International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

Identification of Most Important miRNA Biomarkers for Lung Cancer and Survival Rate Analysis

Hasifa A S[1], Raunak R Kolle[2], Rekha B S[3], Praveen Kumar Gupta[4]

[1]*Student, Dept. of Information Science and Engineering, R.V. College of Engineering, Karnataka, India*
[2]*Student, Dept. of Information Science and Engineering, R.V. College of Engineering, Karnataka, India*
[3]*Assistant Professor, Dept. of Information Science and Engineering, R.V. College of Engineering, Karnataka, India*
[4]*Assistant Professor, Dept. of Biotechnology, R.V. College of Engineering, Karnataka, India*

Abstract - *Lung Cancer is the most prevalent type of cancer worldwide. The present means for identification of lung cancer include imaging tests, lung X-ray, sputum cytology and tissue samples (biopsy), these methods are inapplicable for early detection screenings. A microRNA is a single-stranded non-coding RNA molecule which is small, highly conserved and involved in the regulation of gene expression. Inorder to improve the existing low survival rate of infectants with lung cancer, better identification tools and outcomes have to be introduced. Hence there is a need to develop more accurate identification methods with precise metric by using miRNA. In this work, Different feature selection algorithms were used, in which the combination of 5 tree-based classifiers feature importances were evaluated in comparison to a baseline model (Dummy Classifier). The survival rate for the patients with lung cancer were analyzed for different criteria.*

Key Words*: *Lung Cancer, miRNA, tree-based classifiers, survival rate, liquid biopsy, gene expression

1. INTRODUCTION

Lung cancer is one of the most prevalent variants of cancer, present worldwide. Millions of former or current heavy smokers have an elevated risk of developing lung cancer, in Spite of the decreased smoking rates in industrially developed countries. This is due to the symptoms of lung cancer not being evident during the early testing phases. More than two-thirds of all the cases of lung cancer are detected in late stages where the patient's health has deteriorated to an almost inoperable condition. The current standard diagnostics for lung cancer are X-ray imaging, sputum cytology and tissue samples (biopsy), which are inapplicable for early detection screenings. Liquid biopsy–based strategies are being explored to detect lung cancers, as these results showed potential for early detection. Besides proteins, DNA and common bioinformatics, miRNAs have shown greater potential for the detection of a wide variety of human genetic disorders.

In spite of the research being performed on tumor biomarkers, only a few tissue-based biomarkers and virtually no new blood-borne biomarkers are used in clinical practices. The clinical application of biomarkers is greatly hampered by the relatively small number of cases that have been analyzed in most preclinical studies. The limitation is soared up by the fact that most studies consider only a small number of cases, limiting the validation of the miRNA signatures that were analyzed. In addition to this, different methodologies for the collection, storage and analysis have had a particularly strong influence on the results of the small studies. This downside is also applicable to small studies that use the reverse transcriptase polymerase chain reaction. Alternatively, an extended number of cases are studied using the limited few miRNA signatures mostly via reverse transcriptase polymerase chain reaction.

MicroRNAs (miRNA) are a class of small non-protein-coding RNAs of approx. 22 nucleotides that function as negative gene regulators on post-transcriptional level. Each miRNA may have multiple regulated target genes. These genes play key roles in several biological functions ranging from proliferation, cell cycle progression, differentiation to apoptosis and metastasis formation in vivo. Further, miRNAs can function as tumour suppressors and oncogenes suppressing the expression of important cancer-related genes. Conjointly miRNA expression levels are often altered in various varieties of cancer. Expression profiling of miRNAs has been shown to be a more accurate method of classifying cancer subtypes than using the expression profiles of protein-coding genes.

2. LITERATURE SURVEY AND RELATED WORK

The expression values of the microarray are extracted in [5] with respect to the selected features and its value scores are calculated by combining these extracted expression values according to the ET of the chromosome. The classification scores are assigned and the individual chromosomes are ranked according to their fitness numbers and genetic operations are performed on the same. This process is reiterated until a chromosome with a significantly high fitness number is obtained. This classifier works only on a small set of attributes. RNA-sequencing and 450 K methylation for the microarray profiling was conducted by [8] to both, the non-small cell lung cancers and non-malignant lung tissue and the data was analyzed for 14 target genes, distinctive expression and methylation were analyzed. The results obtained in this suggested a lower PD-L1 & 2 and VEGFR expression in

the Non Small Cell Lung Cancer against the non-malignant lung tissue. Specific patient characteristics had no effect on the overall expression differences because they were consistent with the overall findings.

The dimension reduction was done by [10] using PCA as part of the preprocessing. The covariance-matrix was calculated, followed by calculating the eigenvalues which were sorted descendingly. Min-Max scaling was performed to normalize the data. A hybrid model using back-propagation along with ANN-GA was developed which was able to offset the insufficiency in the model and lift up the predictive accuracy and the AUC value using only a small number of feature input. Descriptive statistics of patients were plotted in [11] by nine categories of the clinical data that were selected for their higher relevance to the Non-small cell lung cancer. The subset of the selected samples were chosen according to their respective vital status. The downstream analyses were performed based on the "Dead" samples. Cox Proportional Hazards Regression (CPHR) was implemented for the determination of the association existing between the survival times and the individual clinical categories. Subsets of raw count data were formed according to the previously described clinical data using the parameters of "pathologic T" and "days to death."

In the work carried out in [15] ,formalin fixed paraffin embedded biopsy samples with Non-small cell lung cancers were included if the tumour section haematoxylin and eosin examination showed at least 70 % of tumour cells. Unsupervised hierarchical clustering was used for the determination of the relationships existing among the various survival groups by the assessment based on their expression profile carried out independently. Analysis of variance (ANOVA) test was carried out on the filtered data for the identification of the gene sequences that had a contrasting expression between short and long survival groups. Predictive models were developed for the gene selection utilizing the 13-gene signature expression measurements in a number of individuals with a known class membership (training cohort). Learning of the classification methods required for the identification of the gene signatures used for the prediction of the survival of the cohort was done using the SVMs.

3. METHODOLOGY

MiRNA Biomarkers relevant top 20 features have to be selected from the original dataset. Further top classification models are selected and further optimized using Randomized Search, Grid Search CV, Ensemble stack and Ensemble Voting techniques. From the best 20 miRNA Biomarkers the optimum 14 have to be reduced. This evaluated model is then used for Survival Rate analysis with the help of different visualization techniques.

The identification of important miRNA Biomarkers for Lung Cancer involve the sequencing of miRNA expressions by pre-processing the data first by eliminating missing values. Followed by introducing Classification models to classify patients into the groups Lung Cancer (LCa), Non-Tumor Lung Disease (NTLD) and CONtrol (CON). The workflow can be seen in Fig-1.

In this work, the top 20 best features were scaled down from 1183 in the feature selection process. The classification model is of medium performance to predict LCa, NTLD and CON classes. In the model selection phase 10 classification algorithms were taken into consideration and the top 3 were selected from it based on accuracy, precision, roc_au and f1 score. Here, five tree based classifiers were used - Decision tree (DT), Random Forest (RT), Extra-Tree (ET), Adaptive Boost (AB) and Extreme Gradient Boosting (XGB) classifier; two Recursive Feature Elimination models were used as classifiers which are logistic regression classifier (LR) and Stochastic Gradient Descent (SGD) classifier; along with these the k-nearest neighbors (KNN), Gaussian Naive Bias (GNB) and Support Vector Machine (SVM). The Dummy Classifier (DC) algorithm is used as the base-line algorithm in comparison to the other chosen models. The top three selected models were Random Forest, Extra-Tree and XGB classifier. Classification model Optimization is done using RandomizedSearch, GridSearchCV, Ensemble Stack and Ensemble Vote (Hard Voting) techniques. Then the top 14 optimum features were then extracted from the best 20 in the feature reduction process. The identified features can be experimentally tested in laboratories for evaluation of the model.

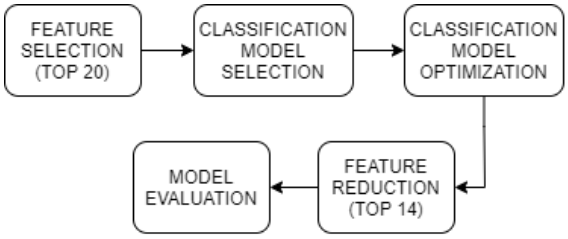

Fig -1: Workflow of Identification of important miRNA Biomarkers for Lung Cancer

The survival probability is visualised with each selected feature to identify the best attribute that helps analyse survival rate and visualise the best selected features with Grades of Lung Cancer stages from 1 to 4. The workflow is described in Fig-2. Missing, unknown and duplicate values were filtered in the preprocessing phase. In the data transformation process the column headers were converted to shorter and meaningful names, transforming attributes such as age, sex and grade of cancer. Columns that were not being used as features are dropped in the Data Reduction process.

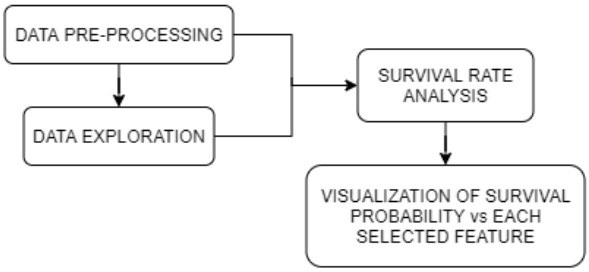

Fig -2: Workflow of Survival Rate Analysis

Support Vector Machine and Linear Regression were used to provide the regression model of the survival rate probabilities of the patients. Univariate linear regression and chi-square test were used for feature selection required in the analysis of the survival rates. Data Visualization techniques were adapted to plot the frequency of survival months as survival probability with respect to features.

4. RESULTS AND ANALYSIS

Table-1: Precision, Roc_AUC values and rank for each algorithm used

Classifier	*Precision value*	*Precision rank*	*Roc_AUC value*	*Roc_AUC rank*
DC	0.3118	11	0.4845	11
LR	0.8068	2	0.8997	4
SGD	0.7316	6	0.7951	8
KNN	0.6957	8	0.8180	7
GNB	0.5773	10	0.7409	10
SVM	0.6524	9	0.8361	6
DT	0.7253	7	0.7908	9
RF	0.7945	4	0.9232	3
ET	0.7954	3	0.9275	2
AB	0.7366	5	0.8492	5
XGB	0.8265	1	0.9385	1

The various classifiers used resulted in providing different values of precision and Roc_AUC which were then ranked according to their values as shown in table-1. The precision and the Roc_AUC value of the XGB classifier was the highest hence it was ranked 1, i.e, it performed the best among the algorithms used, obtaining a precision of 0.8265 and Roc_AUC value of 0.9385.

Although the unoptimized model XGBC_UnOpt showed better performance as compared to the optimized model, it could not prevent the overfitting, hence the optimized model was preferred.

The survival rate probability of the lung cancer diagnosed patients was found to be indeterminent of the sex. Hence the survival rates for males and the survival rates for the females were observed to be quite similar.

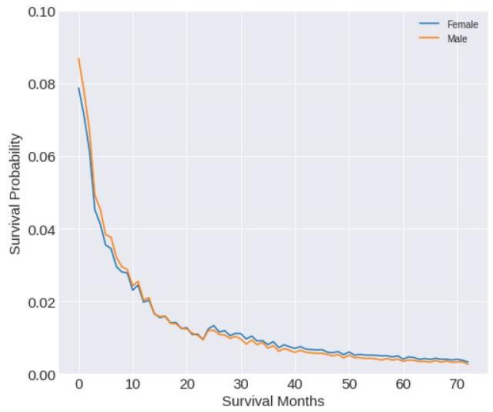

Fig -3: Survival Rate Analysis for the respective sex

The survival rates analyzed showed that the survival probability of the younger generation, i.e, ages in the range 19-38 had a greater survival rate as compared to the survival rate of the older population, i.e, the patients with the age of over 38 had a similar pattern of survival rate as was observed in the graph below.

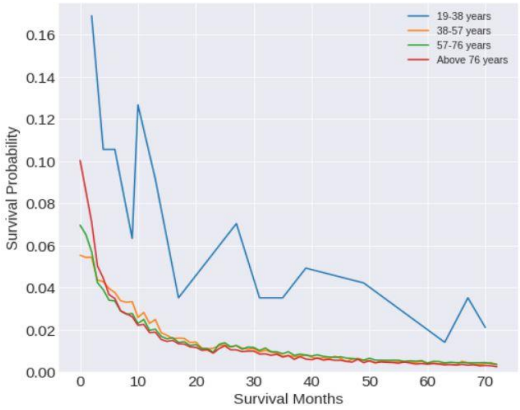

Fig -4: Survival Rate Analysis for the respective age groups.

5. CONCLUSIONS

Majority of the annotation data of samples considered were complete and the label groups Lung Cancer (LCa), Non-Tumor Lung-Disease (NTLD) and CONtrol (CON) were equally represented. Most miRNA-Expression values were not normally distributed except the values for miRNA 223-3p and miRNA 425-5p. The resultant of the

miRNA per sample was found not to be constant.. Various feature selection algorithms were used, among which the combination of 5 tree-based classifiers feature importances resulted in the best results for the model. Ten different classification algorithms were evaluated in comparison to a baseline model, the dummy classifier. Random Forest, Extra Trees and XGBoost Classification algorithms demonstrated the best metric scores and performances for precision and Roc_AUC in combination with TOP20 feature selection. We were able to optimize these models using Randomized Search and GridSearchCV on the chosen metric precision and further increase the performance by ensembling all of them in a Vote Classifier resulting in the Top Model (EVCLF).

The Top Model (EVCLF) demonstrated with ROC class values of 0.97, 0.99, 1.00 for CONtrol group (CON), Lung Cancer (LCa) and Non-Tumor Lung Disease (NTLD) respectively and a micro average ROC (global ROC over all classes) of 0.97 a good performance in comparison to the baseline model. The best Model among the ones used was used to reduce the TOP20 miRNAs, optimizing the conditions in order to save time and costs. The TOP14 had a loss in precision score by 0.867% compared to the best result obtained.

As the first enhancement to this model, reduction of the classification to a binary model can be done for the prediction of Lung Cancer. This could however result in unbalancing of the data as the classes will not remain equally distributed (from 100:100:100 samples in multiclass to 100:200 in binary). Upsampling or downsampling may be a possible solution. Implementation of more regularization penalizing overfitting could be another option to improve the models. Further the selected miRNA can be evaluated in laboratory experiments to validate their role in lung cancer.

REFERENCES

[1] Neubert, P (2020) miRNA Biomarkers for Lung Cancer Diagnostics [Source Code].

https://github.com/Patrick-Neubert/miRNA_LCa_Diagnostics

[2] S. Perakis and M. R. Speicher, "Emerging concepts in liquid biopsies", BMC Medicine, vol. 15, pp. 75, Apr. 2017.

[3] Z. Isik and M. E. Ercan, "Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients", Computers in Biology and Medicine, vol. 89, pp. 397-404, Oct. 2017.

[4] P. A. Jaskowiak, I. G. Costa and R. J. G. B. Campello, "Clustering of RNA-Seq samples: Comparison study on cancer data", Methods, vol. 132, pp. 42-49, Jan. 2018.

[5] Hasseeb Azzawi, Jingyu Hou & Yong Xiang, “Lung cancer prediction from microarray data by gene expression programming” ,Institution of Engineering and Technology, 2016.

[6] Zhang Weisan et al., "Analysis for the mechanism between the small cell lung cancer and non-small cell lung cancer combining the miRNA and mRNA expression profiles", Thoracic Cancer, vol. 6, no. 1, pp. 70-79, 2015

[7] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", CA A Cancer J. Clinicians, vol. 68, no. 6, pp. 394-424, Nov. 2018

[8] Kobe Reynders, Els Wauters, Matthieu Moisse ,Herbert Decaluwé, Paul De Leyn, Stéphanie Peeters, Maarten Lambrecht, Kristiaan Nackaerts, Christophe Dooms,Wim Janssens, Johan Vansteenkiste, Diether Lambrechts, “RNA-sequencing in non-small cell lung cancer shows gene downregulation of therapeutic targets in tumor tissue compared to non-malignant lung tissue”, Radiation Oncology, 2018.

[9] B J Bipin Nair; K J Anju; A Jeevakumar,”Tobacco smoking induced lung cancer prediction by lc-micrornas secondary structure prediction and target comparison.” , 2nd International Conference for Convergence in Technology (I2CT), 2017.

[10] Kusumastuti Cahyaningrum; Adiwijaya; Widi Astuti, “Microarray Gene Expression Classification for Cancer Detection using Artificial Neural Networks and Genetic Algorithm Hybrid Intelligence”, International Conference on Data Science and Its Applications (ICoDSA), 2020.

[11] Kemal Sanli; Sinem Nalbantoglu; Serdar Evman; Volkan Baysungur; Abdullah Karadag, “Pathway-Centric Analysis of the TCGA - NSCLC Transcriptome Data Pertaining”, 1st International Conference on Cancer Care Informatics (CCI), 2018..

[12] M. Sherafatian and F. Arjmand, "Decision tree-based classifiers for lung cancer diagnosis and subtyping using TCGA miRNA expression data", Oncology Letters, 2019.

[13] B. Chen, T. Gao, W. Yuan, W. Zhao, T. Wang and J. Wu, "Prognostic Value of Survival of MiRNAs Signatures in

Non-small Cell Lung Cancer", Journal of Cancer, vol. 10, no. 23, pp. 5793-5804, 2019.

[14] W. Zhu et al., "Diagnostic Value of Serum miR-182 miR-183 miR-210 and miR-126 Levels in Patients with Early-Stage Non-Small Cell Lung Cancer", PLOS ONE, vol. 11, no. 4, pp. e0153046, 2016.

[15] Roslan Harun; Jalal Hadi; Nur Shukriyah Mhazir; Pang Jyh Chyang; Isa Rose; Roslina A Manap; Fauzi M Anshar, “Gene expression profiles predict survival of patients with advanced non-small cell lung cancers” , Fourth International Conference on Modeling, Simulation and Applied Optimization, 2011.