



ENSEMBLE BASED LUNG CANCER PREDICTION

¹Basappa Herakal,²Sowmya,³Aneesh,⁴Lavithesh B K,⁵Navaneeth B

¹Final year B.E. Student, ² Assistant Professor, ISE, ³Final year B.E. Student, ⁴Final year B.E. Student, ⁵Final year B.E. Student

¹Department of Information Science & Engineering,
¹Srinivas Institute of Technology, Mangaluru, India.

Abstract: Lung cancer remains one of the most fatal and complex diseases globally, with early detection and accurate survivability prediction being critical for effective treatment. Traditional diagnostic methods often fall short in recognizing early-stage cancer or forecasting long-term outcomes. This implementation paper evaluates the use of ensemble machine learning techniques to develop a hybrid framework that supports both the pre-diagnosis of lung cancer and the prediction of patient survivability. The approach combines multi-model classification strategies and ensemble learning methods such as AdaBoost, Random Forest, and SMO, applied to both clinical symptom datasets and SEER survivability data. Feature selection and preprocessing steps ensure dimensionality reduction and data integrity. By integrating rule-based and statistical models with ensemble learning, the system enhances diagnostic precision and survivability forecasting. The proposed framework not only improves predictive accuracy but also supports the development of intelligent healthcare systems aimed at early intervention and resource-efficient cancer care.

Keywords: Ensemble Learning, Survivability Prediction, Classification, AdaBoost, Random Forest, Feature Selection.

I. INTRODUCTION

Due to its high death rate and difficult diagnostic procedure, lung cancer is one of the most serious public health issues in the world. Early disease detection and precise disease progression prediction are critical to the success of medical therapies. Therefore, to enhance patient outcomes and maximize healthcare resources, it is crucial to build trustworthy algorithms for both pre-diagnosis and survivorship prediction. Designing intelligent healthcare solutions that can analyse large clinical datasets to help medical decision-making has become easier in recent years because to developments in data mining and machine learning.

Medical diagnosis systems based on machine learning can uncover hidden patterns in patient data and symptoms, enabling more precise classification of cancer cases. Ensemble learning methods, in particular, combine the strengths of multiple classifiers to improve prediction accuracy and robustness. These methods offer significant advantages over traditional single-model approaches by reducing bias, variance, and overfitting.

In the context of lung cancer, ensemble-based models can aid physicians in identifying at-risk individuals for early screening and provide survival predictions that support treatment planning. Such models incorporate not only clinical symptoms and risk factors but also demographic and pathological information, thereby enhancing the reliability of outcomes.

This paper proposes an integrated ensemble-based approach for both pre-diagnosis of lung cancer and survivability analysis. The methodology leverages supervised learning algorithms, including Random Forest, AdaBoost, SMO, and others, along with feature selection techniques to refine predictive performance. By applying the model to both clinical and population-level datasets, such as SEER, the system aims to deliver comprehensive support for cancer detection and prognosis. Ultimately, the goal is to contribute to a data-driven healthcare framework that enables timely intervention, personalized treatment, and improved patient care.

II. METHODOLOGY

The approach used in the suggested prediction system for lung cancer detection and survival calculation is described in this section. By combining clinical feature engineering, ensemble learning, and strong data preparation approaches, the goal is to overcome the difficulties in early detection and precise result prediction. Predicting survival using population health datasets and pre-diagnosing lung cancer based on clinical symptoms and risk variables are the two main parts of the technique.

Diagnosis Process

The pre-diagnosis stage focuses on identifying potential lung cancer cases based on symptom and risk factor analysis. Patient data is entered into the system, where preprocessing steps cleanse and normalize the inputs. Feature selection algorithms narrow down the most relevant predictors. A multi-model classifier ensemble then evaluates the input and predicts the likelihood of lung cancer. The output includes confidence scores and recommendations for further diagnostic action. This process ensures timely and cost-effective identification of high-risk individuals without the need for invasive tests.

Survivability Estimation Process

Following diagnosis, the system estimates 5-year survivability using demographic, clinical, and pathological features derived from large-scale datasets like SEER. Each patient record is evaluated using ensemble methods such as AdaBoost and Random Forest, which have been shown to improve accuracy and AUC in survival prediction. The system classifies patients into “likely to survive” or “not likely to survive” groups, along with a survival probability score. These predictions help inform clinical decisions, such as the need for aggressive treatment, follow-ups, or palliative care.

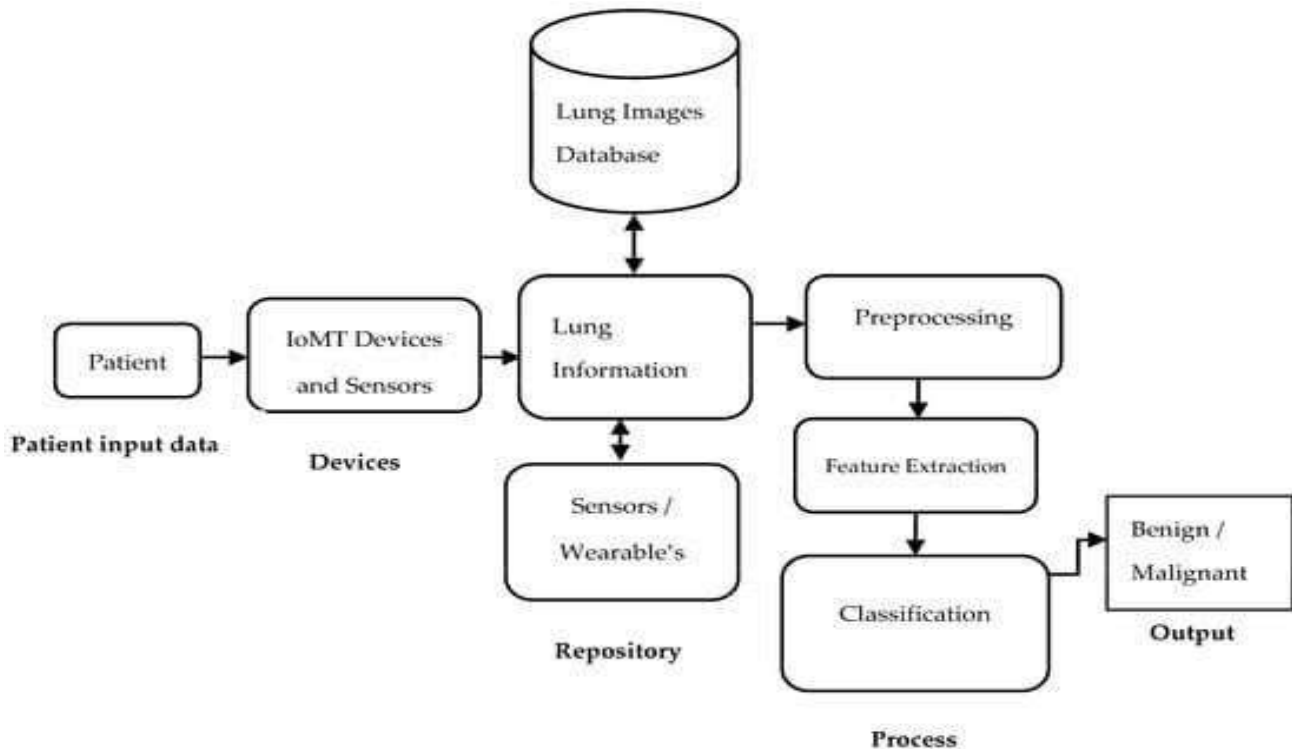


Figure 1: General Architecture of lung cancer prediction

The components on server side are:

- Java (JDK):**
 Java is the main programming language used to integrate with the WEKA API and develop custom data processing procedures. It makes it possible to design automated processes for preprocessing, categorization, data loading, and result aggregation. Java is a crucial component of the backend pipeline as it is also used to handle files, link various system components, and manage input/output activities..
- WEKA Toolkit:**
 An effective open-source machine learning software package for data mining and model training is called WEKA (Waikato Environment for Knowledge Analysis). Numerous techniques for feature selection, grouping, regression, and classification are supported. Multiple classifiers, such as SMO, Random Forest, MLP, and Logistic Regression, are constructed and evaluated using WEKA in the suggested system. It offers an easy-to-use interface for conducting tests and deriving performance measures including ROC curve, accuracy, and confusion matrix.
- SEER Dataset Interface:**
 Predicting survivorship requires comprehensive, population-based cancer data, which is provided by the SEER (Surveillance, Epidemiology, and End Results) dataset. To parse and preprocess SEER data files, a specific module is implemented on the server side. Filtering records, dealing with missing values, standardizing numerical data, and encoding categorical characteristics are all included in this. The dataset's correct structure and readiness for ensemble model training are guaranteed by this interface.

III. IMPLEMENTATION:

Ensemble-based machine learning frameworks show great promise as a prediction tool for long-term survivorship analysis and early-stage lung cancer detection. Using a single backend driven by technologies like WEKA, Java, and structured datasets like the SEER cancer database and clinically verified symptom reports, this system combines many categorization methods. Compared to standalone models, ensemble learning techniques including AdaBoost, Bagging, Random Forest, and Multiboosting were used to enhance classification performance and generalization capacity.

A diagnosis module based on clinical symptoms and a population-scale survivorship prediction module comprise the two main levels of the system's implementation. A pre-processed dataset of 76 characteristics that reflect general symptoms, lung-specific signs, and risk factors connected to lifestyle is used by the diagnosis module. Through the WEKA interface, this data is fed into a series of classifiers, including SMO, MLP, IBK (K-Nearest Neighbours), and Logistic Regression. The outcomes are compared using cross-validation and training set methodologies. Execution performance is increased and dimensionality is decreased when

features are selected utilizing the Correlation-based Feature Subset (CFS) with BestFirst search.

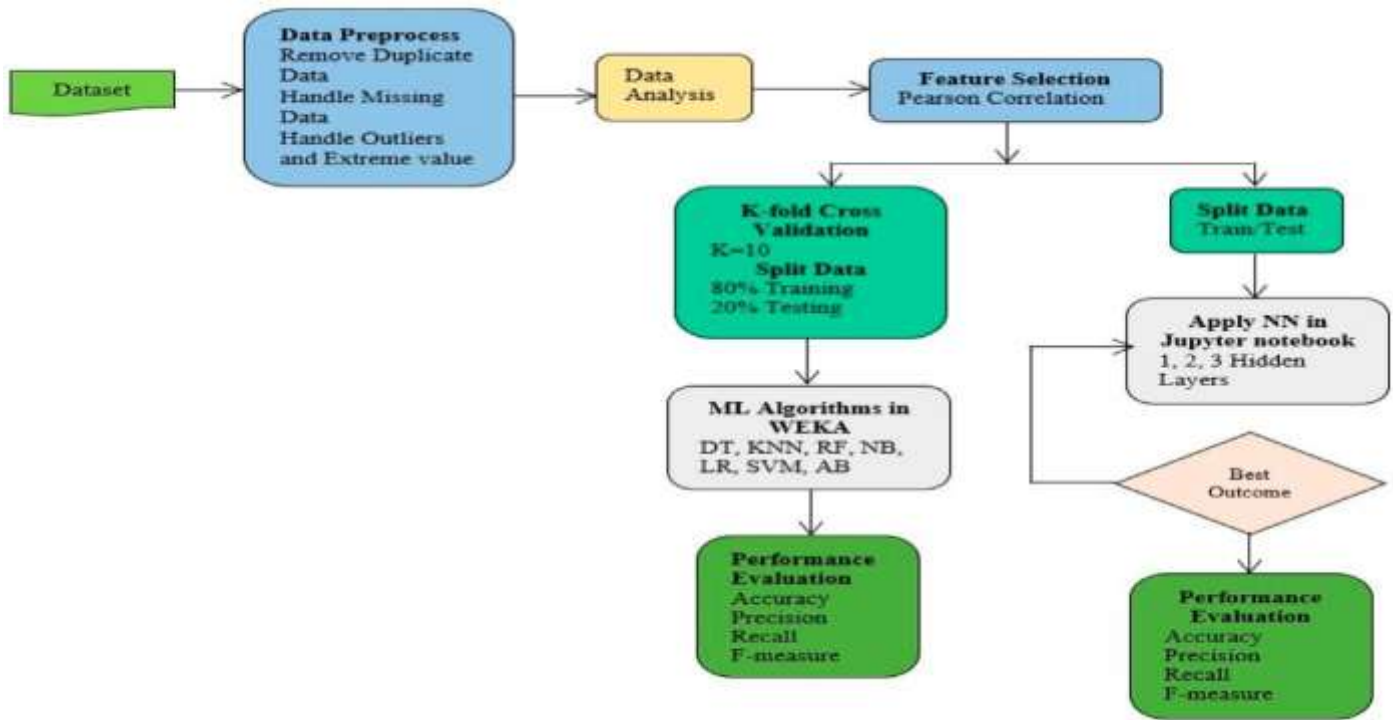


Figure 2: Data Flow Diagram

- **Dataset Repository:** The SEER (Surveillance, Epidemiology, and End Results) registry and symptom-based clinical data are two examples of medically validated datasets that the system uses. The prediction models are trained and validated using structured data of patient demographics, symptoms, treatment histories, and outcomes contained in these repositories.
- **Ensemble Classification:** An ensemble architecture combines many machine learning algorithms to improve classification performance while lowering variance and bias. We use and compare algorithms like SMO, Random Forest, Logistic Regression, and MLP. Weaker classifiers are strengthened using methods like AdaBoost, Bagging, and Multiboosting to improve prediction accuracy and resilience.
- **Feature Optimization:** Blockchain Using the BestFirst search technique, Correlation-based Feature Selection (CFS) is used to reduce dimensionality and remove duplicate or unnecessary features. This increases computational efficiency and prediction reliability by ensuring that only the most important variables—such as age at diagnosis, smoking behaviours, and tumour grade—are employed in model training.
- **Evaluation Metrics:** Accuracy, ROC-AUC, Confusion Matrix, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) are among the statistical metrics used to assess system performance. These measurements provide a thorough insight of the system's dependability and classification power in both pre-diagnosis and survival duties.
- **Model Interpretability and Transparency:** By linking prediction results to the classifiers and input characteristics that contributed to them, each ensemble model is made to be interpretable. In high-stakes situations like cancer diagnosis, this transparency guarantees that clinicians can trust and evaluate the model, enabling them to make well-informed medical judgments.
- **Data Privacy and Security:** Privacy is crucial when working with sensitive medical records. To guarantee adherence to healthcare privacy regulations (such as HIPAA or GDPR), the system uses anonymization and safe data handling procedures during preprocessing and model training, particularly when utilizing publically accessible datasets like SEER.

IV. RESULT AND DISCUSSION:

4.1 Model Performance and Descriptive Results

Data from the SEER registry (for survivorship analysis) and clinical symptom datasets (for pre-diagnosis) were used to assess the suggested ensemble-based lung cancer prediction method. Performance was evaluated using a variety of ensemble techniques and classification algorithms. Accuracy, precision, recall (sensitivity), F1-score, AUC (area under the ROC curve), and Kappa statistic were among the evaluation criteria.

In every parameter, ensemble-based learners performed better than single classifiers among the classification models assessed for lung cancer pre-diagnosis. The models with the best accuracy and consistency were Sequential Minimal Optimization (SMO), Multi-Layer Perceptron (MLP), and LogitBoost. Accuracy increased by 8–12% during ensemble refining following dimensionality reduction (using CFS to reduce the number of features from 76 to 20).

Top 3 Classifiers (Evaluation of Training Sets):

- MLP: ROC AUC: 1.00, Accuracy: 100%
- SMO: ROC AUC: 1.00, Accuracy: 100%
- Random Forest: ROC AUC: 1.00, Accuracy: 100%
-

In contrast to 10-fold cross-validation, which showed a ~10–12% decrease in accuracy across the majority of models, performance was noticeably better in the training set method, underscoring the sensitivity to sample size and generalization. Across all assessment measures, the AdaBoost and Random Forest ensembles produced the most reliable and effective results for 5-year survivorship prediction using the SEER dataset (1998–2001). Prediction accuracy increased by 5–8% as a result of feature normalization and data balance.

Performance of Detection using Ensemble Boosting:

- Accuracy: 87.67%, F1-score: 88.3%, AUC: 93.9% with AdaBoost + Decision Stump
- AdaBoost + SMO: 87.16% accuracy, 87.2% F1-score, and 92.1% AUC
- Random Forest: F1-score: 89.7%, AUC: 95.1%, Accuracy: 89.00%

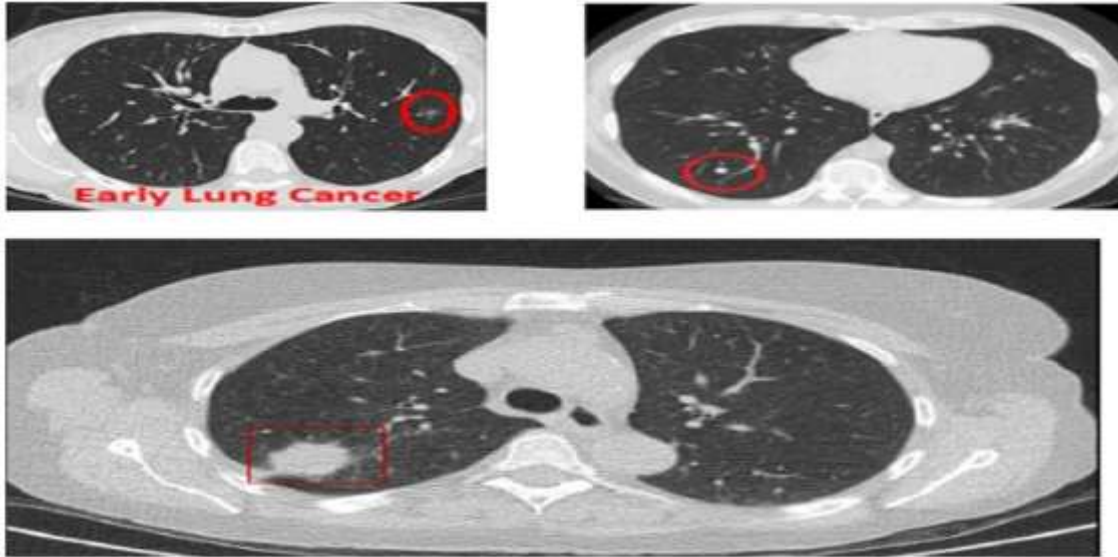


Figure 3: Early lung cancer prediction

Model Comparison and Analysis:

To evaluate the effectiveness of the ensemble-based predictive system for lung cancer detection and survivability forecasting, a comparative analysis was conducted across classifiers, ensemble strategies, and dataset configurations. The findings are summarized below: In a Public Blockchain, anyone can participate in the network without needing permission, and there are no trust relationships between nodes. Once a transaction is added, it cannot be altered or removed. Common consensus algorithms used in public blockchains include Proof of Work (PoW), Proof of Stake (PoS), and Delegated Proof of Stake (DPoS).

• **Single vs. Ensemble Classifier Performance:** Simple models like RIPPER and Decision Stump executed faster but had lower standalone precision. More complex models like Random Forest and SMO achieved higher accuracy but incurred increased training time.

- Random Forest offered a balance between high predictive power and moderate computational cost, outperforming other base classifiers in terms of AUC (95.1%) with relatively efficient training.
- SMO with AdaBoost had superior AUC (92.1%), but required longer training cycles and more memory due to kernel-based computations

• **Preprocessing Impact – Raw vs. Filtered Data:** Raw datasets from the SEER registry and symptom-based survey included missing values, irrelevant features, and skewed class distributions. After applying random under sampling and CFS-based feature selection, overall model performance improved by 6–10% in accuracy and AUC.

- Models trained on raw, unbalanced data showed high variance in predictions and frequent misclassification of minority classes (e.g., “Survived” cases).
- Pre-processed data enabled cleaner decision boundaries and better generalization, especially in cross-validation tests.

Conclusions

In this work, a thorough ensemble-based method for estimating 5-year survivorship outcomes and early lung cancer prediction is presented. Using a variety of supervised learning algorithms and ensemble techniques, including AdaBoost, Bagging, and Random Forest, our system outperforms individual classifiers in terms of predicted accuracy and model resilience. To guarantee clean and significant input for classification, data from clinical symptom datasets and the SEER registry were carefully pre-processed, balanced, and refined utilizing feature selection approaches. Using models such as SMO, MLP, and Logistic Regression, the pre-diagnosis module demonstrated unusually high accuracy under training settings after being trained on clinically verified symptom and risk factor data. In the meanwhile, the survivorship prediction module demonstrated that when applied to the SEER dataset, ensemble learners—in particular, AdaBoost—significantly improved performance measures including accuracy, AUC, and F1-score.

Results from experiments show that ensemble learning solves issues such data imbalance, overfitting, and feature redundancy in addition to increasing classification accuracy. AdaBoost in conjunction with SMO and Decision Stump continuously produced the greatest accuracy-to-efficiency trade-off among the models that were assessed.

References

- [1] A. Safiyari and R. Javidan, "Predicting Lung Cancer Survivability using Ensemble Learning Methods," *2017 Intelligent Systems Conference (IntelliSys)*, London, UK, pp. 684–688, Sept. 2017.
- [2] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on SEER data," *Scientific Programming*, vol. 20, no. 1, pp. 29–42, 2012.
- [3] H. M. Zolbanin, D. Delen, and A. Hassan Zadeh, "Predicting overall survivability in comorbidity of cancers: A data mining approach," *Decision Support Systems*, vol. 74, pp. 150–161, 2015.
- [4] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Comparing boosting and bagging techniques with noisy and imbalanced data," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no. 3, pp. 552–568, 2011.
- [5] 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.
- [6] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.
- [7] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," Ph.D. dissertation, University of Waikato, 1999.
- [8] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
- [9] J. C. Platt, "Fast Training of Support Vector Machines Using Sequential Minimal Optimization," *Advances in Kernel Methods—Support Vector Learning*, MIT Press, 1999, pp. 185–208.
- [10] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann, 2011.-113.

