International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217357

Lung Cancer Prediction Using Ensemble Learning

Vaibhav Narawade, Akash Singh, Mohit Shrivastava, Abhishek Prasad
Department of Computer Engineering, Ramrao Adik Institute of Technology, Navi Mumbai, Maharashtra, India

ABSTRACT

Article Info
Volume 7, Issue 3
Page Number: 477-482

Publication Issue :
May-June-2021

Article History
Accepted : 25 May 2021
Published : 31 May 2021

Lung Cancer is the most commonly occurring type of cancer in the world. Despite all the research in the field of lung cancer it still maintains a extremely high mortality rate and a cure rate of of less than 15%. Majority of lung cancer patients are diagnosed at a very advanced stage which is why randomized clinical trials have come under intense scrutiny from the medical practitioners and have led to a new resurgence of interest in its screening methods and development of newer techniques to improve its efficiency. The existing screening and detection techniques have known to be slow, cost ineffective and have other discrepancies such as false positives. Keeping this in mind we propose to use ensemble learning methods to train our data-set to overcome the drawbacks and improve upon the individual algorithms.

Keywords : Machine Learning, Lung Cancer, Naive Bayes (NB), Decision Tree, Random Forest, KNN Classifier, Soft Voting, Ensemble

I. INTRODUCTION

Ensemble learning is a method of combining various models together to improve the predictive power of the model.

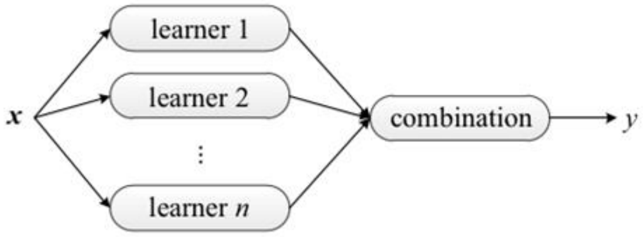


Fig. 1. Ensemble Learning

Ensemble model is known to improve the overall accuracy, efficiency and robustness on single learning methods and can also overcome the constraints of a single method. We use this to obtain a much higher accuracy for our results. Lung Cancer, having many possible expressions and symptoms, is very difficult to detect reliably and the cost of someone without lung cancer on cancer medication or chemotherapy is irreversible. Using multiple models in an ensemble style helps us mix-and-match models and emphasize their strengths while removing their weaknesses. This combined with the possibility of running on inexpensive consumer hardware could be very helpful.

With this project we aim to use Ensemble learning methods utilizing techniques like Random Forest, KNN, Decision Tree and Naive Bayes to predict Lung Cancer more accurately and efficiently as ensemble methods enable us to vastly improve the prediction results by training and testing the datasets against multiple algorithms, including the aforementioned ones, by improving upon their individual results and predictions. We also aim to get rid of the known issues that plague the currently available screening and testing techniques such as false positives which forces patients to undergo more testings inevitably leading to delayed diagnosis which may prove fatal, by making the results more accurate.

We were motivated to choose this topic because we have a joint interest in Machine Learning and new avenues like Ensemble Learning interest us. Lung Cancer presents an exciting opportunity to test it on something of importance that impacts many human lives, will probably grow in the coming years with increasing pollution, and requires correct assessments with inexpensive techniques to enable low-cost diagnosis and early prevention.

Limitations of Existing System

I. Non-ensemble techniques may rely on multiple runs of the same algorithm, which may miss crucial elements, or may take a much longer time than feasible.
II. Small Difference in the Training Data Set results inconsistent results for classification algorithms, i.e, an individual process is non-generalizable.
III. Particle Swarm Optimization is computationally expensive and unsuitable for large datasets.
IV. The existing systems of lung cancer detection are expensive and time consuming resulting in delay in detection and diagnosis which can prove fatal in some cases.
V. Computed tomography(CT) and Low dose computed tomography(LDCT) have been known to give false positives

II. METHODS AND MATERIAL

We aim to improve the efficiency of lung cancer detection by using ensemble learning by comparing and training larger data-sets against multiple algorithms and techniques namely KNN, Random Forest, Naive Bayes and Decision Tree so that a wider demography can be screened and we get a more definite result in a lesser time frame. We also aim to optimize the efficiency and computation for results and hence will use an ideal number of individual results as per our research

- After comparing various algorithms and based on accuracy, our model will give better, faster and accurate results which will help the doctors to diagnose the patients faster.
- Ensemble Learning will help in increasing the predictive power and precisions and will also help in decreasing the error rate which can boost the average precision.
- Data analysis helps public health department in understanding the trends observed in different diseases and helps in building awareness.
- Data analysis in healthcare has be come the driving force for improving the medical loss ratio and better managing clinical decision support systems.
- For our project, we aim to use a data-set with 567 Patients [9], with early symptoms like Air Pollution, Al- cohol Use, Dust Allergy, Occupational Hazards, Genetic Risk, etc. After cleaning and splitting the data-set, we use classification algorithms like Random Forest, Decision Tree, Naive Bayes and KNN on it. The results of these will then be fed to an ensemble. This process will be repeated multiple times to obtain a result with the highest accuracy acquired.
- Google Colab will be used for developing machine learning algorithms in python. It is

based on Jupyter notebook and supports collaborative development.

- Dataset Selection : The dataset is the most important aspect in prediction problems. The dataset has 567 instances and 25 attributes.
- Analyzing and transforming the Variables: Analysis of the attributes of the data is to be done to check the distribution of each of the variables and know the importance of the variable on the final results and check the relationships with other variables and with dependent variables.
- Random Sampling and Splitting the dataset for training and testing purposes: Here the ratio of training and testing dataset will be 7:3. The training dataset will be used for training the model and test for validation purposes by each classification model.
- Implementing the classification algorithms: Ensemble Learning basically combines more than one machine learning algorithms to give better results compared to the individual machine learning algorithms.

Two steps involved in the development of Ensemble learning methods are:

A) The Prediction is made by the following classifiers: KNN, Decision Tree, Random Forest and Naive Bayes.
B) Soft Voting is used, the summing of the Predicted Probabilities for class labels is taken and the prediction with the largest sum probability is made. Hard voting takes prediction from all the classifiers and predicts the class that gets the most votes. Soft voting takes into account how certain each voter is rather than taking binary decisions based on the output of the voter. Soft Voting is more specific and can improve on hard voting since it takes into account more information than the hard voting, it considers each classes uncertainty in the final decision.

Design and Implementation

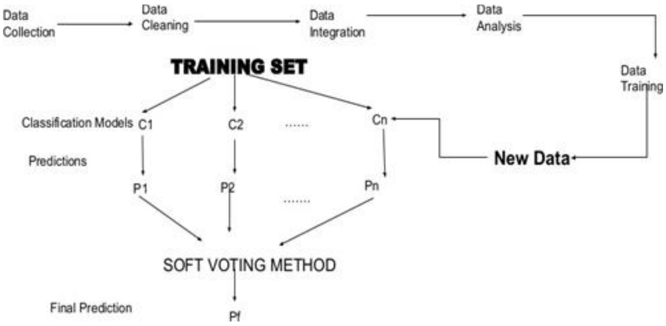


Fig. 2. Design of the System

We first Collect the data, then clean it. Data integration means changing it to a format that is conducive to our work. Which for us is csv. Data Analysis included understanding what the data implied and the different symptoms for Lung Cancer that we could identify.

The Data training phase starts after this. The now new data is processed using multiple machine learning models, and their predictions are saved. There may be multiple runs of a single model with different parameters, depending on the ensemble.

The ensemble is then generated using the Soft Voting Method, which enables us to then get the final prediction for that ensemble. We then repeat this process for as many ensembles as we think is necessary before we find one that suitably demonstrates our hypothesis, that is, it has a higher accuracy than a single model and does use any models that are relatively out of the norm.

The Development Environment used here is Google Collaboratory Notebook, it is a free Jupyter Notebook Environment running on cloud which supports machine learning Libraries. The notebook created can be simultaneously edited by group members and also can be saved on Google drive in .ipynb format. Python is used for implementing machine learning algorithms, various libraries used are Numpy, Pandas, Matplotlib, Scikit-Learn. Numpy is a module available for computing, it contains powerful dimensional array objects and matrices and different mathematical functions which can be implemented on the arrays. Pandas is an Open source Data Analysis Tool in Python Language which comprises Dataframe Data Structure and also used for importing and managing datasets. It is an important part used for statistical operations used in python. Matplotlib is a library used for creating visualizations of data Analysis in the form of graphs. Scikit-Learn is used for implementing Different Machine learning classification models. Scikit- Learn provides easy access for the classification algorithms used in this Project.

Data cleaning and Preprocessing is done after the loading of the data, the data-set available should be in the proper format before it can be used for the training the model. The classification algorithms require numerical data so for the Column Gender ’1’ was replaced in place of Male and ’2’ in place of Female. The attributes in the dataset have the values in nominal range.

Data Analysis is done to understand the dataset and the pattern of the attributes. Bivariate Analysis was done where the importance of the variable on the Target column, to know the pattern of the results. The correlation among the variables was calculated and visualized in the form of Heat Map.

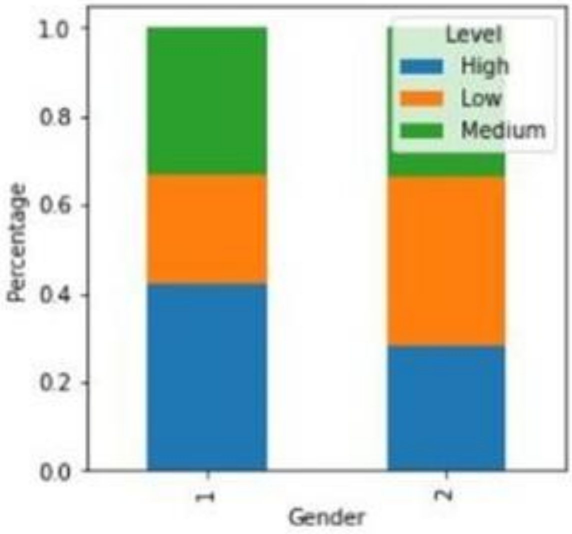


Fig. 4. Correlation between Lung Cancer and Gender

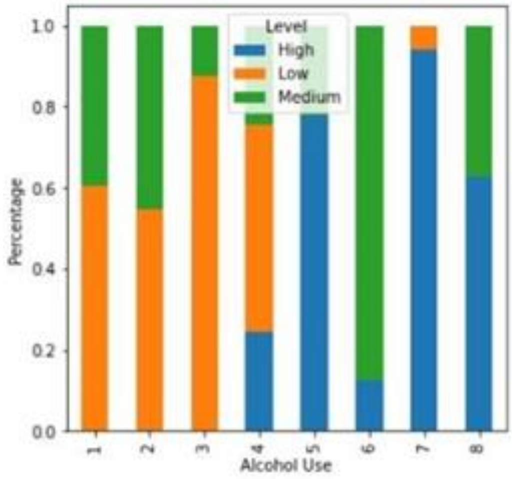


Fig. 5. Lung Cancer chances on Alcohol Use

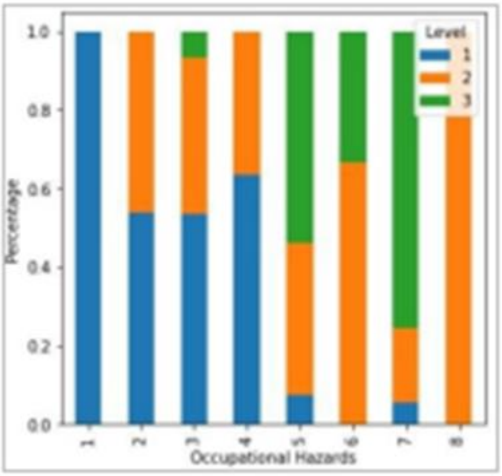


Fig. 6. Cancer chances on Occupational Hazards

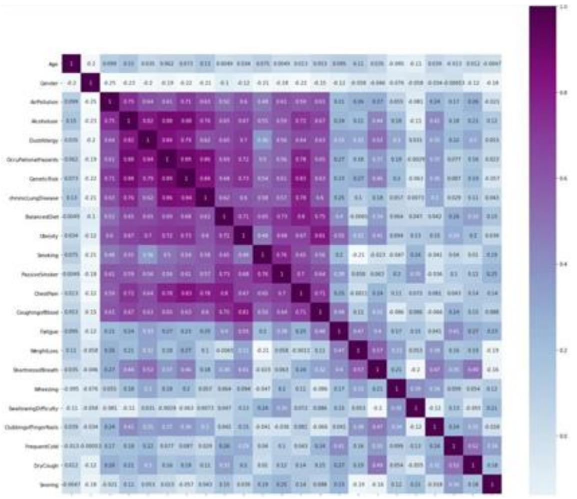


Fig. 7. Attribute Heatmap

III. RESULTS AND DISCUSSION

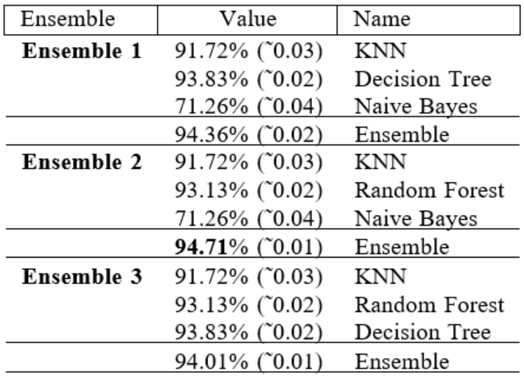
We compared three ensembles, and these ensembles were:

- Decision Tree, KNN, Naive Bayes
- Random Forest, KNN, Naive Bayes
- Random Forest, KNN, Decision Tree

There are 24 attributes in the data set responsible for the Results. Understanding the dataset and knowing the dependent variables before implementing the classification algorithm. The Heat Map showing the correlation of the variables shows that the most correlated variables are Occupational Hazards, Dust Allergy, Genetic Risk, Chronic Lung Disease. Bivariate Analysis of Alcohol Use with the Result shows that higher the value of this variable more the chances of getting Lung Cancer. The Dataset was split in 7:3 ratio for training and validation purposes.

KNN classification model implemented gives accuracy of 91.72%, the number of neighbours set is 5. Decision Tree implemented gives an accuracy of 93.83%, all the parameters are set to default. Random forest model gives accuracy of 93.13%. Naive Bayes gives accuracy of 71.29%.

Three ensembles were used. The Decision Tree, KNN and Naive Bayes Ensemble gave us an accuracy of 94.36%, while the Random Forest, KNN and Naive Bayes Ensemble gave us an accuracy of 94.71%. The Random Forest, KNN, Decision Tree Ensemble gave us an accuracy of 94.01%.



Ensemble	Value	Name
Ensemble 1	91.72% (˜0.03)	KNN
	93.83% (˜0.02)	Decision Tree
	71.26% (˜0.04)	Naive Bayes
	94.36% (˜0.02)	Ensemble
Ensemble 2	91.72% (˜0.03)	KNN
	93.13% (˜0.02)	Random Forest
	71.26% (˜0.04)	Naive Bayes
	94.71% (˜0.01)	Ensemble
Ensemble 3	91.72% (˜0.03)	KNN
	93.13% (˜0.02)	Random Forest
	93.83% (˜0.02)	Decision Tree
	94.01% (˜0.01)	Ensemble

The Random Forest, KNN and Naive Bayes Ensemble has the highest accuracy and should be used over the others.

IV. CONCLUSION

- In Lung Cancer treatment delay results in high mortality rate. Therefore detection of cancerous cells at early stage is very much important.
- Several Techniques are there predicting Lung Cancer but are more expensive time consuming, and have less capability of detecting the lung cancer.
- Ensemble Learning Model provides considerably better predictions performance than single models.
- The Random Forest , KNN and Naives Bayes. Ensemble has the highest accuracy and should be used over the other ensemble.
- In the future we hope to increase the accuracy of our system and also to hopefully find an ensemble that utilizes models that require low computational power over to give a comparatively more accurate outcome.

V. REFERENCES

[1]. E. Y. V. Chandra, K. R. Teja, M. C. S. Prasad, and M. Ismail, “Lung cancer prediction using data mining techniques,” Inter- national Journal of Recent Technology and Engineering, 2019.A

[2]. X. Ying, “Ensemble learning,” 05 2014.

[3]. N. Nai-arun and P. Sittidech, “Ensemble learning model for diabetes classification,” Advanced Materials Research, vol. 931- 932, pp. 1427–1431, 05 2014.

[4]. S. Dr. Senthil and A. B, “Lung cancer prediction using feed for- ward back propagation neural networks with optimal

features," International Journal of Applied Engineering Research, vol. 13, pp. 318–325, 10 2018.

[5]. P. Mung and S. Phyu, "Effective analytics on healthcare big data using ensemble learning," pp. 1–4, 02 2020.

[6]. P. Pintelas and I. E. Livieris, "Special issue on ensemble learning and applications," Algorithms, vol. 13, p. 140, Jun 2020.

[7]. D. Crankshaw, X. Wang, G. Zhou, M. J. Franklin, J. E. Gon- zalez, and I. Stoica, "Clipper: A low-latency online prediction serving system," in 14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17), (Boston, MA), pp. 613–627, USENIX Association, Mar. 2017.

[8]. J. Mendes-Moreira, A. M. Jorge, C. Soares, and J. F. de Sousa, "Ensemble learning: A study on different variants of the dy- namic selection approach," in Machine Learning and Data Mining in Pattern Recognition (P. Perner, ed.), (Berlin, Heidelberg), pp. 191–205, Springer Berlin Heidelberg, 2009.

[9]. J.D. Hunter, "Matplotlib: A 2d graphics environment," Com- puting in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.

[10]. L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller,O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler,R. Layton, J. Vanderplas, A. Joly, B. Holt, and G. Varoquaux, "Api design for machine learning software: experiences from the scikit-learn project," 2013.

[11]. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel,B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Ma- chine learning in Python," Journal of Machine Learning Re- search, vol. 12, pp. 2825–2830, 2011.

Cite this article as :

Vaibhav Narawade, Akash Singh, Mohit Shrivastava, Abhishek Prasad, "Lung Cancer Prediction Using Ensemble Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7, Issue 3, pp.477-482, May-June-2021. Available at
doi : https://doi.org/10.32628/CSEIT217357
Journal URL : https://ijsrcseit.com/CSEIT217357