International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Identification of Lung Cancer Using Ensemble Methods Based on Gene Expression Data

K. Mary Sudha Rani[1]*, Dr. V. Kamakshi Prasad[2]

Submitted: 28/05/2023 **Revised**: 06/07/2023 **Accepted**: 25/07/2023

Abstract: Lung cancer is consistently classified as the most dangerous form of the disease since the beginning of recorded history. Patients with lung cancer who receive appropriate medical care, such as a low-dose CT scan, have a far better chance of survival since the disease is detected and diagnosed early. Nonetheless, there are certain drawbacks to this attempt. The gene expression level in hundreds of genes or cells within each tissue may now be determined because of developments in DNA microarray technology. Even though machine learning (ML) is rapidly being used in the medical field for lung cancer detection, the shortage of interpretability of these models remains a significant hurdle. Machine learning can be used to analyze gene expression data (DNA microarray) to predict whether or not a patient has lung cancer. The Collective Random Forest and Adaptive Boosting were employed to determine who was responsible for the harm. KPCA, or Kernel principal component analysis, was used for the feature reduction procedure. We calculated the correlation between each feature and the target using the statistical parameters provided by KPCA. Determining the proportion of the correct predictions for a given data set is one way to calculate the accuracy of a classification model. We tested the validity of the proposed technique in this work using a dataset including information about lung cancer. The dataset includes GSE4115 from the Gene Expression Omnibus (GEO) database, as well as the expression profiles it contains. The findings demonstrate the Identification of Lung Cancer (IOLC) model's potential to detect lung cancer in terms of accuracy, precision, recall, F-Measure, and error rate, with results indicating an accuracy of 81%, the precision of 81.2%, recall of 78.9%, F-Measure of 77.7%, and error rate of 0.29%, respectively.

Keywords: *Gene Expression, Lung cancer, Ensemble machine learning Random forest, AdaBoost*

1. Introduction

Cancer is a disease that causes cell destruction in the body. Cells develop and increase in a controlled manner; nevertheless, this control may fail if an error occurs in the cell's genetic blueprints. A variety of factors can cause this mistake. Lung cancer is the most common and lethal malignant tumor seen worldwide. In 2012, around 1,800,000 new lung cancer cases were detected, with 1,600,000 people dying due to the condition. Lung cancer is more common in women and is the leading cause of cancer death. Although smoking is the primary cause of lung cancer, around 15% of male and 53% of female lung cancer patients did not smoke. Furthermore, it is estimated that 25% of lung cancer patients worldwide did not get the disease due to smoking. Previously, the primary resource in biology was gene networks GNs [1], commonly depicted as graphs with nodes and rods, with nodes representing genes and rods signifying gene interactions. These rods may be assigned a numerical number or weight based on the strength of the relationships between the parties involved. As a result, GNs can uncover genes linked with biological processes and their interactions, providing a complete picture of the processes under inquiry. GNs are widely utilized in many fields,inquiry. GNs are widely utilized in many fields, including but not limited to biology, healthcare, and bioinformatics.

Furthermore, when it comes to non-smokers have a different carcinogenic pathway, clinic pathological features, epidemiology, and natural history than smokers. Lung cancer is the most common type diagnosed worldwide and the leading cause of cancer fatalities wheezing, hoarseness, chest tightness, coughing, and spitting up blood are all indications of lung cancer. Indications and symptoms include chest discomfort, shortness of breath, and wheezing [2]. To avoid this dreadful situation, we require machine learning algorithms to aid in the early detection and prevention of lung cancer. Treatments can be more effective and less likely to recur if started early in lung cancer [3]. As a result, preventative lung cancer screening and detection may be therapeutically beneficial, particularly for patients with undiagnosed lung disease. Experiments have uncovered the genes responsible for lung cancer mutagenicity and pathogenesis, albeit most genes have only a tenuous link to the disease. To determine whether a gene is linked to lung cancer, one must run several trials, which would necessitate a considerable financial commitment.

On the other hand, machine learning (ML) algorithms can prioritize disease-triggering genes where their significant

[1]*Research scholar, Dept. of CSE, JNTUH, Assistant Professor, CSE Dept., Chaitanya Bharathi of Technology Hyderabad, Telangana, India
Email: kmarysudha_cse@cbit.ac.in
[2]Professor, Dept. of CSE, JNTUH, Telangana, India

studies offer the ability to uncover the association amid cancer identification genes. These findings can potentially be used in the early detection of cancer. In particular, successful ML approaches are described works. These algorithms include artificial neural network-based computer-aided diagnosis [4, 5], ensemble approaches [6, 7], and hybrid methods.

Because of its extensive prevalence, it has been examined for cancer biomarkers that can predict a disease's prognosis. To be more exact, lung carcinoma is one of the most common types of cancer, with tobacco use accounting for over 85 percent of cases. Regrettably, the vast majority of instances are fatal. This is due, in part, to a delayed diagnosis, which necessitates specialized medical procedures such as bronchoscopy. As a result, lung cancer biomarkers are seen as critical in the disease's early identification; as a response, numerous initiatives have investigated non-invasive approaches for testing these biomarkers [8]. Ensemble learning could be effective in our investigation because it can increase a model's robustness and accuracy by merging multiple imperfect classifiers. Ensemble learning, which includes bagging and boosting, can be viewed as a general bagging technique for enhancing cancer classification. Random forest is another popular bagging technique. Gene microarray (GMA) recently appeared as a promising cancer detection and classification technique. Statistical analysis and machine learning methods were used to uncover accurate gene characteristics that can be used as inputs for cancer classification models. The lung cancer data's limited sample size makes the interpretation and training of microarray data problematic. The presence of noise in the samples can have a negative impact on the training models' performance. Furthermore, the random forest was utilized to search the classifiers at random, and in the training stage, a better judgment was generated.

One technique like self-paced learning, while another develops a novel formulation to uncover samples [10]. As the SPL regularizer's penalty steadily increases during optimization, more samples during the training phase are selected modes. Adoption has been quick, especially in multi-task learning [11], image categorization [12], and molecular descriptor selection [13, 14]. So, in the current article, we extract high-quality samples using this strategy. The following contributions were made by this paper, which is given below.

- We initially proposed using the IOLC to train a cancer detection and classification model using DNA microarray technology. The samples' degrees of confidence ranged from high to low.
- We built a machine learning prediction model using the Ensemble Methods Random Forest, and AdaBoost was the second stage of this project.
- The suggested approach's accuracy, F1-score, and recall are much greater than previously utilized classifiers.
- Furthermore, the proposed technique chooses a small number of genes (less than 1%) that are extremely important in predicting the disease's early prognosis.

The following are the organization of the paper: We describe the related work of lung cancer identification models based on genes data in section 2. Section 3 proposed work. The section 4 covers the results and discussion. Finally, in Section 5, concludes the paper.

2. Background and Related work

In [15], the authors investigated the link between socioeconomic status and the prevalence of lung cancer in several locations of the world, using educational degrees as a proxy for socioeconomic class. This study's data came from 18 prospective cohorts dispersed over 15 countries, including the United States, Europe, Asia, and Australia. They examined the link between educational level and the incidence of lung cancer in people who had never smoked and those who now or had previously smoked using Cox proportional hazards models. The International Standard Classification of Education was used to harmonise education data, which was then modeled as an ordinal variable divided into four categories. The models were modified to consider age, gender, whether the individuals smoked currently or previously, as well as smoking duration, quantity of cigarettes per day, and time since leaving.

In [16], the authors examined various methods, including machine learning, Ensemble learning, deep learning approaches, and numerous ways based on image processing techniques and text information that contribute significantly to determining cancer malignancy degree. Lung cancer has been listed as one of the most deadly diseases humans have faced since the species' inception. It is even one of the malignancies that causes the most continuous fatalities and contributes significantly to the overall mortality rate. The number of persons diagnosed with lung cancer continues to rise. In India, roughly 70.0 thousand cases are reported each year. It is impossible to identify early because the disease is often asymptomatic in its early stages. As a result, discovering cancer earlier is beneficial to save lives. Learning about a patient's illness as soon as possible will improve their chances of rehabilitation and recovery. Cancer diagnosis frequently relies substantially on technical breakthroughs. They intended to use this to combine or bring together Ensemble learning techniques such as stacking, blinding, Max voting, boosting, and XGBoost to provide a comprehensive methodology for evaluating and investigating the outcomes. Compared to other strategies, the Blinding ensemble learning methodology emerges as

the most successful way based on performance criteria such as accuracy, F1 score, precision, and recall.

In recent years, computer technology has been used to resolve various diagnostic concerns. To accurately predict the lung cancer severity level, these newly designed systems include several deep learning and machine learning tactics and specific image processing methods. As a result, this methodology aims to provide a new and unique approach to lung cancer diagnostics. The initial stage in data collection is to download two benchmark datasets. These datasets contain attribute information extracted from the medical records of a range of individuals. To extract features, the techniques of "Principal Component Analysis (PCA)" and "t-Distributed Stochastic Neighbour embedding (t-SNE)" were used. Furthermore, the deep characteristics are derived from what is known as "the pooling layer of Convolutional Neural Network (CNN)." The Best Fitness-based Squirrel Search Algorithm (BF-SSA), also known as optimal feature selection, is used to pick the features themselves in addition to the important features. This is referred to as feature selection. This hybrid optimization strategy is advantageous in many industries because it more efficiently explores the search space and performs better using feature selection.

In [18], the authors evaluated relevant surveys, underlining the need for a further study focusing on Ensemble Classifiers (ECs) utilized for cancer diagnosis and prognosis. By integrating several types of input data, learning methods, or characteristics, ensemble approaches strive to increase performance. They are being used in cancer detection and prognosis, among other things. Nonetheless, the scientific community needs to catch up in this technological sphere. A systematic evaluation of ensemble methodologies used in cancer prognosis and diagnosis, coupled with a taxonomy of such methods, can help the scientific community keep up with technology and, if comprehensive enough, even lead the trend. The following stage will thoroughly review the possible methodologies, both classical and deep learning-based. In addition to identifying the well-studied cancer kinds, the best ensemble methods used for the linked purposes, the most common input data types, the most common decision-making strategies, and the most common assessment methodology, the review creates a taxonomy. All of this happens as a result of the evaluation. Furthermore, they recommend future directions for scholars who want to continue existing research trends or concentrate on areas of the subject that have yet to receive less attention.

In [19], the authors developed Neural Ensemble-based detection for automatic disease detection (NED). An artificial neural network ensemble detects lung cancer cells in patient needle samples. They used Neural Ensemble-based Detection (NED) to achieve autonomous anomalous detection. The ensemble was divided into two levels. Because each network has two outputs, normal and malignant, the first-level ensemble can accurately detect normal cells. There are five types of lung cancer cells produced by each network: adenocarcinoma, squamous cell carcinoma, small cell carcinoma, prominent cell carcinoma, and normal. The second-level ensemble deals with cancer cells discovered by the first. To include network predictions, plurality voting is employed. NED has a high detection rate and a low false negative rate, which occurs when cancer cells are misidentified as normal. This reduces the number of undiagnosed cancer patients, hence saving lives.

3. Proposed Model

This section comprehensively explains the methods and materials used in this work, beginning with the architecture of the proposed Identification of Lung Cancer (IOLC) model. Figure 1 depicts the various processes involved in implementing and utilizing the model in the form of major blocks. The following subsections provide a complete overview of the architecture's fundamental building blocks, which include data set collection (genomic data from mutant and normal genes), data preparation (label encoding), feature extraction, and classification.

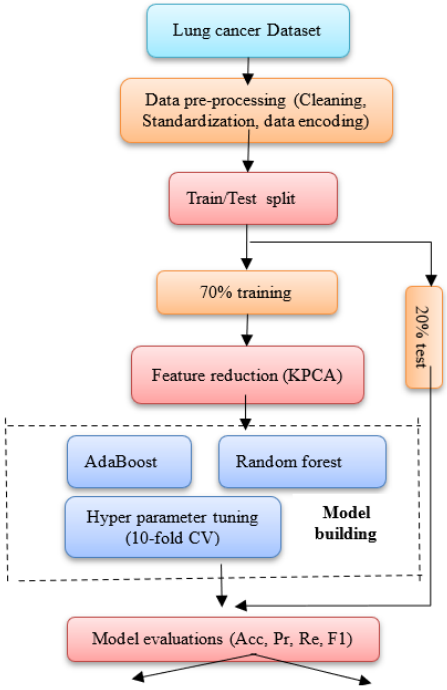

Fig 1: Working procedure of Lung cancer identification model

Description of the dataset:

The dataset for discussion here is comparable to one used in a prior study at the Boston University Medical Centre [20, 21]. In these investigations, a microarray was used to examine the level of gene expression found in epithelial cells originating in smokers' respiratory tracts. This dataset was used to extract the levels of expression of 22284 genes collected from 192 smoking subjects. Tissue samples were gathered and isolated from tissue samples. Patients were separated into 3 groups: those who had already been diagnosed with lung cancer (97), those who had not yet been diagnosed with lung cancer (90), and those who were thought to be at a high risk of developing cancer (5). This dataset was chosen specifically for its ability to perform in-depth research into the underlying genetic abnormality in smokers who acquire lung cancer. Although it was created on an older platform (the Affymetrix U133A array), it was chosen carefully. The dataset, known as GDS2771 and associated with the reference number GSE4115, is freely available for download from the NCBI's Gene Expression Omnibus (GEO) database [22]. The Affymetrix Human Genome U133A Array (HG-U133A) was used as the screening platform to gather this data. This array provided the information on the probesets.

The working procedure of the Lung cancer identification model is was repeated for each dataset containing gene expression data. After extracting the probe annotations, it gets represented to a respective gene, and those were not match any of the genes in the dataset were excluded from further consideration. If the gene has more than one probe, the gene's expression was calculated by taking the mean of all probe expressions. If the gene does not have many probes, the value was calculated by taking the mean of only one probe's expression. The Lung Cancer gene expression dataset is listed in Table 1.

Table 1 Lung Cancer gene expression dataset

GEO accession number	Disease	Number of samples (Disease/Control)	No. of Genes	Micro array Platform	Platform
GSE 4115	Lung Cancer	187 (97/90)	22,215	Affymetrix Human Genome U133A Array	GPL96 (HG-U133A)

Data Preprocessing:

Before using the data for training, 163 samples with complete clinical features were removed from the sample, including 85 smokers who did not have lung cancer and 78 smokers with lung cancer. Clinical data were gathered for future research, including patients' ages, genders, smoking histories, smoking indices, tumor diameters, and the presence or absence of lymphadenopathy. The Affymetrix Human Genome U133A Array platform was

used for the expression profile analysis. Each probe ID was matched to the symbol of a matching gene based on the information saved on the platform (GPL96-15653.txt). Because multiple probes could be associated with the equivalent gene sample, the domino effect was combined and averaged. The *Z*-score was used to normalize all gene expression values, which was calculated using the standard deviation (SD and mean of every gene symbol and then correcting the *X* value. This was done to mitigate the effect of variances in the quantities of intrinsic expression found in different genes. The new equation produces the value *X*, which is the mean/standard deviation ratio. The expression levels of all genes in each dataset were normalized using the methods described below.

$$z_{ij} = \frac{g_{ij} - mean(g_i)}{std(g_i)} \quad (1)$$

where g_{ij} represents the expression value of gene *i* in sample *j*, and $mean(g_i)$ and $std(g_i)$ respectively represents mean and standard deviation of the expression vector for gene *i* across all samples.

Feature reduction:

Kernel Principal Component analysis: PCA is widely used when one wants to minimize the dimensionality of a dataset while retaining as much information as possible. The entire dataset (with m dimensions) is mapped onto a new subspace (with *j* dimensions). *j* is smaller than *x*. This projection approach is useful for reducing both computing costs and the errors that can occur while estimating parameters ("the curse of dimensionality"). Suppose the data cannot be separated linearly. In that case, a nonlinear technique must be used to reduce the dimensionality of the dataset with KPCA, or Kernel Principal Component Analysis, which is a method for analyzing linearly inseparable data. PCA improves output by generating a feature subspace which reduces variance and normalizes the dataset to a unit scale (with mean = 0 and variance = 1). This is required for a wide range of ML methods to perform correctly. The main task is to transform the m-dimensional dataset (represented by A) into a new sample set (represented by B) with a lower dimension (k less than m). In this situation, B will stand in for the most important part of A, designated by A.

$$B = PC(A) \quad (2)$$

With *X* comprises of n vectors (x_1, x_2 ,…., x_n), each x_i signifies dataset instance, so:

$$\sum_{j=1}^{x} \delta(a_l) = 0 \quad (3)$$

To compute covariance matrix (CM) we take

$$C = \frac{1}{x}\sum_{j=1}^{x} \delta(a_l)\delta(a_l)^T \quad (4)$$

The eigenvectors are:

$$Cu_k = \alpha_k u_k, k = 1, \dots, M \quad (5)$$

After constructing the Eigen space from the covariance matrix and removing the less relevant regions, the original data will have a better chance of being accurate. To avoid access to the feature area and instead concentrate on kernels, which are as follows:

$$J(a_{l,}, a_p) = \delta(a_l)^T \delta(a_l) \quad (6)$$

Ensemble Learning:

Ensemble learning enhances generalizability and resilience over a single model by aggregating multiple models, like the J-node regression tree. This is achieved by combining the predictions of several simple models or base learners. Bagging and boosting are two typical tactics in group music.

Random Forest classifier

This was built using a modified version that generates a huge sample of uncorrelated trees and takes the average of those trees. This enabled the formation of a Random Forest [23]. This was formed as a result of this change. During the tree generation process, choose *p* input variables in the random subset from the entire set of *v* input variables to be examined for their appropriateness as a candidate for a split. This is referred to as the tree generation process. Because time series forecasting is being performed, each new training set is generated without replacing previous data. As a result, a single regression tree named T_b is produced by iteratively repeating the steps described below for each node in the tree until the smallest possible node size is obtained. This technique is repeated several times until the tree achieves the appropriate level of precision.

This procedure is performed on every *B* tree. The function can be expressed as an average of all *B* trees.

$$\hat{F}(x_i) = \frac{1}{B}\sum_{b=1}^{B} T(x_i; \Theta_b) \quad (7)$$

where Θ_b denotes the tree for node split. In [23], the findings show that using randomness and diversity in tree construction results in a lower generalization error and an overall better model with less variance.

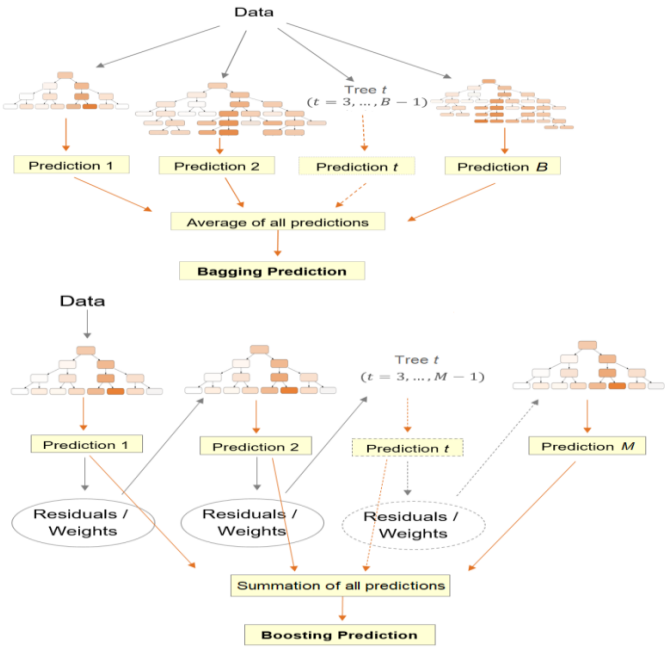

Fig 2: Tree structures of Bagging and Boosting

The boosting method entails sequentially developing the trees by combining the knowledge gained from previously formed trees with modified training data (Figure 2). This can be stated as follows:

$$F_m(x_i) = F_{m-1}(x_i) + \sum_{y=1}^{I_m} \gamma_{jm}\left(x_i \epsilon R_{jm}\right) \qquad (8)$$

So, the updated model will put in the form as

$$\hat{F}(x_i) = F_M(x_i) = \sum_{m=1}^{M} T(x_i; \Theta_m)$$

$$= F_{m-1}(x_i) + \sum_{y=1}^{I_m} \gamma_{jm}\left(x_i \epsilon R_{jm}\right) \qquad (9)$$

$\Theta_m = \{R_{jm}, \gamma_{jm}\}_1^{j_m}, F_{m-1}(x_i)$ signifies the prior model, while $\widehat{F}(x_i) = F_M(x_i))$ signifies the current tree.

AdaBoost classifier

The classifier in [24] updates by attaching weights $\{w_1, w_2,..., w_N\}$ for every training instance (x_i, y_i) (Figure 2). As a result, a total of N weights will be used. At the start of the procedure, each weight is assigned the value $w_i = 1/N$, indicating that the data is being trained in the usual manner. This is known as the learning period. The weighted observations training approach will be continued until all stages have been completed at each subsequent step ($m = 2, 3$, etc.). This will be repeated until all phases have been accomplished. The weights of the various components are changed at each of these steps. To be more exact, the weights for the observations that were mistakenly predicted in the previous step are given higher priority *in* step *m*, whereas the weights for the observations that were correctly predicted are given lower priority. This arises because the weights for the erroneously predicted observations are more likely to contain errors. As a result, as the iterations advance, the findings that are hardest to predict gain increasing emphasis. Finally, as shown in Equation (9), the final prediction is formed by combining the weighted predictions from each tree. This yields the final projection. Gradient boosting, a technique that may be applied to any arbitrary differentiable objective function, can be used to extend boost. Initial training data are used to instruct a tree in the first step of the procedure. As a result, the gradient may be determined to be [25] for all *i* values ranging from 1 to *N* inclusive.

$$-g_{im} = -\left[\frac{\partial L}{\partial F(x_i)}\right] F = F_{m-1} \qquad (10)$$

For the squared error loss, the negative gradient signifies the residual$-g_{im} = y_i - F_{m-1}(x_i)$.

Model Selection and Validation

An optimization strategy is employed during the learning phase to forecast the values of various parameters based on the collected data. These parameters contain the splitting variable and the splitting point value. On the other hand, each learning algorithm includes a set of hyperparameters that are not learned and must instead be tailored to the unique modeling task at hand. The hyperparameters govern both the model's architecture and its level of complexity. The data and the problem at hand decide their ideal values. However, the training data residual sum of squares cannot be calculated because

doing so weakens a model's capability to generalize to new data. This is because doing so would reduce the size of the training set. As a result, three distinct data sets were used: the training set, the validation set, and the test set. The training set was used to train the model, while the validation set was used to evaluate and fine-tune the model's parameters and hyperparameters. Ultimately, the test set was only used to estimate the generalization error. As a result of this, we were able to select machine learning models with hyperparameter values.

Performance evaluation

Several validation metrics were discovered during our inquiry. Accuracy (Acc), F1-score (F1), Precision (Pr), and Recall or sensitivity (Re) were among them. The formula for each validation parameter is presented in equations (11) through (14). The abbreviations TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative outcomes, respectively.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (11)$$

$$Re = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

$$Pr = \frac{TP}{TP + FP} \times 100\% \quad (13)$$

$$F1 = \frac{2 \times (Pr \times Re)}{\mathrm{Pr} + Re} \times 100\% \quad (14)$$

4. Results and Discussion

In this section, the first step is to divide the data into two categories: training (70%) and testing (30%). Several machine learning algorithms, such as feature scaling, KPCA, ROS, and hyperparameter tuning, are utilized to determine the optimum model that delivers the highest level of accuracy. Good classification was picked by combining all the ML algorithms used in this study. This experiment necessitates the use of specified resources. The suggested system's environment configuration includes an Intel® CoreTM i-3-1005G1 CPU running at 1.20GHz, 8GB of RAM, the Anaconda tool, and the Python programming language, which was utilized to construct the model for this study. As shown in Figure 3, Cancer and Non-cancer data of the lung cancer dataset used in this study.

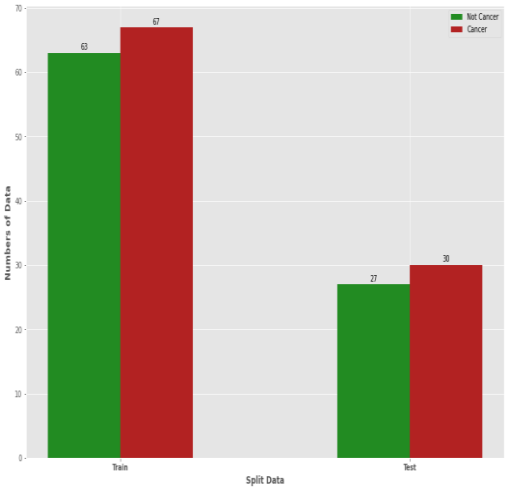

Fig 3: Cancer and Non-cancer data

In this paper, we use stable LASSO operation to provide more proof of the efficacy of our technique in computer-assisted diagnostics. Table 1 shows the ten genes that received the highest rankings after being submitted to stable LASSO analysis across all datasets. Most stability ratings are close to one, indicating that the genes chosen are tough. Furthermore, obtain the important p-values that ate statistically best for this study. Much research is being done on the functional analysis of gene expression. USP6NL, is one such protein that acts as a GTPase activator for RAB5A.

Table 1: Best 10 genes from GSE 4115 dataset

Gene Name	Gene Symbol	Stable Score	p-Value
USP6 N-terminal like	(USP6NL)	1	<0.01
acyl-CoA oxidase 2	(ACOX2)	0.98	<0.01
agouti related neuropeptide	(AGRP)	0.53	<0.01
HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase	(HUWE1)	0.99	<0.01
calcium/calmodulin dependent protein kinase II beta	(CAMK2B)	1	<0.01
tripartite motif containing 5	(TRIM5)	1	<0.01
Janus kinase 3	(JAK3)	1	<0.01
sperm antigen with calponin homology and coiled-coil domains 1 like	(SPECC1L)	0.96	<0.01
sperm antigen with calponin homology and coiled-coil domains 1 like	(EML3)	1	<0.01
glycosylphosphatidylinositol anchor attachment protein 1 homolog (yeast) pseudogene	(LOC100288 570)	1	<0.01

Figure 4 shows the heat map correlation discovered between the genes in the meantime. Red is used when there is a positive correlation, and violet is used when there is a negative correlation. The stronger the correlation, the greater the degree of resemblance. As seen in Figure 4, most identified genes have a positive relationship.

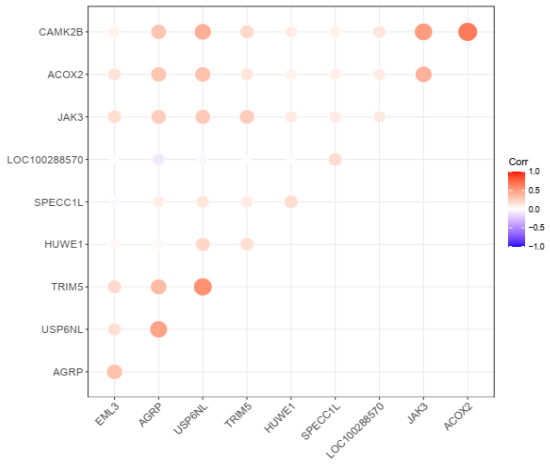

Fig 4: Graph showing heat map

Several well-known matrices, such as accuracy, recall (also known as sensitivity), precision, and F1-score, are used to assess the classification algorithms' performance. Table 2 shows the performance of RF and AdaBoost in terms of various evaluation metrics.

Table 2: Performance analysis of RF and AdaBoost models

Model	Accuracy	Precision	Recall	F1-score
AdaBoost	0.789	0.810	0.781	0.766
Random forest	0.810	0.812	0.789	0.777

In the case of the GSE4115 example presented in Figure 5, the best model obtained is a Random forest, with an accuracy of 0.810 and an F-1 score of 0.777, respectively. This demonstrates that ML is a suitable strategy for working with the dataset. Meanwhile, we noticed that the AdaBoost model that performed the lowest had an accuracy of 0.789 and an F-1 score of 0.766. Meanwhile, the best recall score in MI for the Random forest classification approach is 0.789, implying that all models can reliably predict genuine positives while avoiding false pessimistic predictions.

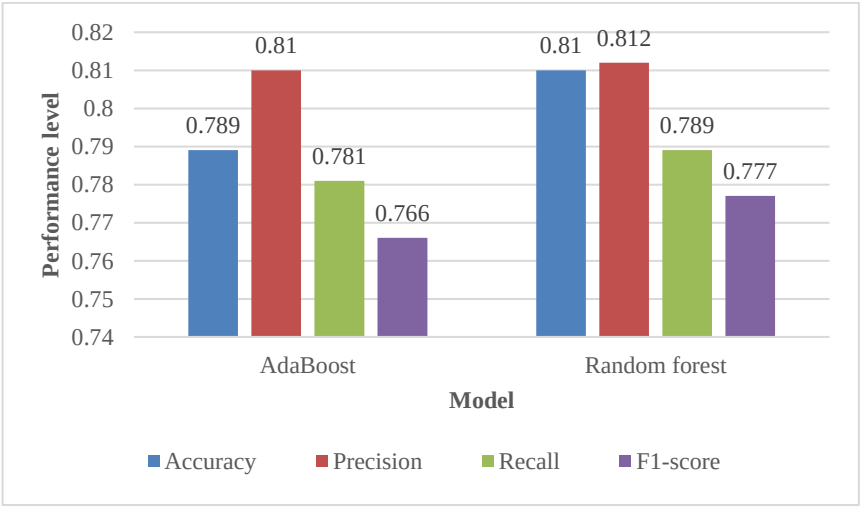

Fig 5: Performance comparison of AdaBoost and Random forest on GSE 4115 dataset

Results on Feature Selection operation:

We used 5-fold cross-validation to investigate the effect of feature number on overall model performance. This enabled us to reduce the overall amount of features. The ideal number of attributes was determined by examining a range of values from 2 to 10. Figure 6 shows that the variation of the scores produced using AdaBoost and Random Forest is substantially more considerable than that achieved using any other approaches for GSE4115. This indicates that the AdaBoost approach's performance highly depends on the number of features utilized. In an unexpected turn of events, the Random forest approach revealed a significant decline in a feature's overall score.

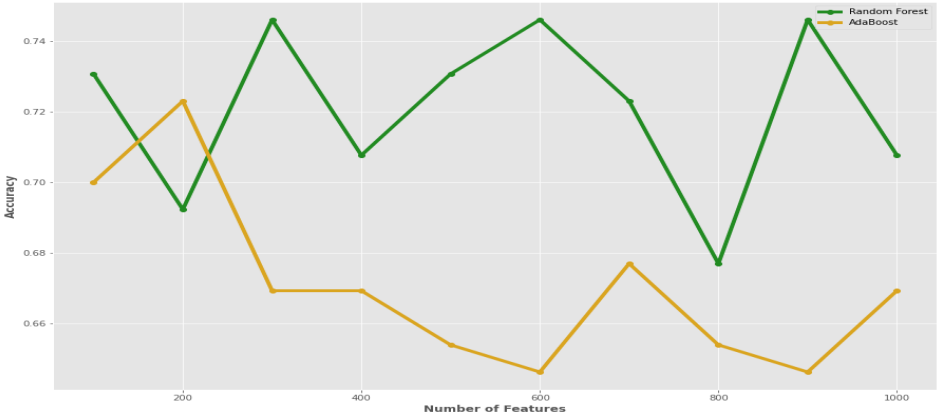

Fig 6: Performance comparison of AdaBoost and Random forest after applying feature selection on GSE 4115 dataset

5. Conclusion

In this article, we propose a novel method for detecting lung cancer by building an ensemble classifier and comparing its findings to the RF classifier. In the Ensemble-Classifier, we used two machine learning models: AdaBoost and Random Forest. We begin by extracting features from the dataset, then divide it into 70:30 proportions for training and testing. We classified cancer as Tumor or Normal using the confusion matrix and then provided a classification report that contained accuracy, precision, recall, and F1-score. The feature selection procedure involved calculating the correlation between the feature and the target using statistical parameters, also known as KPCA. Deep learning techniques, such as CNN, may one day aid in diagnosing lung cancer. Images from many scanning modalities, including MRI, CT, PET, and X-ray, can be considered. This can increase precision, allowing the medical sector to provide rapid prevention at a minimal cost. In addition to categorized information, continuous information can be used.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2018. CA: a cancer journal for clinicians, 68(1):7–30, 2018.

[2] Lindsey A. Torre, Rebecca L. Siegel, and Ahmedin Jemal. Lung Cancer Statistics. Springer International Publishing, 2016.

[3] Howard Lee and Yi Ping Phoebe Chen. Image based computer aided diagnosis system for cancer detection. Expert Systems with Applications 42(12):5356–5365, 2015.

[4] Azian Azamimi Abdullah and Syamimi Mardiah Shaharum. Lung cancer cell classification method using artificial neural network. Information engineering letters, 2(1), 2012.

[5] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. M. Ngai, and J. Shao. Classification of lung cancer using ensemble-based feature selection and machine learning methods. Molecular Biosystems, 11(3):791–800, 2015.

[6] Maciej Zięba, Jakub M Tomczak, Marek Lubicz, and Jerzy Świątek. Boosted svm for extracting rules from imbalanced data in application to predictionof the post-operative life expectancy in the lung cancer patients. Applied soft computing, 14:99–108, 2014.

[7] Golrokh Mirzaei, Anahita Adeli, and Hojjat Adeli. Imaging and machine learning techniques for diagnosis of alzheimerâĂŹs disease. Reviews in the Neurosciences, 27(8):857–870, 2016.

[8] Aboul Ella Hassanien, Hossam M Moftah, Ahmad Taher Azar, and Mahmoud Shoman. Mri breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier. Applied Soft Computing Journal, 14(1):62–71, 2014.

[9] Qingyong Wang, Liang-Yong Xia, Hua Chai, and Yun Zhou. Semi-supervised learning with ensemble self-training for cancer classification. In 2018IEEE

Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud& Big Data Computing, Internet of People and Smart City Innovation (Smart World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pages 796–803. IEEE, 2018.

[10] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In Advances in Neural Information Processing Systems, pages 1189–1197, 2010.

[11] Changsheng Li, Junchi Yan, Fan Wei, Weishan Dong, Qingshan Liu, and Hongyuan Zha. Self-paced multi-task learning. In AAAI, pages 2175–2181, 2017.

[12] Ye Tang, Yu Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. In ACM International Conference on Multimedia, pages 833–836, 2012.

[13] Liang-Yong Xia, Qing-Yong Wang, Zehong Cao, and Yong Liang. Descriptor selection improvements for quantitative structure-activity relationships. International Journal of Neural Systems, pages 1–16, 2019.

[14] Abiezer, Otniel & Nhita, Fhira & Kurniawan, Isman. (2022). Identification of Lung Cancer in Smoker Person Using Ensemble Methods Based on Gene Expression Data. 89-93. 10.1109/IC2IE56416.2022.9970035.

[15] Onwuka, Justina & Zahed, Hana & Feng, Xiaoshuang & Alcala, Karine & Johansson, Mattias & Robbins, Hilary & Consortium, Lung. (2023). Abstract 1950: Socioeconomic status and lung cancer incidence: An analysis of data from 15 countries in the Lung Cancer Cohort Consortium. Cancer Research. 83. 1950-1950. 10.1158/1538-7445.AM2023-1950.

[16] Fatima, Fayeza Sifat & Jaiswal, Arunima & Sachdeva, Nitin. (2023). Lung Cancer Detection Using Ensemble Learning. 10.1007/978-3-031-23724-9_15.

[17] Zolfaghari, Behrouz & Mirsadeghi, Leila & Bibak, Khodakhast & Kavousi, Kaveh. (2023). Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning. ACM Computing Surveys. 55. 10.1145/ 3580218.

[18] Pradhan, Kanchan & Chawla, Priyanka & Tiwari, Rajeev. (2022). HRDEL: High Ranking Deep Ensemble Learning-based Lung Cancer Diagnosis Model. Expert Systems with Applications. 213. 118956. 10.1016/ j.eswa.2022.118956.

[19] Zhou, Zhi & Yang, Yu & Chen, Shi. (2002). Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles. Artificial intelligence in medicine. 24. 25-36. 10.1016/S0933-3657(01)00094-X.

[20] Spira, A.; Beane, J.E.; Shah, V.; Steiling, K.; Liu, G.; Schembri, F.; Gilman, S.; Dumas, Y.M.; Calner, P.; Sebastiani, P.; et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. Nat. Med. 2007, 13, 361.

[21] Gustafson, A.M.; Soldi, R.; Anderlind, C.; Scholand, M.B.; Qian, J.; Zhang, X.; Cooper, K.; Walker, D.; Mc Williams, A.; Liu, G.; et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. Sci. Transl. Med. 2010, 2, 26ra25–26ra25.

[22] Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002, 30, 207–210.

[23] Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32.

[24] Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci. 1997, 55, 119–139.

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

[26] Leila Abadi, Amira Khalid, Predictive Maintenance in Renewable Energy Systems using Machine Learning , Machine Learning Applications Conference Proceedings, Vol 3 2023.

[27] Martin, S., Wood, T., Hernandez, M., González, F., & Rodríguez, D. Machine Learning for Personalized Advertising and Recommendation. Kuwait Journal of Machine Learning, 1(4). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/156

[28] Raghavendra, S., Dhabliya, D., Mondal, D., Omarov, B., Sankaran, K. S., Dhablia, A., . . . Shabaz, M. (2022). Development of intrusion detection system using machine learning for the analytics of internet of things enabled enterprises. IET Communications, doi:10.1049/cmu2.12530