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

V. Nisha Jenipher\textsuperscript{a}, Aruna Jasmine\textsuperscript{b}, K. Ravindran\textsuperscript{c}, and J. S. Richard Jimreeves\textsuperscript{d}

\textsuperscript{a} Assistant Professor, CSE Department, St. Joseph’s Institute of Technology, Chennai, India
\textsuperscript{b} Assistant Professor, IT Department, Jeppiaar Institute of Technology, Chennai, India
\textsuperscript{c} Assistant Professor, IT Department, Easwari Engineering College, Chennai, India
\textsuperscript{d} Assistant Professor, IT Department, Easwari Engineering College, Chennai, India

\textbf{Article History:} Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

\textbf{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.

\textbf{Keywords:} histology, accuracy, classification, prediction

\section*{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 of cancer 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 deadly disease.

The aim of this paper are as follows

\begin{itemize}
  \item This work identifies appropriate features that are related to histological categorization of cancer cells.
  \item 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 prediction algorithm.
  \item Performance of these two models are evaluated to determine the better model.
\end{itemize}

\section*{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.

Dmitriy Zinovevet al. [8] evaluated an algorithm where Area under the curve (AUC) was used as a performance metric and it provided 69\% performance. Dmitriy Zinovevet 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 of nodule.

Sakshi Wasnik 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. Subrato Bharatiet 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. System architecture

Fig. 1. shows the workflow methodology of our system architecture which includes
1. Histological categorization of lung cancer dataset
2. Cleaning the missing value
3. Data Visualization and Feature selection
4. Supervised ML algorithm and Result

3.1 Histological categorization of lung cancer dataset

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.

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

Table 1 Lists of features in the dataset.
Comparing the performance of algorithm with relevant features for histological categorization of Lung cancer

Figure 1. System Architecture

3.2 Cleaning the missing value
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 Feature Selection
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 estees 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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Type</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age in years</td>
<td>numeric</td>
<td>40</td>
<td>88</td>
</tr>
<tr>
<td>Gender</td>
<td>0: male, 1: female</td>
<td>categorical</td>
<td>0 (male)</td>
<td>1 (female)</td>
</tr>
<tr>
<td>Height</td>
<td>Patients height in cm</td>
<td>numeric</td>
<td>72</td>
<td>200</td>
</tr>
<tr>
<td>Weight</td>
<td>Patients weight in Kg</td>
<td>numeric</td>
<td>43</td>
<td>168</td>
</tr>
<tr>
<td>BMI</td>
<td>Patients Body mass index</td>
<td>numeric</td>
<td>16.61</td>
<td>324.07</td>
</tr>
<tr>
<td>Tumor Size</td>
<td>Patient tumor in cm</td>
<td>numeric</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Histology (target variable)</td>
<td>Patient histology type (The types range from 0 10 5)</td>
<td>numeric</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 2. Feature description

<table>
<thead>
<tr>
<th>Table 2. Feature description</th>
<th></th>
</tr>
</thead>
</table>

Figure 2. Distributions in the dataset

Figure 3. Pearson Correlation matrix using heatmap
3.4 Comparison of model (WRF & RF) and Result

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.
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 missing values.
2. Data visualization is done by ScikitLearn.
3. Features are selected using PCA.
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 into RF.
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 prediction algorithm.

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 Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. 12 September 2018.
2. India against Cancer: “http://cancerindia.org.in/lung-cancer/”.
Comparing the performance of algorithm with relevant features for histological categorization of Lung cancer

9. DmitriyZinovev, Jacob Furst, Daniela Raicu. Building an Ensemble of Probabilistic Classifiers for Lung Nodule Interpretation. 10.1109/ICMLA.2011.44.


19. MichaelJ.Hassett,MD,MPH,HajimeUno,WAngelM.Cronin,MS,NikkiM.Carroll,MS,ZMarkC.Horbrown,PhD,YandraDebraRitzwoller,PhDz


