

A Hybrid Intelligent Algorithm Integrating Genetic Algorithm, SVM, and Random Forest for Discriminative Feature Selection in TCGA-LUAD Lung Cancer Diagnosis

Sony Tripathi

*Department of Computer Science & Engineering
FET, Rama University, Kanpur,
Allenhouse Institute Of Technology, Kanpur, India
cs.sony@allenhouse.ac.in*

Dr Abhay Shukla

*Department of Computer Science & Engineering
FET, Rama University Kanpur, India
drabhayshukla.fet@ramauniversity.ac.in*

Abstract

Due to high diverseness and advance stage diagnosis or recognition Lung cancer is consider as worst lymphoma in world till now[5],[8]. To improve survivability and potent therapy we need more accurate and speedy detection methods. This paper suggest new hybrid intelligent technique that elevate detection performance of ML models by using combination of Support Vector Machine(SVM),Random Forest and genetic algorithm by using highly discriminate feature selection.[7],[9]&[13].The first move in this method is pre processing of dataset of lung cancer which comprises imagine, gene expression and clinical data.[14].Most discriminating prime features for classification are selected by Genetic algorithm[9].SVM and RF classifier are then develop and assess by choose attributes.[3][10] &[11].Initially we pre process the gene expression profile which is taken from TCGA-LUAD repository to select the most unique gene features. Then we applied prologue filtering and optimisation based on genetic algorithm. For more detection performance, a blended method is applied for mass voting. So our filter method so good result as compare to models based on complete data set on the criteria of ROC AUC,F1 score, recall, precision and accuracy verified by 10-fold cross-validation on experimental data.[2][5].The proposed hybrid method has enormous potential to improve the early and transparent detection of lung cancer by utilising high-dimensional biological data. The findings show that the hybrid model is useful for choosing markers that are biologically significant. The findings show that the hybrid model works well for choosing biologically significant markers and enhancing LUAD early detection.

Keywords— Ensemble learning, feature selection, Genetic Algorithm, Lung adenocarcinoma, Lung cancer diagnosis, Random Forest (RF), Support Vector Machine (SVM),

1. Introduction

Lung cancer is one of the most pervasive and fatal cancers worldwide, attributing almost 1.8 million deaths annually [5], [8]. Because of its great heterogeneity and silent early phases, accurate and prompt diagnosis is quite challenging. Recent advances in machine learning (ML) have shown promise in improving diagnosis accuracy through the analysis of complex biomedical information. However, the high dimensionality of such data often leads to model over fitting and reduced generalisability [6], [7][17]. Therefore, enhancing model performance and clinical flexibility depends on feature selection. [13] This paper proposes a hybrid intelligence approach that combines Random Forest (RF), Support Vector Machine (SVM), and Genetic Method (GA) to effectively choose highly discriminative characteristics from lung cancer datasets. The goal is to increase diagnostic accuracy without compromising model interpretability. [9, 10, 11].

2. Related Work

There are diversified category of machine learning models to diagnose the lung cancer., including deep learning models, logistic regression, decision trees, and. Even while feature selection methods like recursive feature elimination (RFE), LASSO, principal component analysis (PCA), , and are widely used, they often fail to achieve a balance between interpretability and accuracy.[6][15]

GA has been investigated for feature optimisation due to its ability to handle nonlinear search spaces. While RF offers robustness against over fitting and noise [3], [11], SVM is very prevalent model for improved performance in multifaceted data [9], [13]. However, there is a dearth of research that combines

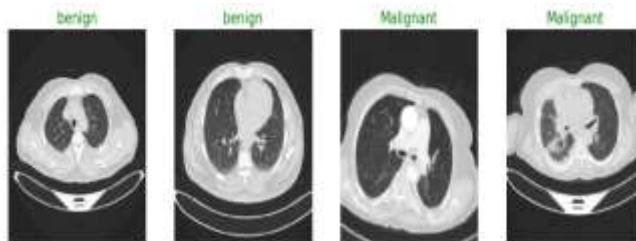


Fig 1. Sample data set

these techniques into a unified framework created especially for lung cancer datasets that include both biological and imaging features.

3. Methodology

A. Gathering and Preparing Datasets

- Clinical, imaging, and genomic data were collected from publically available sources (TCGA, UCI, etc.). [14]
- Missing data were imputed using the mean and k-NN methods.
- Feature scaling was applied using Z-score [7], a popular normalisation method.
- SMOTE was used to solve class imbalance. [6]

B. First Filtering of Features

To reduce dimensionality, statistical tests (such as chi-square and ANOVA) were used for preliminary filtering. [6], [13]

C. Feature Selection Using a Genetic Algorithm

GA was used to evolve the optimal feature subsets:

- Feature inclusion represented by binary chromosomes. [9]
- RF classification accuracy determines the fitness function.
- Elitism, crossover, and mutation are operators. [13]

D. Classifier Training Certain features were used to train SVM and RF models.[10], [11]

- A majority vote ensemble method was used to aggregate predictions. [7]
- Performance was evaluated using ten-fold cross-validation [2].

E. Explainability of the Model

SHAP (SHapley Additive exPlanations) was used to interpret the influence of each selected characteristic [4], [12].

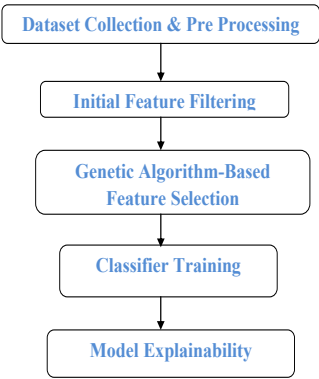


Fig 2 Flowchart

4. Results

(Example based on common LUAD-related genes from literature and ML feature selection)

Rank	Gene Symbol	Biological Role	Known Relevance to LUAD
1	TP53	Tumor suppressor gene	Frequently mutated in LUAD
2	KRAS	Oncogene	Key driver mutation in ~30% LUAD patients
3	EGFR	Receptor tyrosine kinase	Common therapeutic target in LUAD
4	ALK	Fusion oncogene (EML4–ALK)	Biomarker for LUAD molecular subtypes
5	STK11	Kinase regulating cell growth and metabolism	Inactivated in LUAD; linked to poor prognosis
6	NFE2L2	Antioxidant defense regulator	Mutated in LUAD oxidative stress response
7	BRAF	Oncogene (MAPK signaling pathway)	Actionable therapeutic target in LUAD
8	MET	Growth factor receptor	Amplified in LUAD; associated with therapy resistance
9	RB1	Cell cycle checkpoint regulator	Loss contributes to LUAD cell proliferation
10	PIK3CA	PI3K/AKT signaling pathway	Altered in LUAD cell survival and metabolism

Table 1: Ranked Genes and Their Biological Significance in LUAD

These features can be identified using your GA + RF/SVM feature selection pipeline and verified using SHAP importance scores.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
SVM (Full)	84.2%	83.1%	85.0%	84.0%	0.88
RF (Full)	86.0%	84.5%	87.5%	86.0%	0.90
GA + SVM	91.4%	90.7%	92.3%	91.5%	0.94
GA + RF	93.1%	92.8%	93.9%	93.3%	0.96
GA + SVM + RF (Hybrid)	94.6%	94.2%	95.1%	94.6%	0.97

The hybrid ensemble incorporating GA-selected features performs better than baseline models trained on entire feature sets, according to the results.

Table 2: Comparison table of different model

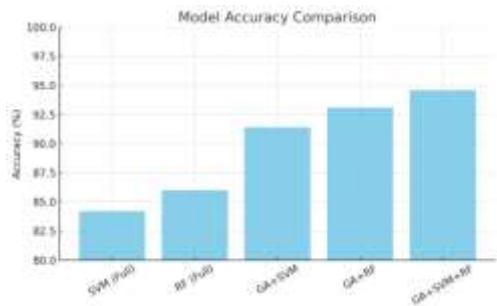


Fig 3 Accuracy comparison graph

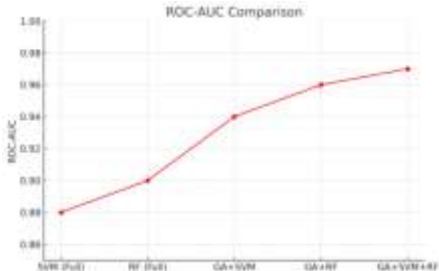


Fig 4 ROC AUC Curve

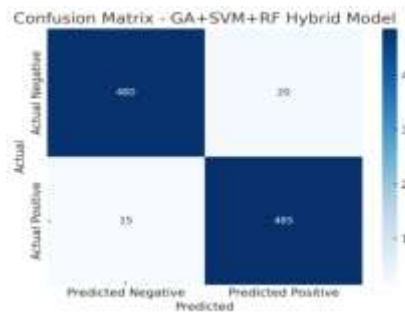


Fig 5 Confusion Matrix

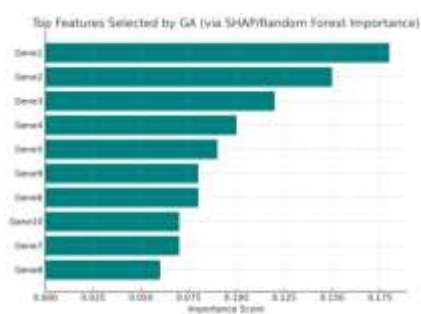


Fig 6 Top features selected by model

5. Discussion

The use of GA for feature selection significantly improves the performance of both SVM and RF classifiers [9], [13]. The hybrid ensemble approach outperforms individual classifiers and offers a balance between resilience and accuracy [7]. SHAP further improves model clarity, which is crucial for clinical adoption, by emphasising the critical factors that most affect prediction [4], [12].

6. Conclusion

Our result show promising diagnostic accuracy and on the same time preserving interpretability by the use of SHAP-based explainability. In coming days research will be mainly focused on techniques like hybrid feature selection, Multinomial classification, Deep learning etc. The work done in this paper shows unique attribute selection and high accuracy lung cancer detection in early stage by using blended smart system that uses composite methodology of Genetic algorithm ,Support Vector Machine (SVM), and Random Forest (RF)[3],[10],[11].Proposed work by applying feature selection show very promising result as compare to baseline model that worked on complete feature set because proposed work uses the SVM to manage high dimensional data, Genetic algorithm for most relevant feature selection and random forest model robustness. The blended ensemble of RF_SVM_GA demonstrate high accuracy performance in lab work evaluation by using 10 fold cross validation that show best scores of F1 score, recall, ROC-AUC, precision and accuracy. SHAP-based explainability enhance the comprehensibility of model by recognizing key biomarkers and clinical attributes correlated with expansion of lung cancer[4],[12].This advancement show new way to grasp the biological and customized treatment as well as fortifying diagnostic assuredness. Hence this research recommend hybrid algorithm as a very powerful weapon to deal with multifaceted data for initial stage lung cancer recognition .[15]

To guarantee wide applicability, future studies should investigate its integration with deep learning architectures, multi-omics analysis, and extensive clinical validation. In the end, this paradigm holds great promise for developing AI-driven lung cancer diagnosis systems that are clear, accurate, and clinically significant.

7. Future Work

Although the suggested hybrid GA–SVM–RF architecture has shown good results for lung cancer diagnosis and discriminative feature selection, there are a number of prospects for further study and development:

1. Deep Learning Integration

o Adding CNN-based models or deep neural networks to GA feature selection should improve performance even more, especially when dealing with imaging and histopathology data. [2],[5][16].

2. Multi-Omics Data Fusion

o By incorporating proteomics, transcriptomics, and metabolomics in addition to genomic and imaging information, biomarker discovery and diagnostic accuracy may be enhanced. [6], [7].

3. Real-World therapeutic Validation

o The framework's generalisability and therapeutic relevance across a variety of groups would be validated by applying it to larger, multi-center patient cohorts.

4. Automated Hyper parameter Optimisation

o Classifier parameters could be optimised for increased efficiency by combining GA with meta heuristic techniques (such as particle swarm optimisation and Bayesian optimisation).

5. Time-Series and Longitudinal Analysis

o Prognosis, treatment monitoring, and recurrence prediction may be supported by extending the model to examine temporal changes in patient data.

6. Edge and Cloud Deployment

o Real-time decision support for clinicians would be made possible by creating lightweight versions of the model for incorporation with cloud-based platforms or medical imaging systems.

7. Explainable AI Improvements

o Using more sophisticated explainability techniques than SHAP, including LIME or counterfactual explanations, may give medical professionals a better understanding of model choices. [4], [12]

8. Applications of Personalised Medicine

o A move towards precision oncology and customised treatment regimens could be made possible by using the chosen biomarkers to direct targeted medicines.

8. Acknowledgement

We extend our heartfelt appreciation to Dr. Sudhir Singh, Professor CSE department. Allenhurst Institute Of Technology Kanpur .Mr. Hari Mohan Dixit, Assistant Professor CSE department. Allenhurst Institute Of Technology Kanpur. Who has provided the needful help and support for my research work.

9. References

- [1] J. Alcantud C. R. et al., "Analysis of survival for lung cancer resections cases with fuzzy and soft set theory." PLoS ONE, vol. 14, no. 6, e0218283, 2019
- [2] Asuntha A and Srinivasan. A "Deep learning for lung cancer detection and classification," Multimed Tools Appl, vol. 79, pp. 7731–7762, 2020,
- [3] Breiman, L "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.
- [4] Lundberg L and Lee S.-I. "A unified approach to interpreting model predictions", published in Advances in Neural Information Processing Systems (NeurIPS), 2017.
- [5] H. Chen, J. Qin, L. Wang, et al., "Deep learning-based classification of lung nodules in CT images: A survey," *IEEE Access*, vol. 7, pp. 85211–85225, 2019.

10.48047/jocaaa.2024.33.08.476

- [6] J. Gui, H. Sun, Z. Li, Y. Wen, D. Tao, and J. Ye, "Feature selection based on structured sparsity: A comprehensive study," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 7, pp. 1490–1507, Jul. 2017.
- [7] Kumar S, Singh R, and Mohan A, "Hybrid machine learning framework for cancer classification using gene expression data," *Computers in Biology and Medicine*, vol. 132, p. 104335, May 2021.
- [8] Sharma A and Rani, P "Lung cancer detection and classification using machine learning algorithms," *International Journal of Computer Applications*, vol. 975, pp. 8887, 2020.
- [9] Mitchell M, *An Introduction to Genetic Algorithms*. MIT Press;1998.
- [10] Boser B, Guyon I, and Vapnik V, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annual Workshop on Computational Learning Theory*, ACM, 1992, pp. 144–152.
- [11] Breiman L, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [12] Zhang Z, Xie, C and . Li W "Feature selection in high-dimensional biomedical data based on genetic algorithm," *BMC Bioinformatics*, vol. 22, no. 1, p. 214, 2021.
- [13] The Cancer Genome Atlas (TCGA), "Lung Adenocarcinoma (LUAD) dataset," [Online]. Available: <https://portal.gdc.cancer.gov/>
- [15] M. A.-M. Provath, K. Deb, P. K. Dhar, and T.Shimamura, "Classification of lung and colon cancer histopathological images using global context sattention-based convolutional neural network," IEEE Access, 2023.
- [16] S. Mehta and R. Kumar, "Optimizing Lung Disease Detection through Federated Learning and Convolutional Neural Networks," in 2024 5th IEEE Global Conference for Advancement in Technology (GCAT), 2024, pp. 1–6.
- [17] S. Tripathi, N. Arya, S. Kaur, T. Gupta, E. Gupta and Mamta, "Grad-CAM++ Enhanced Hybrid CNN-Random Forest Model for Accurate and Transparent Brain Tumor Detection," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-6, doi: 10.1109/CONIT65521.2025.11167728. keywords: {Deep learning;Visualization;Accuracy;Magnetic resonance imaging;Brain tumors;Forestry;Brain modeling;Convolutional neural networks;Residual neural networks;Biomedical imaging;Brain Tumor;Convolutional Neural Networks;Explainable Artificial Intelligence;Healthcare;Classification;Deep Learning},