Research Article

Comparing the Performance of Algorithm with Relevant Features for Histological Categorization of Lung Cancer

V. NishaJenipher[a], Aruna Jasmine[b], K. Ravindran[c], and J. S. Richard Jimreeves[d]

[A] Assistant Professor,CSE Department, St. Joseph's Institute of Technology,Chennai,India
[b]Assistant Professor, IT Department, Jeppiaar Institute of Technology, Chennai,India
[C]Assistant Professor,ITDepartment, Easwari Engineering College,Chennai, India
[d]Assistant Professor, IT Department, Easwari Engineering College, Chennai,India

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Due to increasing cancer cases around the world, Lung cancer has become the favorite topic of research for a long period of time. The actual reason is due to the increasing rate of new cases across the globe. Therefore, many researchers used prediction or classification algorithm to identify the factors that contribute to the increase of this deadly disease. Two models were built namely WRF and RF. RF model provides the result of features selected by a predominant feature selection method whereas WRF model provides result of all features without performing any selection process. A comparison is made to inform the importance of selecting the feature for classification or prediction algorithm. The accuracy provided by WRF model is higher than RF model which highlights the importance of selecting the feature for classification algorithm.

Keywords: histology, accuracy, classification, prediction

1. Introduction

Around the globe, Lung cancer (LC) is most repeatedly identified cancer in 37 countries and it is responsible for high death rate in males [1]. Unlike other cancer cells, lung cancer patients have higher survival rate, once detected earlier. There are many histological categorizations in the Lung cancer cells [2]. Based upon the size of the cancer cell, they are classified into many types [3]. Certain type ofcancer cells is frequently found in heavy smokers than non-smokers, also the progress of particular type of lung cancer cell is higher in non-smokers [4]. Though there are many parameters contributing to the development of Lung cancer, the exact reason is not known. Therefore, many prediction and classification algorithms are used to find out features that contribute to this deadlydisease.

The aim of this paper are as follows

- This work identifies appropriate features that are related to histological categorization of cancercells.
- This work has created two models. One model provides the result of features selected by a predominant feature selection method and another model provides result of all features without performing any selection process. A comparison is made to inform the importance of selecting the feature for classification or predictionalgorithm.
- Performance of these two models are evaluated to determine the bettermodel.

II. Related works

Wail A.H Mousa et al. [5] used an SVM classifier that provided sensitivity of 87.5%. Swati P. Tidke et al. [6] developed a model to classify the cancer cells. Input image is preprocessed. segmentation using thresholding is done followed by certain operations and an accuracy of 92.5% was shown. Elmar Rendon-Gonzalez et al. [7] employed SVM algorithm for classification. The model developed includes preprocessing step, segmenting lung parenchyma , identifying nodule and produced 78.08% accuracy.

DmitriyZinovevet al. [8] evaluated an algorithm where Area under the curve (AUC) was used as a performance metric and it provided 69% performance. DmitriyZinovevet al. [9] built a classifiers for Lung Nodule Interpretation. It included some learning approach. Different strategy was employed and probabilistic labels are learned , therefore using them to form classifiers. M H Hasnaet al. [10] created a classifier that gave an accuracy of 80%. Sarah Soltaninejad et al. [11] built a classifier for detection ofnodule.

SakshiWasnik et al. [12] made used of k-nearest neighbors (KNN) algorithm classifier which provided an accuracy 96.25%. Three stage of implementation was done by P. Bhuvaneswariet al. [13] and the accuracy obtained was 90%. S.L.A. Lee etal.[14] provided 100 % true positive and 1.27 false positive per scan by random forest. SubratoBharatiet al. [15] gave a high accuracy texture and spiculation. Jose et al. [16]proposed medical image classification where Random forest performed well and produced an 92%accuracy.

3. Systemarchitecture

Fig.1. shows the workflow methodology of our system architecture which includes

1. Histological categorization of lung cancerdataset
2. Cleaning the missingvalue
3. Data Visualization and Feature selection
4. Supervised ML algorithm andResult

3.1 Histological categorization of lung cancerdataset

The dataset has been collected from Cancer Imaging Archive (TCIA). National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) [17] and Clark. K. et al. [18] used the same dataset for their research purpose. The clinical dataset contain 113 patient clinical data with 43 features including the prediction variable. Table 1 lists the features in the dataset.

S. No.	Feature Name	S. No.	Feature Name
1	Tumor code	23	Vital_status_at_24months_follow_up
2	Case Id	24	Residual_tumor
3	Gender	25	Alcohol_consumption
4	Age	26	Tobacco_smoking_history
5	Height in cm	27	Number_of_pack_years_smoked
6	Weight in kg	28	Tumor_status_at_12months_follow_up
7	BMI	29	Cause_of_death_at_12months_follow_up
8	Race	30	Days_from_initial_pathologic_diagnosis_to_death_at_12months_follow_up
9	Ethnicity	31	Tumor_status_at_24months_follow_up
10	Tumor_site	32	Cause_of_death_at_24months_follow_up
11	Tumor_site_other	33	Days_from_initial_pathologic_diagnosis_to_death_at_24months_follow_up
12	Tumor_size_in_cm	34	Days_from_initial_diagnosis_to_last_contact_at_12months_follow_up
13	Histologic_type	35	Days_from_initial_diagnosis_to_last_contact_at_24months_follow_up
14	Histologic_type_other	36	Specimens_specimen_id
15	Histologic_grade	37	Specimens_slide_id
16	Tumor_stage_pathological	38	Specimens_tissue_type
17	AJCC_or_TNM_cancer_staging_edition	39	Number_of_years_consumed_more_than_2_drinks_per_day_for_men_or_more_than_1_for_women
18	Pathologic_staging_primary_tumor_pt	40	Specimens_percent_tumor_surface_area
19	Pathologic_staging_regional_lymph_nodes_pn	41	Specimens_percent_tumor_nuclei
20	Pathologic_staging_distant_metastasis_pm	42	Specimens_percent_necrotic_surface_area
21	Clinical_staging_distant_metastasis_cm	43	Specimens_weight_in_mg
22	Vital_status_at_12months_follow_up		

Table 1 Lists of features in the dataset.

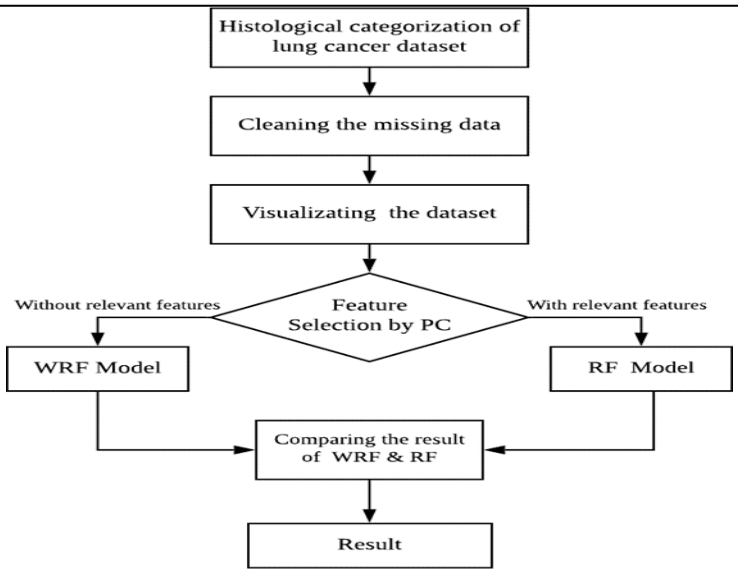
Histological categorization of lung cancer dataset
Cleaning the missing data
Visualizating the dataset
Feature Selection by PC
Without relevant features
With relevant features
WRF Model
RF Model
Comparing the result of WRF & RF
Result

Figure 1. System Architecture

3.2 Cleaning the missingvalue

The real-world data contain unsuitable, irrelevant values and missing values. Michael J. Hassett et al. [19] in his research work dropped the missing values. In our dataset there are features which have more missing values and the feature that has more than 20% of missing values are dropped. The qualified features are age, gender, Height, Weight, BMI, tumor size and histology and these features have only small portion of values as missing and they are replaced by attribute mean. Table 2 provide the features description of the dataset to be used in the following steps.

3.3 Data Visualization and FeatureSelection

Data is managed, prepared and cleaned to make it available for visualization. Many data exploring techniques are available to know and to infer conclusions based on the requirements. Some of the visualization tools used are scikit learn, tableau, Qlikview, FusionCharts and HighCharts. Data Visualization uses presentation, to gain added understanding about the information within the data. Scikit learn is used for implementation purpose. Figure 2 depicts the bivariate distributions in the dataset. Figure 4, Figure 5, Figure 6 provides distribution of data in the dataset and Figure 7, Figure 8, Figure 9 provides relationship between features.

The primary goal of selecting the features is to recognize those attributes or highlights which are associated with yield esteems where the qualities rely on a particular information which is gathered by applying some valuable test. Usually in statistics, the correlation used are Pearson correlation (PC), kendall rank correlation, Spearman correlation and Point-biserial correlation. In our work we use PC to measure of the strong point of a linear association between variable. The rightness and adequacy of histological categorization of lung cancer can be done by selecting the right features. Selecting the dominant features by PC has been done by AnimeshHazra et al. [20] in predicting the survivability of lung cancer patient dataset. The negative correlation with histology is considered as irrelevant features whereas all the positive correlation is considered as important features. The feature selected are age, height and weight while BMI and tumor size are considered as irrelevant features by PC. Figure 4 shows the feature selection process of PC.

Feature	**Description**	**Type**	**Min**	**Max**
Age	Age in years	numeric	40	88
Gender	0: male, 1: female	categorical	0 (male)	1 (female)
Height	Patients height in cm	numeric	72	200
Weight	Patients weight in Kg	numeric	43	168
BMI	Patients Body mass index	numeric	16.61	324.07
Tumor Size	Patient tumor in cm	numeric	1	10
Histology (target variable)	Patient histology type (The types range from 0 10 5)	numeric	0	5

Table 2. Feature description

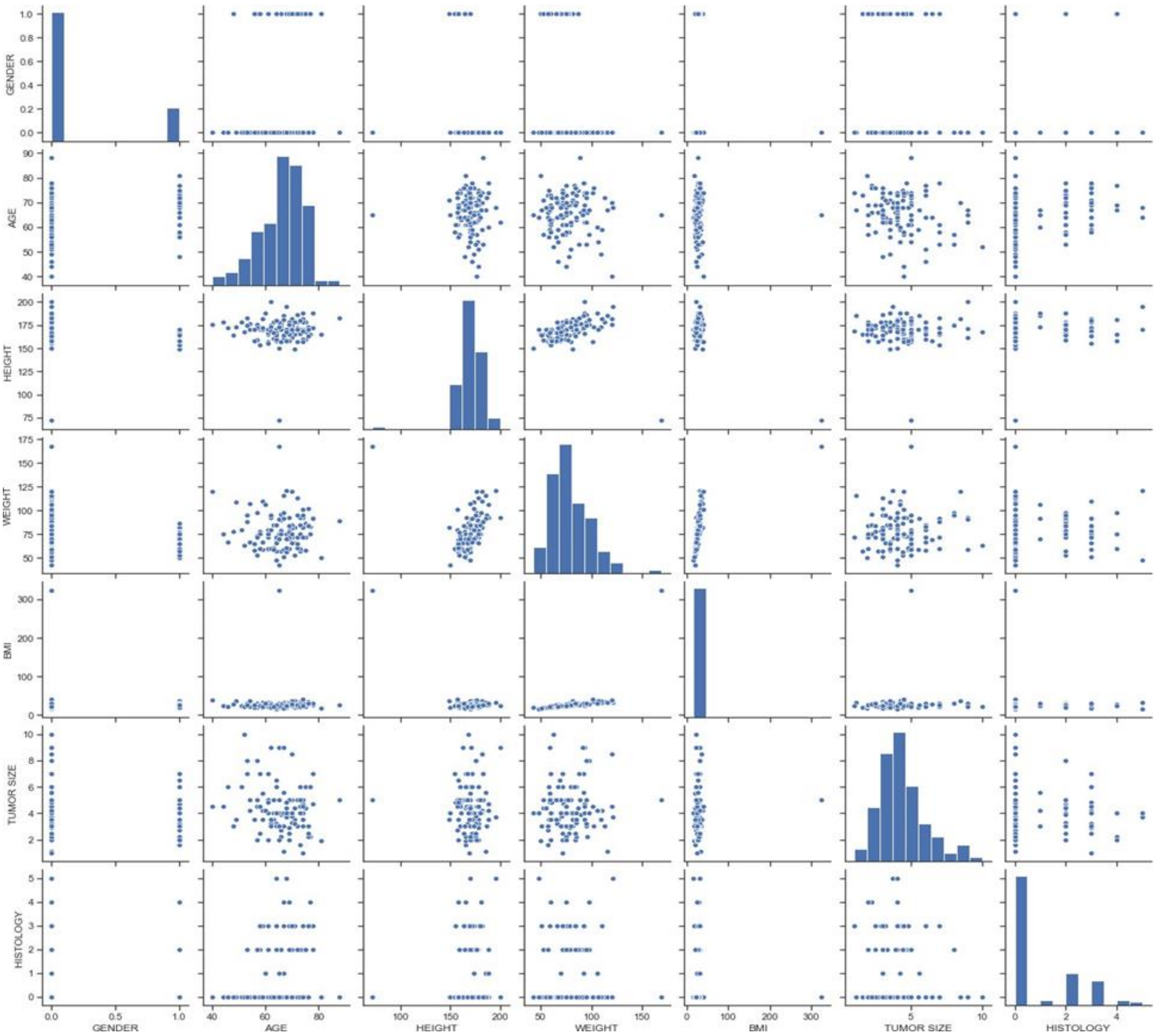

Figure 2. Distributions in the dataset

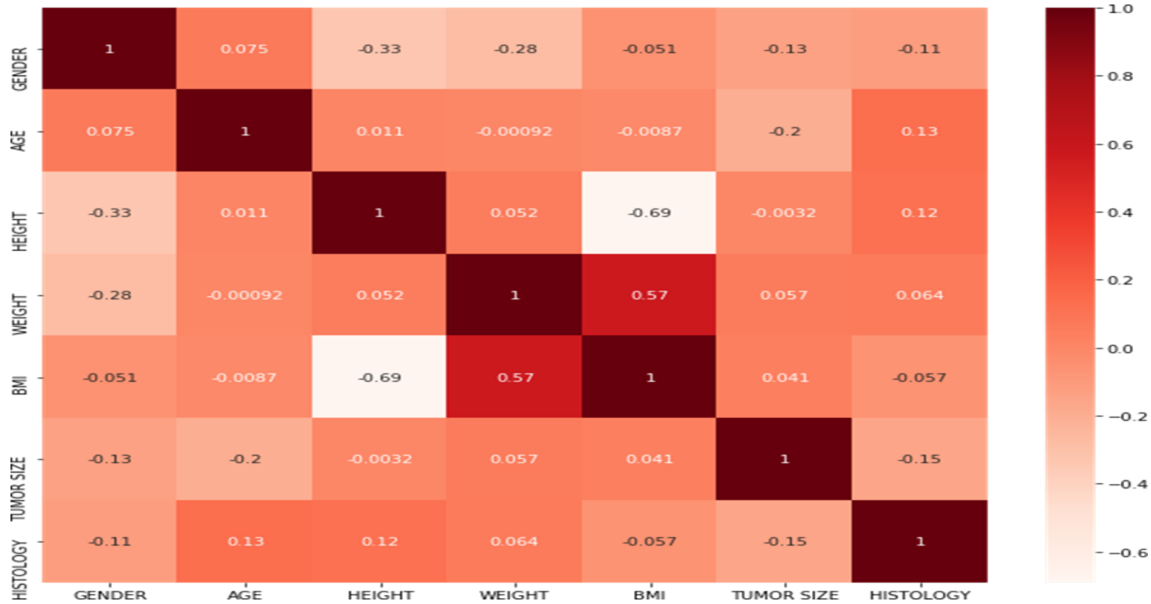

Figure 3. Pearson Correlation matrix using heatmap

3.4 Comparison of model(WRF & RF) andResult

Two models are created namely WRF and RF for histological classification. Dataset with all features are loaded to WRF and dataset with features selected by PC are loaded into RF.

WRF and RF model comprises of machine learning algorithm such as support vector machine (SVM), Logistic Regression (LR), Decision tree (DT), K-Nearest Neighbor (KNN) and Random Forest (RAF). The selected features by PC are given as input to RF model whereas all the features without undergoing feature selection by PC are given as input to WRF model. The accuracy produced by RF model and WRF model are compared. SVM and RF algorithm with feature selection produced greater accuracy of 73.529% than other algorithm. Figure 10 provide the comparison of algorithm with and without feature selection.

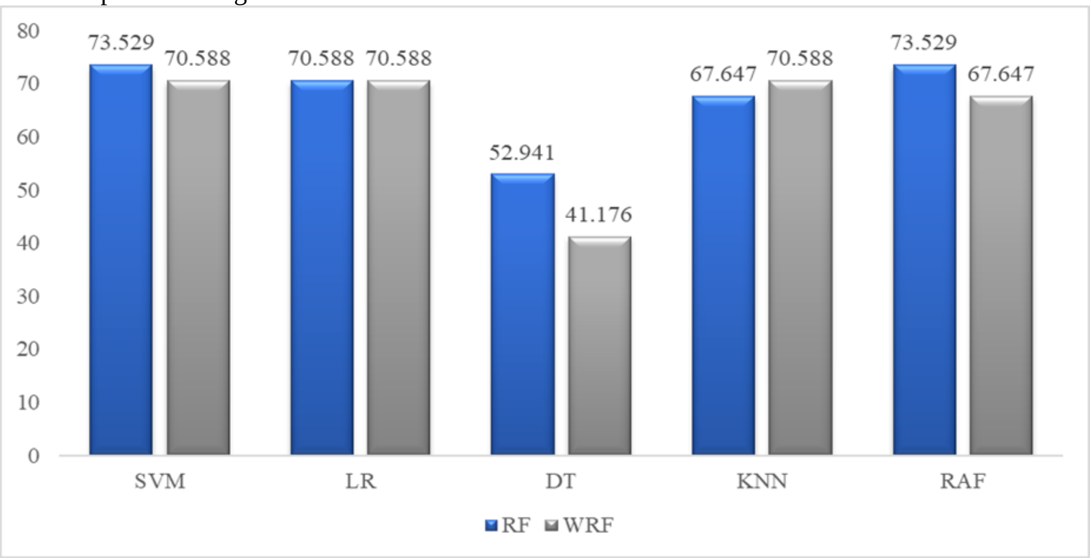

Figure 10. Comparison of result produced RF and WRF model

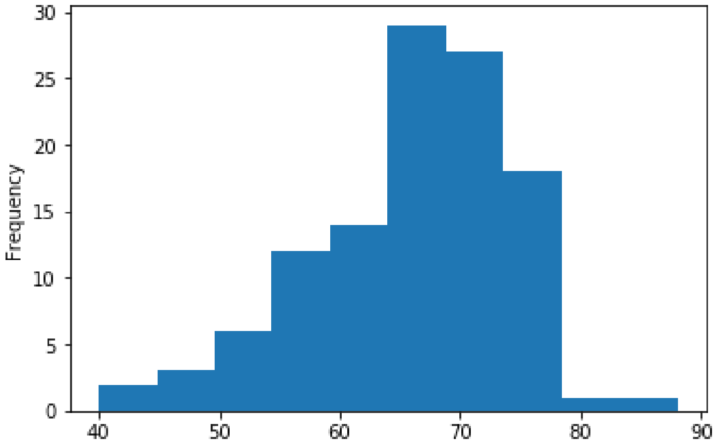

Figure 4. Age in Frequency

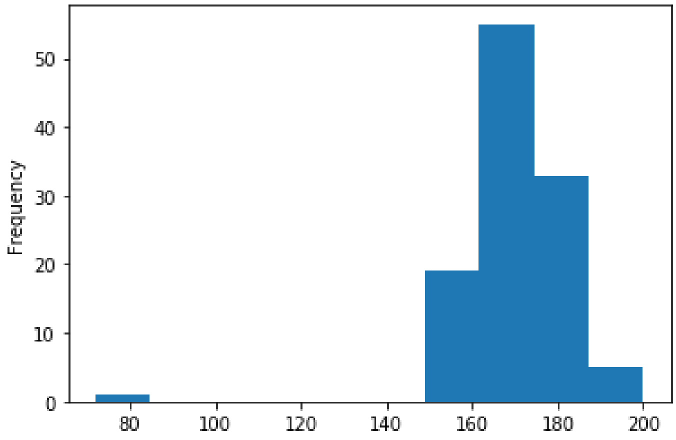

Figure 5. Height in frequency

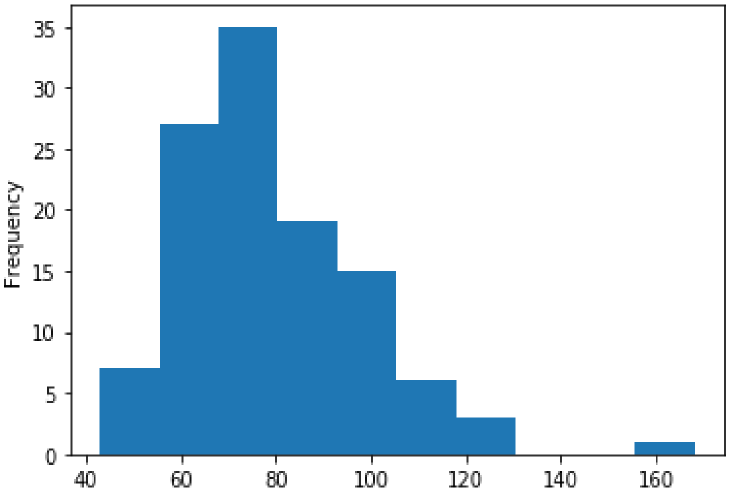

Figure 6. Weight in frequency

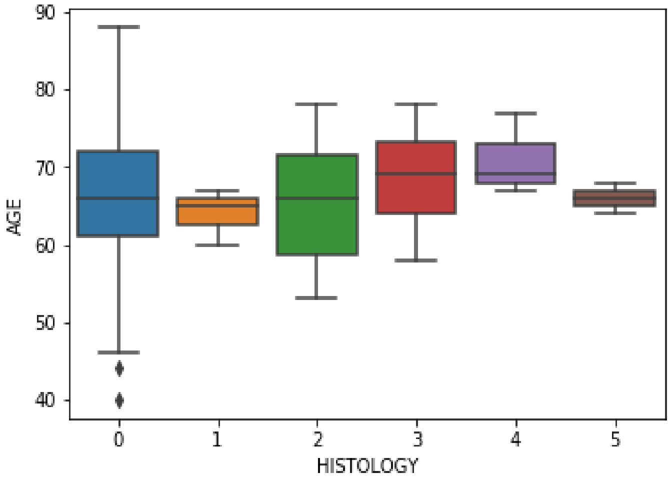

Figure 7. Relationship between histology and age.

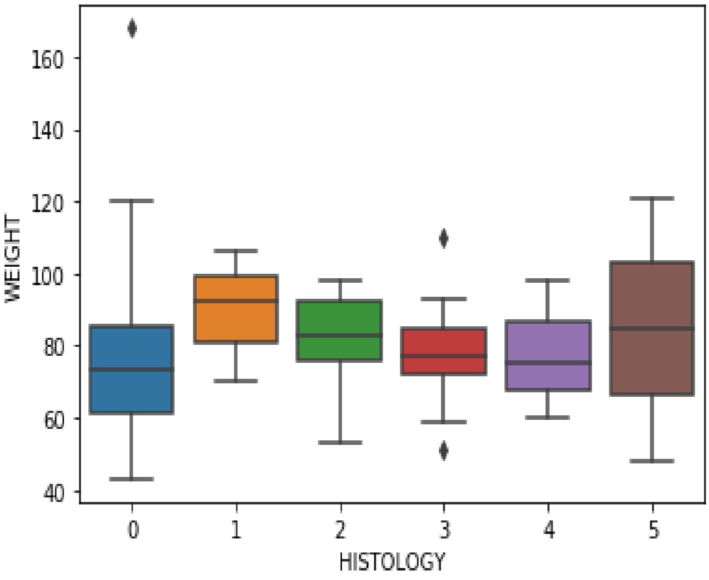

Figure8. Relationship between histology andweight.

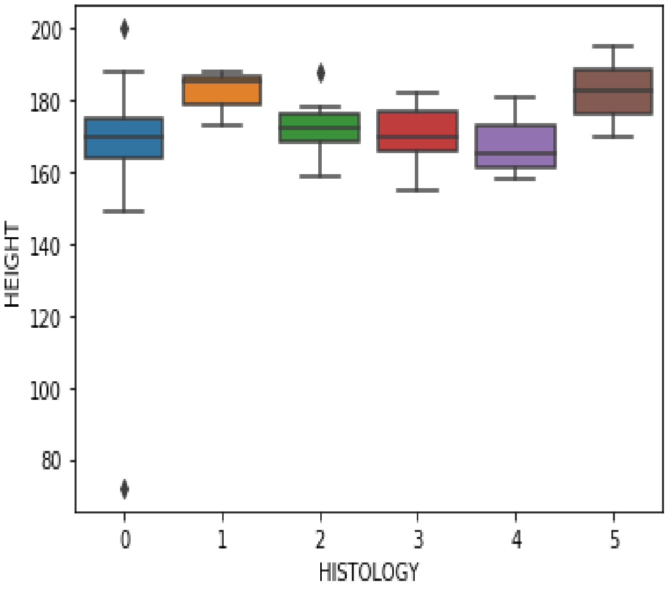

Figure9. Relationship between histology andheight.

4. Summary of Current Work

In this section, we summarize our current research work as follows:

1. Input data collected from cancerimagingarchive.net undergoes cleaning process to eliminate missingvalues.
2. data visualization is done by ScikitLearn
3. Features are selected usingPC.
4. Two models are created namely WRF and RF for histological classification. Dataset with all features are loaded to WRF and dataset with features selected by PC are loaded intoRF.
5. Accuracy provided by WRF model algorithm is compared with RF model algorithm. WRF and RF model comprises of machine learning algorithms. SVM and RF algorithm with feature selection produced greater accuracy of 73.529% than other algorithm.

5. Conclusion

In this paper, we have created two models WRF and RF which comprises of machine learning algorithm. Dataset with all features are loaded to WRF and dataset with features selected by PC are loaded into RF model. Accuracy provided by WRF model algorithm and RF model algorithm are compared. SVM and Random Forest algorithm with feature selection produced greater accuracy of 73.529% than other algorithm. This informs the need of selecting the feature while predicting some deadly disease like lung cancer. In future, research work can be made to improve the accuracy of classification or predictionalgorithm.

REFERENCES

1. Freddie Bray, BSc, MSc, PhD; Jacques Ferlay, ME; Isabelle Soerjomataram, MD, MSc, PhD; Rebecca L. Siegel, MPH; Lindsey A. Torre, MSPH; AhmedinJemal, PhD,DVM:Global CancerStatistics2018:GLOBOCANEstimatesofIncidenceandMortalityWorldwidefor36Cancersin185Countries. 12September2018.
2. India against Cancer:“http://cancerindia.org.in/lung-cancer/”.
3. Kiruthika, S. U., S. K. S. Raja, and R. Jaichandran. "IOT based automation of fish farming." Journal of Advanced Research in Dynamical and Control Systems 9 (2017): 50-57.
4. Noronha V,DikshitR1 ,Raut N2 , Pramesh CS3 , Karimundackal G3 , Agarwal JP4 , Munshi A4 , Kumar P. Epidemiology of lung cancer in India: Focus on the differences between non-smokers and smokers: A single-center experience , DOI:10.4103/0019-509X.98925.
5. Wail A.H Mousa and Mohammad A. U Khan. Lung nodule classification utilizing support vector machines. International Conference on Image Processing.doi:10.1109/ICIP.2002.1038927.
6. Ms. Swati P. Tidke, Prof. Vrishali A. Chakkarwar. Classification of Lung Tumor Using SVM. International Journal of Computational Engineering Research Vol. 2 Issue.5.
7. Elmar Rendon-Gonzalez, VolodymyrPonomaryov Automatic Lung Nodule Segmentation and Classification in CT Images Based on SVM. -24 June 2016, Kharkiv, Ukraine. DOI:10.1109/MSMW.2016.7537995.
8. DmitriyZinovev, Jonathan Feigenbaum, Jacob Furst, and Daniela Raicu. Probabilistic Lung Nodule Classification with Belief Decision Trees. doi:10.1109/IEMBS.2011.6091114.

9. DmitriyZinovev , Jacob Furst , Daniela Raicu .Building an Ensemble of Probabilistic Classifiers for Lung Nodule Interpretation.10.1109/ICMLA.2011.44.
10. M H Hasna, Jobin Jose. International Research Journal of Engineering and Technology. Lung nodule classification using multilevel patch-based context analysis and decision tree classifier. Volume: 02 Issue: 03June-2015.
11. Sarah Soltaninejad, Mohsen Keshani, FarshadTajeripour. Lung Nodule Detection by KNN Classifier and Active Contour Modelling and 3D Visualization. doi: 10.1109/AISP.2012.6313788. 27 September2012.
12. SakshiWasnik, PallaviParlewar, PrashantNimbalkar. Detection of Cancerous Nodule in Lung Using KNN Classifier. doi10.29042/2019-5779-5783.
13. P.Bhuvaneswari,Dr.A.BrinthaTherese. DetectionofCancerinLungWithK-NNClassificationUsingGeneticAlgorithm. doi.org/10.1016/j.mspro.2015.06.077.
14. LeeS.L. A.,kouzani,A. Z.andHu, E.J.2008:ARandomforesetforlungnoduleIdentification,IEEEregion10conference,IEEE,Piscataway, N.J.,pp.1-5.
15. SubratoBharati, PrajoyPodder and Pinto Kumar Paul. Lung cancer recognition and prediction according to random forest ensemble and RUS Boost algorithm using LIDC data. International Journal of Hybrid Intelligent Systems 15 (2019) 91–100. doi10.3233/HIS-1902.
16. D. Jose, A. N. Chithara, P. Nirmal Kumar, and H. Kareemulla, "Automatic Detection of Lung Cancer Nodules in CT Images," National Academy Science Letters, vol. 40, no. 3, pp. 161–166,2017.
17. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). (2018). Radiology Data from the Clinical Proteomic Tumor Analysis ConsortiumLungSquamousCellCarcinoma[CPTAC-LSCC]Collection[Dataset].TheCancerImagingArchive.https://doi.org/10.7937/k9/tcia.2018.6emub5l2.
18. Clark, K., Vendt, B., Smith, K. et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging 26, 1045– 1057 (2013).https://doi.org/10.1007/s10278-013-9622-7.
19. MichaelJ.Hassett,MD,MPH,wHajimeUno,WAngelM.Cronin,MS,NikkiM.Carroll,MS,zMarkC.Hornbrook,PhD,yandDebraRitzwoller,PhDz
20. .Detecting Lung and Colorectal Cancer Recurrence Using Structured Clinical/Administrative Data to Enable Outcomes Research and Population Health Management. Doi: 10.1097/MLR.0000000000000404.
21. AnimeshHazra ,NanigopalBera , AvijitMandal. Predicting Lung Cancer Survivability using SVM and Logistic Regression Algorithms. International Journal of Computer Applications (0975 – 8887) Volume 174 – No.2, September2017.
22. Sampathkumar, A., Murugan, S., Sivaram, M., Sharma, V., Venkatachalam, K. and Kalimuthu, M., 2020. Advanced Energy Management System for Smart City Application Using the IoT. In Internet of Things in Smart Technologies for Sustainable Urban Development (pp. 185-194). Springer, Cham.
23. Sampathkumar, A., Murugan, S., Rastogi, R., Mishra, M.K., Malathy, S. and Manikandan, R., 2020. Energy Efficient ACPI and JEHDO Mechanism for IoT Device Energy Management in Healthcare. In Internet of Things in Smart Technologies for Sustainable Urban Development (pp. 131-140). Springer, Cham.