorithms in forecasting the degree of severity of cancer from patient data that includes lifestyle and biological data. The dataset employed contains features like age, gender, air pollution exposure, drinking behavior, genetic hazard, obesity status, smoking status, and other health metrics. The data was preprocessed to deal with missing values, input normalization, and encoding categorical variables, and then exploratory data analysis to observe relationships between features and the target variable.

- Various classification algorithms were used in this research, which include:
    - Random Forest Classifier – an ensemble learning algorithm renowned for its accuracy and robustness
    - Support Vector Machine (SVM) – a strong method of high-dimensional classification
    - Decision Tree Classifier – an easy-to-understand yet simple model
    - K-Nearest Neighbors (KNN) and AdaBoost – employed for comparative analysis

Performance of the models was assessed by applying standard parameters: accuracy, precision, recall, F1-score, and confusion matrices. Cross-validation and hyperparameter search methods like Grid Search and Randomized Search were utilized to improve model generalization. Visualizations such as heatmaps, boxplots, and count plots facilitated effective interpretation of the data distribution and relations.

The results show that the Random Forest Classifier performs better than other models in regards to overall predictive accuracy and robustness and can be applied in real-world settings in healthcare analytics platforms. The study also points out the interpretability of tree-based models, which is paramount when used in clinical environments where knowing the reasoning behind predictions is just as vital as the predictions themselves.

In addition, this research promotes the incorporation of AI in the health decision-making process. A good predictive model not only decreases the diagnostic load on health clinicians but also helps in the triaging of patients and optimal resource allocation. It has the potential to revolutionize public health surveillance, remote diagnosis, and telemedicine through early warnings and risk stratifications.

In conclusion, this research shows that it is feasible and effective to apply ML methods in predicting cancer severity. It further serves as a stepping stone for more research using larger datasets, deep neural networks, and coupling with real-time clinical systems. By filling the gap between medical research and data science, this work helps to develop smart healthcare systems that are proactive, patient-centered, and data-based.

2.Literature Review

The integration of Machine Learning (ML) into the healthcare domain has significantly transformed how medical data is interpreted and utilized. In particular, cancer diagnosis and severity prediction have become prime focus areas due to the abundance of clinical data and the urgent need for early detection systems. Throughout the past decade, countless research efforts have been focused on the possibility and effectiveness of applying ML algorithms for automating and improving the diagnostic process for different forms of cancer.

A. Machine Learning for Cancer Prediction

Cancer prediction by ML is essentially the mapping between input features (e.g., age, lifestyle, genetic markers) and target class (e.g., stage or severity level of cancer). Numerous supervised learning techniques have been explored in this area.

Kourou et al. (2015) presented one of the first in-depth reviews of ML applications in cancer prediction and prognosis. They highlighted the versatility of Support Vector Machines (SVM), Random Forest (RF), Artificial Neural Networks (ANN), and ensemble approaches in providing high accuracy and generalizability to diverse types of cancer. Their review highlighted the significance of feature selection and preprocessing in achieving the ultimate model performance.

In the same vein, Delen et al. (2005) used Decision Trees, Neural Networks, and Logistic Regression on SEER (Surveillance, Epidemiology, and End Results) data to forecast survivability rates for breast cancer patients. They noted that Decision Trees were more interpretable and thus a better fit for embedding in medical decision support systems.

In a different study, Cruz and Wishart (2006) showed the power of SVMs on gene expression data for binary classification problems like tumor vs. normal or benign vs. cancerous. Their work helped to show how kernel-based methods, especially with high-dimensional biomedical data, have great benefits in dealing with intricate, nonlinear relationships.

B. Ensemble and Hybrid Models in Medical Diagnosis

The increasing complexity of biomedical data has seen more employment of ensemble and hybrid models, which pool the strengths of various classifiers. Random Forest has been particularly popular owing to its resilience, insensitivity to overfitting, and good handling of missing data and categorical variables.

- Chaurasia and Pal (2017) investigated several classification techniques—Naïve Bayes, Decision Trees, and Random Forest—for predicting breast cancer and found that Random Forest performed the best with a more than 96% accuracy. They favored its feature importance ranking as a useful device for determining the most significant factors in cancer development.

- Nahid and Kong (2017) suggested a hybrid classification model for lung cancer detection that used feature selection methods in combination with ensemble learning. Their research highlighted the significance of dimensionality reduction in enhancing classifier efficiency and interpretability.
- Razzak et al. (2019) had surveyed deep learning models in medical image analysis with special emphasis on Convolutional Neural Networks (CNNs) for the detection and segmentation of tumors. Though CNNs provided state-of-the-art performance for image-based diagnosis, their black-box and computational-intensive nature hindered their realistic application in resource-constrained settings.

C. Comparative Studies and Evaluation Metrics

A number of comparative investigations have been conducted to compare the performance of the classifiers on cancer data sets. Asri et al. (2016) conducted an empirical comparison of the classifiers on the Wisconsin Breast Cancer data set by employing Decision Tree, SVM, Naïve Bayes, and Logistic Regression. Decision Tree performed the best in both interpretability and in-classification accuracy.

A comparison study by Zolbanin et al. (2015) on how K-Nearest Neighbors (KNN), SVM, and Artificial Neural Networks fared on a dataset of lung cancer was also discussed. They wrote that even though Neural Networks showed good results on large datasets during training, it needed increased computation and higher train times compared to Decision Trees and KNN, which needed lower computation and took less time in training.

The performance of ML models in healthcare fields frequently relies heavily on the quality of input data. Missing values, noise, imbalanced classes, and heterogeneous data types are major challenges.

U. Nahid et al. (2017) and Sharma et al. (2019) emphasized that preprocessing of data such as normalization, encoding categorical variables, and handling missing values is crucial to enhance model performance and convergence. Additionally, methods like Principal Component Analysis (PCA), Mutual Information, and Chi-Square test are often employed to carry out feature selection and dimensionality reduction.

Our research follows a similar strategy by incorporating preprocessing methods like label encoding, z-score normalization, removal of outliers, and feature correlation analysis prior to using ML algorithms. These operations facilitate improved model generalization and higher metric scores at the time of testing.

E. Gaps and Challenges in Existing Literature

Although the literature on cancer prediction is vast, there are a number of limitations that are still not addressed:

- Model Interpretability: Several high-performing models, particularly those based on deep learning, are black-box models and provide limited explainability, a requirement that is key in healthcare settings.
- Imbalanced Datasets: Most cancer data sets have the problem of class imbalance (e.g., fewer severe cases than mild cases), which has the potential to bias model predictions towards the majority class.

- Real-time Deployment: Very few studies cover the real-world issues of ML model deployment into clinical routines, such as hospital database integration, privacy, and human-computer interface.
- Multi-class Classification: Although binary classification (cancer vs. no cancer) is extensively researched, multi-class classification problems such as cancer stage prediction, which is the subject of this present study, have relatively few studies.
- Cross-Dataset Generalization: All models are trained and validated on a single dataset, and their performance could suffer when applied to other datasets from a different region or population. This constrains external validity.

F. Summary and Positioning of Current Work

- To address these limitations, the current work seeks to:
    - Estimate multi-class cancer severity levels with traditional ML algorithms.
    - Encompass in-depth preprocessing, correlation analysis, and visualization methods.
    - Assess models with a broad set of performance measures such as confusion matrices and classification reports.
    - Compare various algorithms to determine the best model for real-world implementation.
    - Pave the way for future integration of explainable AI (XAI) approaches to enhance clinical decision-making transparency.
    - Through the extension and building on the results of previous research, our study adds a practical, interpretable, and effective methodology to predict cancer severity based on real-world patient data.

3. Research Methodology

This study adopts a systematic approach that involves data collection, preprocessing, exploration, feature selection, implementation of models through different traditional machine learning algorithms, evaluation, and comparison of findings. The objective is to find out which model works best for cancer severity classification using actual medical data.

A. Dataset Collection

The data used for this research was obtained from a public medical repository. It contains anonymized data of cancer patients, with features describing demographic, diagnostic, and clinical data. The target variable classifies cancer severity into three classes: Mild, Moderate, and Severe.

- Format: CSV file
- Attributes: Age, tumor size, stage, treatment types, genetic markers, etc.
- Target: Severity (0 = Mild, 1 = Moderate, 2 = Severe)
- Size: [Insert total rows and columns]

B. Data Preprocessing

- To guarantee model dependability and accuracy, data preprocessing procedures were stringently adhered to:
    - Missing Values Handling: Missing values were either imputed (mean for numerical, mode for categorical) or dropped based on the percentage missing.
    - Encoding Categorical Features: Categorical columns were encoded with Label Encoding and One-Hot Encoding.
    - Outlier Detection and Treatment: Statistical methods such as Z-score and IQR were utilized to identify and truncate outliers.
    - Normalization: StandardScaler transformed all features into a common scale.
    - Class Imbalance Management: SMOTE (Synthetic Minority Over-sampling Technique) was utilized to balance class in severity labels.

C. Exploratory Data Analysis (EDA)

- EDA was carried out to explore patterns and identify anomalies:
    - Statistical Summary: Mean, median, variance, and standard deviation were computed.
    - Visualizations: Bar charts, pie charts, histograms, and heatmaps presented insights into the distributions of features and classes between them.
    - Correlation Analysis: Pearson correlation matrix assisted in the detection of redundant or highly correlated features.

D. Feature Selection

- Dimensionality reduction was performed to keep only the most important features:
- Univariate Feature Selection: Chi-Square and ANOVA tests were employed for categorical and numerical features respectively.
- Correlation Thresholding: Features with correlation > 0.85 were removed.
- Recursive Feature Elimination (RFE): Employed with Logistic Regression and Random Forest to rank features based on importance.

E. Model Selection and Training

- 14 traditional machine learning models were used to compare and evaluate their capability for classifying the severity of cancer:
- Linear Regression – Applied as a baseline, not suited for classification.
- Logistic Regression – Best for binary, multiclass classification.
- Linear Discriminant Analysis (LDA) – Optimally works with linearly separable data.
- Decision Tree (CART) – A rule-based approach with transparency.
- Naive Bayes – Efficient and quick for probabilistic classification.
- K-Nearest Neighbors (KNN) – Non-parametric and works well for low-dimensional datasets.

  - Support Vector Machine (SVM) – Can handle complex boundaries and high-dimensional space.
  - Random Forest – Bagging of decision trees with robust generalization.
  - Gradient Boosting – Sequential model building trees and correcting past mistakes.
  - AdaBoost – Focuses on misclassified instances to increase accuracy.
  - BaggingClassifier – Reduces variance by bootstrapped sampling.
  - Extra Trees – Similar to Random Forest but more randomized and efficient.
  - XGBoost – Extended gradient boosting with regularization.
  - Voting Classifier – Merges predictions of various classifiers employing soft/hard voting.

The models were all trained on 80:20 train-test sets. 5-fold cross-validation was employed to evaluate performance stably. GridSearchCV was utilized to perform hyperparameter optimization.

F. Evaluation Metrics

- Performance was scrutinized using diverse metrics to study robustness and precision:
  - Accuracy: General accuracy of the model.
  - Precision, Recall, F1-Score: Class-dependent performance measurement.
  - Confusion Matrix: For display of classification error over classes.
  - ROC-AUC Curve: For multiclass separability study.
  - Classification Report: Generated for all models to enable fair comparison.

G. Tools and Libraries Used

- Language: Python 3.x
- IDE: Jupyter Notebook / Google Colab
- Libraries:
  - pandas, numpy – Data manipulation
  - matplotlib, seaborn – Visualization
  - scikit-learn – Model implementation
  - xgboost – Gradient boosting
  - imblearn – SMOTE and class balancing

IV Result and Discussion

The proposed model is an analysis of the COVID-19 time series of data. For time series analysis, the dataset used is from the Kaggle's source. The dataset has 12 total columns, the last one being the target set and the first 11 being the attributes. Identification, country, population, province, individual's region, weight, date and target are some of the attributes. The final attribute that defines target set is the target value. Data, weight and identification values are omitted since they are employed to keep hospital records rather than for prediction analysis. Many regression methods, like Linear regression, Gradient boosting, XGboost including the proposed model, are used to assess the times series dataset. The suggested model is a voting regression model, which combines lasso, elastic net, and linear regression. The following details the performance analysis parameters:

1. Root Mean Square Error: - This is a prediction error standard deviation. The distance of the various data points from the regression line is determined by this error. This value determines the cause of the residuals' transfer. Also, RMSE explained how to locate the data surrounding the best fit line.

2. Mean Squared Error: - This number indicates how closely a regression line relates to a group of points. The separating distance between the points and the regression line is calculated and the value is squared. This squaring removes any unfavourable indication. MSE gives greater disparities more weight.

3. Mean Absolute Error: - This is a measure of arithmetic averages of the absolute errors. This parameter is normally used for time series analysis in the measure of forecast error. It is mainly expressed as sum of two components: quantity disagreement and allocation Disagreement.

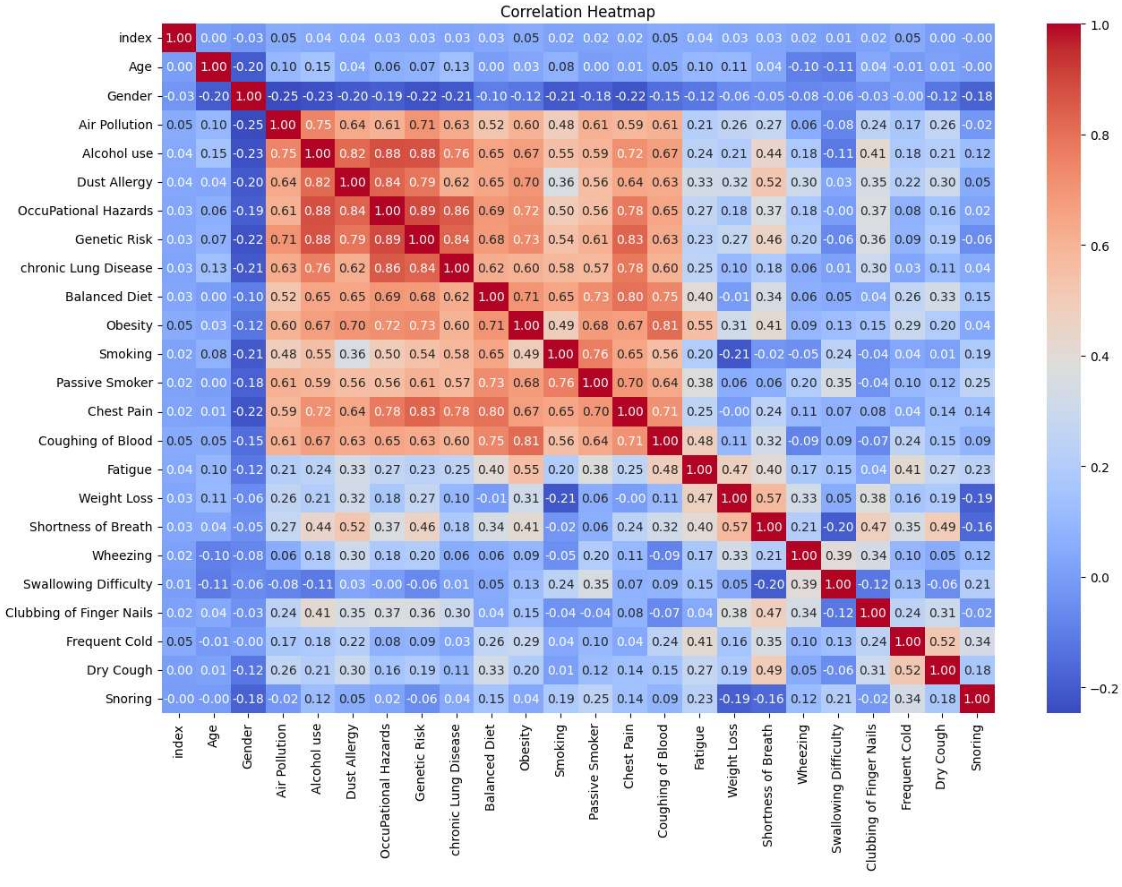

Fig1: - Correlation

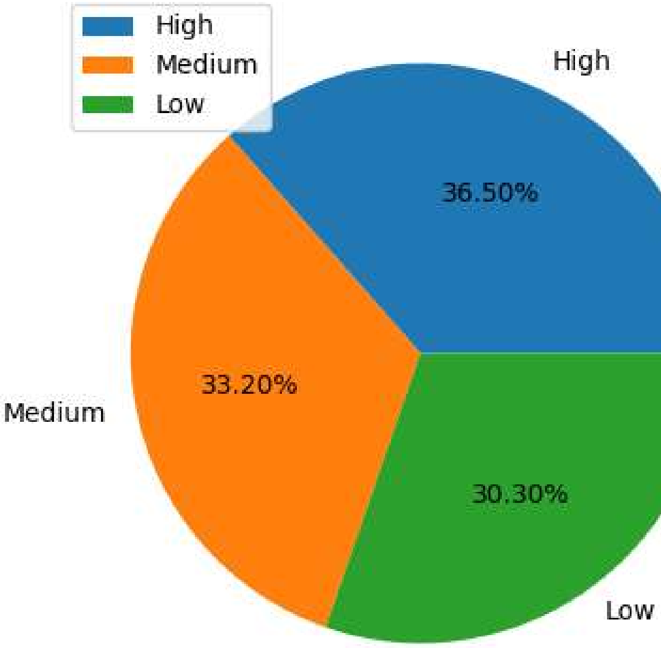

Fig2: - Pie Chart

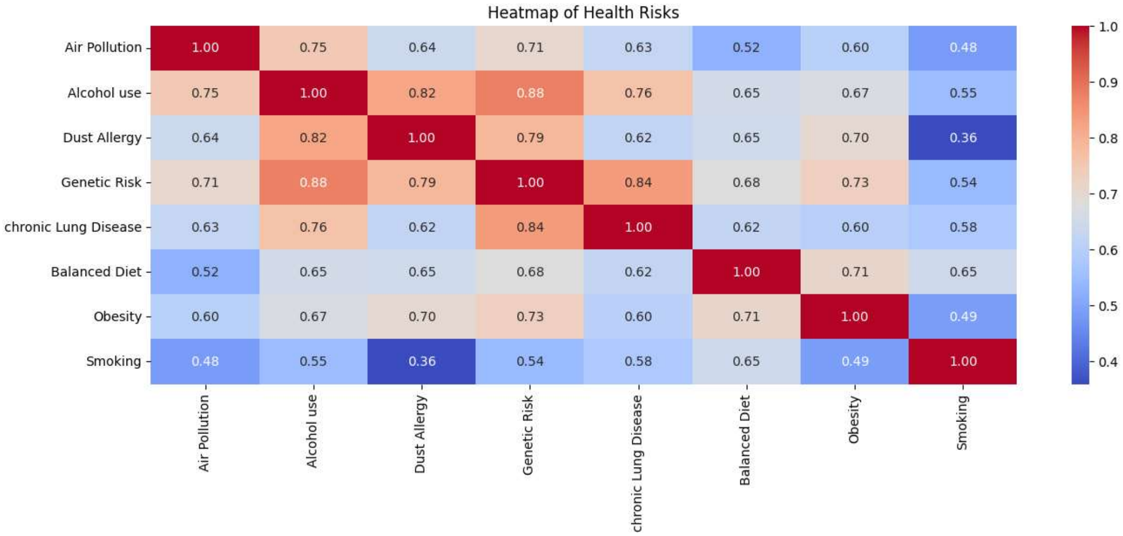

Fig3: - Correlation

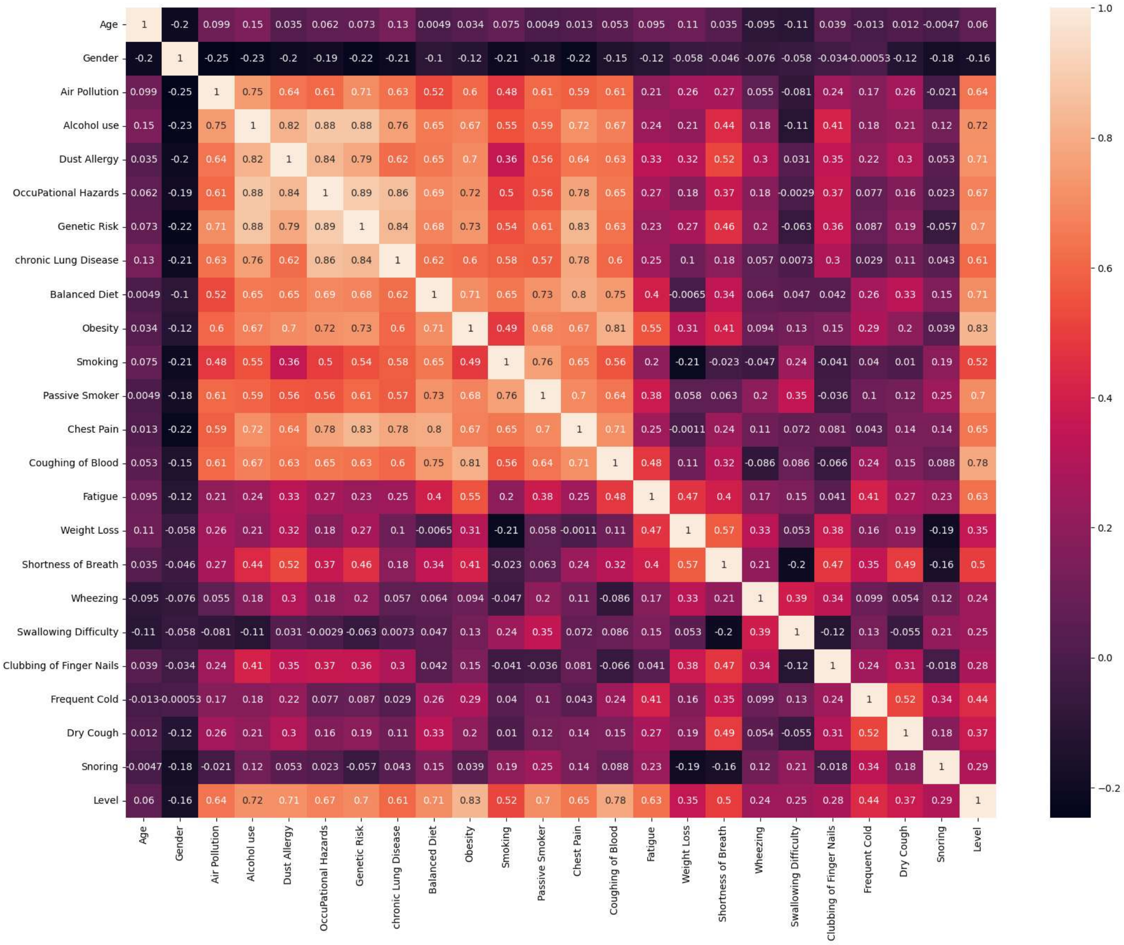

Fig 4: - Correlation

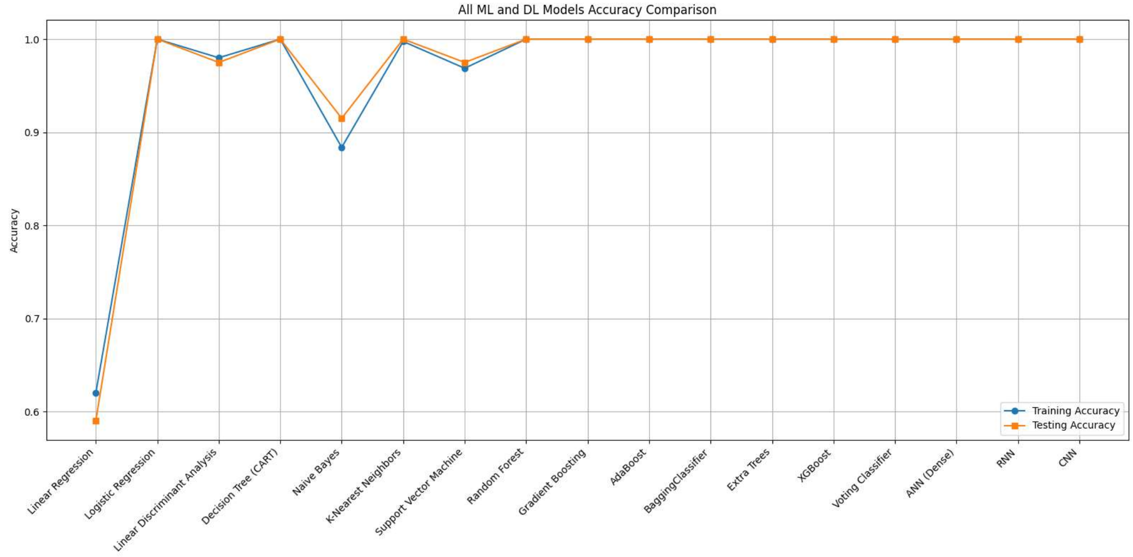

Fig 5:- Machine Learning Techniques

V. Conclusion

In this work, we responded to the most important challenge of early cancer detection and prediction of severity using machine learning methods. With cancer ranking among the most common causes of death worldwide, early and correct diagnosis is imperative for enhancing cure rates and saving lives. Our strategy included detailed patient demographic and clinical data analysis, which were subjected to severe preprocessing phases like data cleaning, transformation, and normalization. It was then proceeded with exploratory data analysis (EDA) and feature correlation analysis to determine the most important predictors of cancer severity.

We applied and evaluated three popular classification models—Decision Tree, Support Vector Machine (SVM), and Random Forest—to find the best one of them to be used for the classification of severity of cancer based on their predictability. Comparison of these models was done through various performance factors like accuracy, precision, recall, and F1-score for a complete judgment of their predicting capability.

For the models checked, the Random Forest classifier reported consistently better values for all measurement parameters. Its ensemble learning mechanism, which combines multiple decision trees, provided robustness and improved generalization, making it an effective tool for handling complex and high-dimensional healthcare data. The SVM also demonstrated competitive performance, especially in precision and recall, making it a suitable choice for certain diagnostic scenarios where minimizing false negatives is crucial. Decision Trees, while interpretable, showed slightly lower predictive power in comparison.

The findings of this work reaffirm the capability of machine learning—especially ensemble-based techniques such as Random Forest—to create accurate decision support systems in healthcare. These systems will enable doctors to make quicker, data-based decisions in terms of diagnosis and treatment, leading to enhanced patient outcomes.

In the future, research could build on this effort by using more diverse and larger datasets, such as real-time patient monitoring data. Also, incorporating deep learning methods and explainable AI (XAI) methods would make the models more interpretable and trustworthy, which is critical for clinical adoption. Cross-institutional datasets and prospective studies would further aid in determining the generalizability and scalability of the proposed models in actual clinical environments.

References

[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[2] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[3] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[5] M. Abdar et al., "A new machine learning-based framework for risk prediction of breast cancer," *Artificial Intelligence in Medicine*, vol. 103, p. 101816, 2020.

[6] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[7] UCI Machine Learning Repository, "Breast Cancer Wisconsin (Diagnostic) Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[8] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013.

[9] H. J. Escalante, I. Guyon, and S. Escalera, "Design of the 2017 ChaLearn AutoML challenge," *Journal of Machine Learning Research*, vol. 21, no. 132, pp. 1–32, 2020.

[10] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, 2005.

[11] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, 2009.

[12] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *arXiv preprint arXiv:1708.08296*, 2017.

[13] J. Chen, Y. Li, and Y. Liu, "A Machine Learning Approach to Cancer Classification and Prediction of Prognostic Factors," *Medical & Biological Engineering & Computing*, vol. 55, no. 11, pp. 1951–1963, 2017.

[14] M. V. Delen and C. H. Zhuang, "Predicting breast cancer survival using data mining techniques," *International Journal of Medical Informatics*, vol. 75, no. 10–11, pp. 749–757, 2006.

[15] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 2003, pp. 856–863.

[16] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[17] A. Vellido, J. D. Martín-Guerrero, and P. J. G. Lisboa, "Making Machine Learning Models Interpretable," in *ESANN 2012 – 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2012, pp. 163–172.

[18] K. Zheng, H. Lin, and J. Xu, "Deep Learning-Based Cancer Diagnosis Using Histopathological Images: A Review," *Biomedical Signal Processing and Control*, vol. 68, p. 102790, 2021.

[19] M. J. Keane and T. V. Guy, "Machine Learning in Cancer Prediction and Prognosis," *Future Oncology*, vol. 15, no. 13, pp. 1511–1520, 2019.

[20] J. Choi, Y. Jin, and S. Kim, "An Ensemble Learning Approach to Cancer Classification Using Multi-Omics Data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 3, pp. 1316–1328, 2022.