14th National Convention on Statistics (NCS)
Crowne Plaza Manila, Ortigas Center, Quezon City
October 1-3, 2019

LUNG CANCER STAGE CLASSIFICATION USING RANDOM FOREST AND ARTIFICIAL NEURAL NETWORK

by

**Stanley T. Dalagan, Ella Joyce S. Paragas,
Aileen Joy V. Ramos, Clarissa Jewel B. Alota**

For additional information, please contact:

Author's name	Stanley T. Dalagan
Designation	Data Analyst
Affiliation	Proview Global Administration Inc.
E-mail	stanleydalagan@gmail.com

Author's name	Ella Joyce S. Paragas
Designation	Instructor
Affiliation	Central Luzon State University
Email	ellaparagagas@clsu.edu.ph

LUNG CANCER STAGE CLASSIFICATION USING RANDOM FOREST AND ARTIFICIAL NEURAL NETWORK

by

Stanley T. Dalagan, Ella Joyce S. Paragas,
Aileen Joy V. Ramos, Clarissa Jewel B. Alota

ABSTRACT

Lung cancer is considered the most common and the deadliest cancer type. The survival of a lung cancer patient is affected by a variety of factors including lung cancer stage or the measure of the spread of cancer within the body. To analyze the occurrence of the four lung cancer stages, Random Forest (RF) and Artificial Neural Network (ANN) were used to develop classification models. The predictor variables used are age at diagnosis, average number of cigarettes consumed per day, number of years smoked/ smoking, primary diagnosis, site of origin, gender, vital status, and year of birth. Both methods can be used to predict values of an ordinal dependent variable. From the model feature selection, RF with 100 decision trees with 2 factors per tree (or mtry) was used. The final ANN model for lung tumor stage classification was a 17-1-4 network. The ANN model was found to be better at classifying lung cancer stage than the RF model based on accuracy rate and Kappa statistic.

1. Introduction

Lung cancer or lung carcinoma is the leading cause of cancer death for men and women and is the most common cancer worldwide (Eldridge et al., 2019). In fact, 14% of all new cancer diagnoses are lung cancer and causes an average of 1.37 million deaths annually. In 2018, it was predicted to be the deadliest cancer type in the Philippines since it was responsible for 10,000 deaths of Filipinos in the previous year.

The Cancer Institute defines lung cancer as a malignant tumor characterized by uncontrolled cell growth in the tissues of the lungs. Cancers that start in the lung are Primary Lung Cancer which are mostly carcinomas (WHO, 2014). Carcinomas are tumors that may spread to adjacent tissues and other parts of the body in the process of metastasis. As a common place for tumor spread, primary lung cancers mostly metastasize to the brain, bones, liver and adrenal glands in cancer's latter stages (Lu et al, 2010).

The discoveries of the presence of lung cancer are sometimes accidental thus definitive diagnosis can be done through medical imaging, tissue biopsy and histological examination. Treatments vary according to the specific cell type, its spread, and the patient's overall health status. Surgery, Chemotherapy and radiography are common treatment options for lung cancer while targeted therapy are being applied to advanced lung cancer nowadays and several newly developed treatments like Immunotherapy are still on trials. Unfortunately, most cases of lung cancer are non-curable thus, avoidance of prognosis is estimated to be less than 20% and is affected by a range of influences such cancer stage, age, gender, race, time of diagnosis, and reaction to treatments.

Assessment of how much cancer is in an individual's body and its location is cancer staging. It describes the severity of the disease based on the magnitude of the original tumor (primary) as well as the extent of its spread in the body which considered the most important predictor of survival.

Tumor stage is a prognostic or predictive factor when it comes to lung cancer, it greatly influences treatment options and plans for a patient.

With the optimal goal of bettering survival rate of lung cancer patients, giving them seamlessly tailored treatments and prediction of prognostic factors including tumor stage significantly need improvement. Prediction in medicine is no longer new. Big clinical data in combination with modern machine learning enables the rapid generation of these clinical predictive models for thousands of medical questions (Chen & Asch, 2018). Deep learning algorithms' potential applicability makes analyzation of complex big data that was previously unimaginable, possible. Thus, it is critical for anyone practicing medicine in this age of such data to shift their attention to new statistical tools in the field of machine learning (Obermeyer & Emanuel, 2016).

The growing popularity of machine learning methods such as Random forest (RF) and Artificial Neural Network (ANN) on medical area has defined a new standard of medical diagnostic and prediction in terms of accuracy and usage of immense clinical datasets. These two powerful algorithms that are able to perform classification, regression and other statistical methods are extensively used in health research that covers medical image classification, cancer diagnosis, genomics, and DNA matching hence, its application on lung cancer data is already being practiced for years. In research, classical statistical methods infirmly handle classification of multi-level variables due to its limitation on the maximum number of categories to be categorized. Hence, the use of machine learning methods were explored.

2. Random Forests

Random Forest or random decision forest is a tree- based, ensemble algorithm that is useful in classification (shown in Figure 1). Decision trees mainly work on the principle of high level logic that humans have that is achieved by parting data set and writing rules which the later use to make decisions.

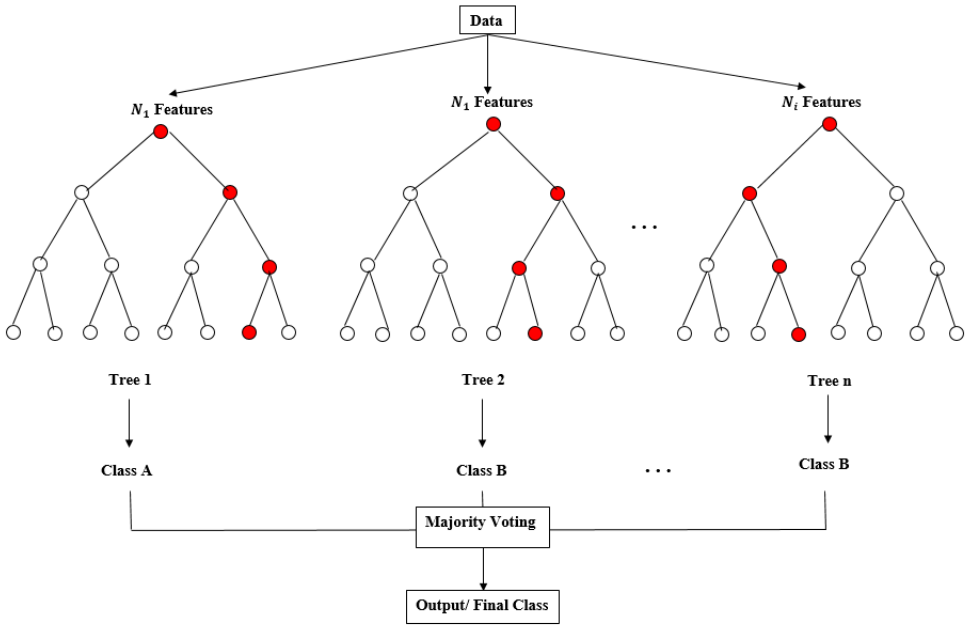

Figure 1. Random forest structure

Random forest is known to perform better or similar to deep learning methods. It can handle large and small data, easier to train and much faster to come up with results compared to some machine learning methods, less hyperparameters are required to tune, easy for distribution system and can deal with many features. RF can also be extended to unsupervised learning and have the GINI index feature that gives out the importance of features on the model. It works on unbalanced data, removes outliers and is robust to noise.

A decision tree rule can be viewed as a simple "if statement". These rules are organized in a tree structure where if the sample meets the first rule (some kind of a threshold), it goes down to the tree structure and is re-verified against the next rule until no rules are left and decisions are made. More trees, better generalizations. Random forests consisted of these decision trees were trained on a random bootstrapped sample of both the observations and the features. In this way each tree are approximately uncorrelated. If used in classification (when output is categorical), a consensus vote is taken to come up with an output/ decision.

3. Artificial Neural Networks

Artificial Neural Network tries to emulate how a biological brain works. There is no standard procedure of creating a neural network hence, developing a model is usually done in an iterative manner (Sydenham & Thorn, 2005).

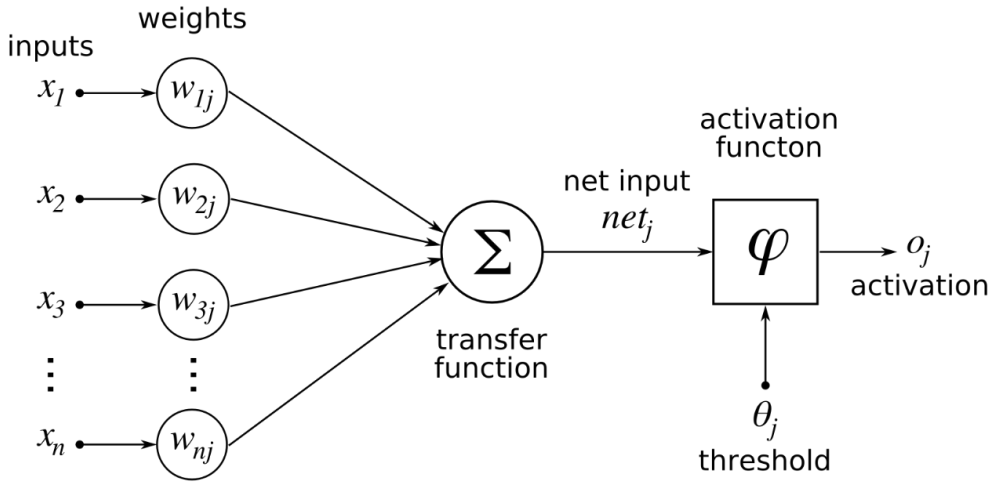

Figure 2. Architecture of a simple neural network

As indicated in Figure 2, predictor variables X_i will be fed into the input layer and associated weights will be adjusted to match the corresponding output value (feed-forward). Prediction will be propagated back to the network. This method known as Back Propagation determines a gradient to be used in weights calculation. Further adjustments of the weights will be done per new inputs introduction. This training process from which the model learns, will be performed repeatedly on the same training set of data until all the weights settled and converge (Sydenham & Thorn, 2005).

In the abundance of big data and computers, the increasing use of machine learning algorithms in performing more complex tasks has led to the explosion of multifaceted data analyses

application on every industry including the burning field of medical research. Its classification and predictive facility enables it to perform diagnosis, medical image cataloguing and prognosis estimation with superb accuracy. The following literature applied Artificial Neural Network (ANN) and Random Forest (RF) on lung cancer data from detection and diagnosis to filtering patients for aggressive treatments. All of which aims to better medical care thus, improving the chances of survival and cancer death reduction in general.

From available literatures on Random Forest and Artificial Neural Networks, researchers have used images and clinical features such as age, gender, ethnicity, tumor grade, tumor type, and tumor stage in their exploration of lung cancer data. It is also imperative to emphasize that some of these papers focused on how these algorithms can be used in deciding treatment options and predicting patient's prognosis – two of the most important aspects of survival. In line with this, this study tested the prediction capability of using clinical data alone with minimal number of features compared to previous studies.

Needless to say, most of these studies on lung cancer used a number of machine learning algorithms with the purpose of comparing each models' performance in terms of accuracy since no specific algorithm is superlative for a certain scenario (Wolpert, 1996). Researchers rely on model accuracy metrics and each models individual features. Some of the most widely used algorithms are Random Forest (RF) and Artificial Neural Network (ANN) which are two of the most powerful, highly accurate and flexible systems that can perform both regression and classification. Implementation of these researches in actual medical practice can significantly reduce overall cost of staging given that this will need minimal human involvement along with it is the reduction of human error. It will improve staging efficiency and in general, will revolutionize medical field.

4. Methodology

This study used a clinical data of 4,086 observations obtained from the National Cancer Institute's Genomic Data Commons Portal – a public cancer database that allows researchers to download harmonized datasets for analysis. The Lung cancer dataset (version 0.15), with cases from projects has been updated in February 2019. It was originally consisted of more than 25 features and was manually amputated to eliminate features that are mainly comprised of missing values or unreported observations. It was further filtered by removing observations without indicated tumor stage as it is the output. The final dataset is consisted of 383 cases and 9 features.

The output/dependent variable is the tumor stage of NSLC patients which is a four-category multinomial ordinal variable that ranges from Stage I, II, III, and IV. The predictor variables used are age at diagnosis, average number of cigarettes consumed per day, number of years smoked/ smoking, primary diagnosis, site of origin, gender, vital status, and year of birth.

Development of ANN and RF models both started with the random division of the dataset into two: training and testing set. The training set was composed of 80% of the original dataset while the remaining 20% percent of the observations was assigned in testing the models for validation (subjective allocation). Since algorithms learn from the training set, it was allocated with the bigger number of observations. Throughout the model training or development, the following optimizations was done.

The application of both algorithms required fine tuning of its parameters, in building an ANN model there were huge number of parameters to choose from like the number of layers, number of neuron, activation function and learning rate. Fine-tuning of these parameters were done in a repetitive manner until a desired model was obtained.

Implementation of the Random Forest to the training set to come up with the desired model started with tuning the parameters of the Random forest. Machine learning literatures noted the small numbers of hyperparameters to be tuned in RF making it one of the most used learning method.

To evaluate features for ANN classification, classical criteria use probabilistic distances or entropy measures, often replaced in practice by simple interclass distance measures. A variable subset which allows the best separation of the data was chosen. Variable selection is usually performed by considering a class separation criterion for the choice criterion and an associated F-test as stopping criterion. Leray & Gallinari (1999) cited that data separation is usually computed through an inter-class distance measure and the most frequent discriminating measure is the Wilks lambda.

Feature selection using Random forest comes under the category of Embedded Methods that combine the qualities of filter and wrapper methods implemented by algorithms that have built-in feature selection methods. It is known for its high accuracy, generalizability and interpretability. The importance of each feature is derived from how “pure” each of the buckets is. For classification, the measure of impurity is either the Gini impurity or the information gain/entropy.

It is possible to compute how much each feature decreases the impurity when training trees. The more a feature decreases the impurity, the more important the feature is. In random forests, the impurity decrease from each feature can be averaged across trees to determine the final importance of the variable. To give a better intuition, features that are selected at the top of the trees (parent tree nodes) are in general more important than features that are selected at the end nodes of the trees, as mostly the top splits lead to bigger information gains.

Model evaluation is done by producing the confusion matrix, accuracy and error rate, sensitivity and specificity. Cohen’s Kappa Statistic was also used to assess model’s accuracy as we are dealing with a multinomial output.

5. Results and Discussion

The lung cancer clinical dataset was filtered with indicated tumor stage and other variables that resulted to 383 cases without missing values. Exploration of the dataset had led to the exclusion of some pre-considered variables. The independent variable Morphology for example, is just encrypted/coded Primary Diagnosis (cancer types), which is the current catalogue from the World Health Organization (2004) hence, it was excluded in the model building. Also, days to birth, days to last follow up and ethnicity were removed due to the prevalence of missing values.

Majority of the patients are male (56.7%) and most were born on the 1930’s and 1940’s. More than half of the observed cases fall under the age group of 60 to 79 years old and with mean 67.69 years old, which is in line with the fact that 70-year old is its usual onset (National cancer Institute, 2016). More than half were reported alive on the time of data collection.

Majority of the patients' tumors had originated from the upper lobe (53%) and lower lobe (33.7%) of the lung, wherein surgical operations were done.

On exposure to cigarettes, 49.9% of the patients smoke 2-3 cigarettes daily. The patient's average daily cigarette consumption is 2.57 or 3 sticks. Majority of the cases smoked cigarettes for 40 to 49 years, 60 years being the longest. The average number of years the patients smoked was 36 years with a standard deviation of 13.08 in years.

More than half (52.7%) of the patients' lung cancer were under stage I. The frequency were noticeably less distributed in the last stages where only 3.1% of the observations fell under stage IV. One possible explanation is that on the first stages, complete surgical removal of the tumors is still possible compared to the advanced stages where removal is already unlikely as the cancer already metastasized or spread beyond the lungs.

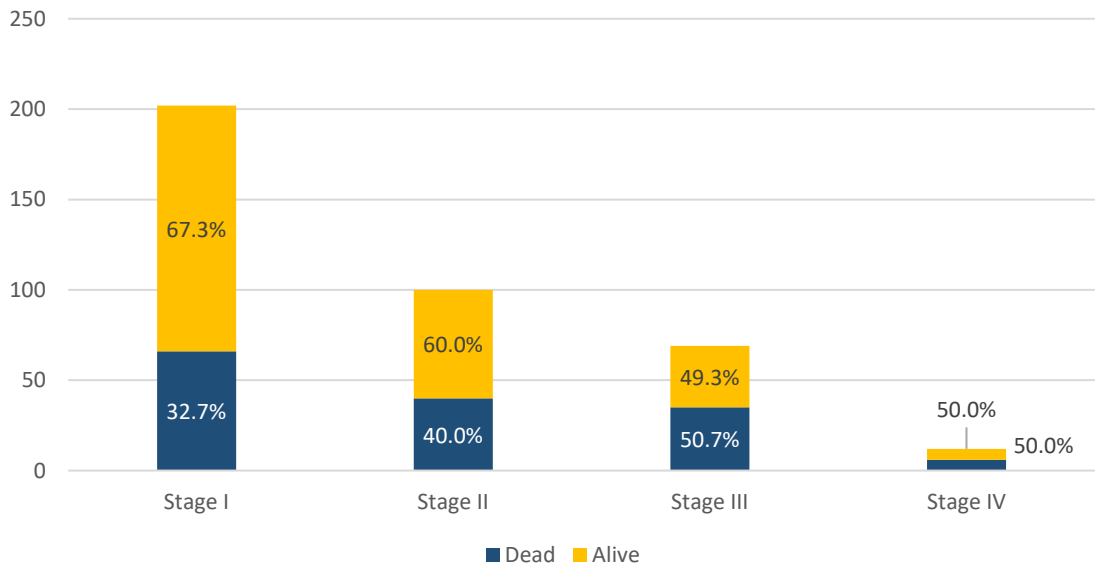

Figure 3. Frequency of tumor stage and percentage of vital status

The percentage of "alive and "dead" status per tumor stage is highlighted in Figure 3. The percentage of reported "alive" status declines in percentage compared to "dead" status relative to the tumor stage. The status met on the last tumor stages. Examining the data, majority of the cases under the stages I and II were reported "alive".

Random Forest Model

Pre-considered variables where examined to be included in the model. The output variable, tumor stage has 4 levels (Stages I, II, III and IV). Primary diagnosis was re-encoded as Adenocarcinoma, Squamous cell carcinoma, or other types. Eight factors (gender, birth year, age at diagnosis, vital status, primary diagnosis, tissue or organ of origin, average daily cigarette consumption, and number of years smoked/smoking) were considered as inputs and subjected to

the analyses in classifying lung cancer stage. Preparation of the data in building a random forest includes data partitioning.

The model's actual accuracy rate is 46.87% while the Kappa coefficient of -0.03 points out that the default classification model with 500 trees is prone to produce misclassifications. Hence, different number of trees were tried and 100 trees have been considered for its higher prediction.

Table 1 shows that only 39 of the 79 cases' tumor stage were classified correctly and they were 40 misclassifications.

Table 1. Random forest model prediction on evaluation set confusion matrix

		ACTUAL				TOTAL
		Stage I	Stage II	Stage III	Stage IV	
PREDICTED	Stage I	33	9	11	0	53
	Stage II	13	6	1	0	20
	Stage III	4	2	0	0	6
	Stage IV	0	0	0	0	0
	Total	50	17	12	0	79

Table 2 shows the statistics by stage as computed from the confusion matrix. Random forest model has a moderate sensitivity on classifying Stage I cases, but weakly sensitive for all the other classes. Its specificity in stage II and III cases are good yet not so specific in determining other stages. However, looking at the confusion matrix, the cases were extremely unbalanced and the random selection of observations as part of the evaluation resulted for some classes to have zero observations thus the high specificity on this classes were nullified. The balanced accuracy is a more appropriate accuracy measure to be observed when dealing with imbalanced classes. In this case, accuracy on predicting Stage II is the highest.

Table 2. Random forest model statistics by stage.

STATISTICS	STAGE I	STAGE II	STAGE III	STAGE IV
Sensitivity	0.6600	0.35294	0.00000	NA
Specificity	0.3103	0.77419	0.91045	1
Balanced Accuracy	0.4852	0.5635	0.45522	NA

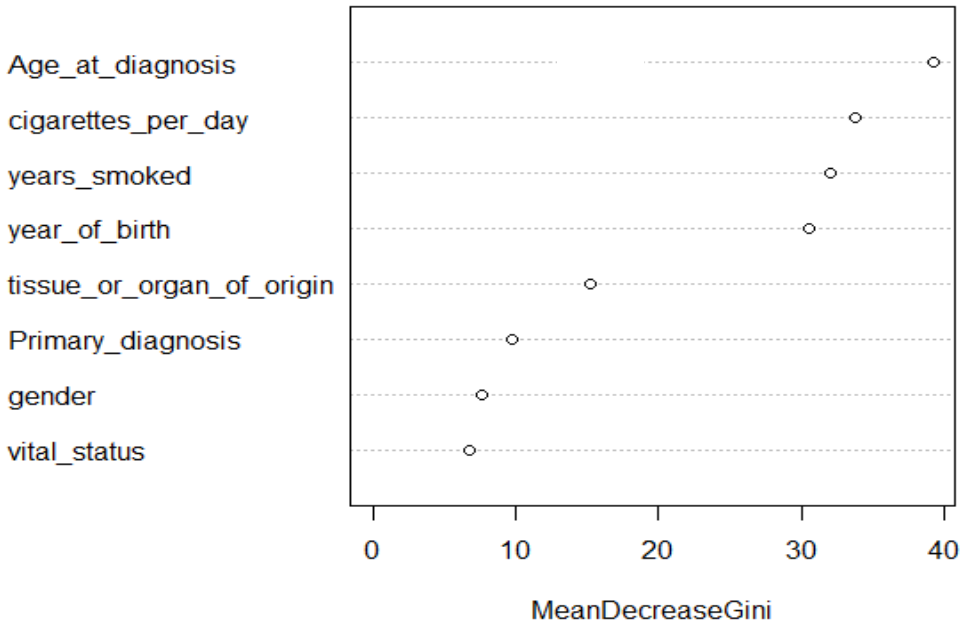

Figure 4. Mean GINI decrease for each variable

Predictor variables importance was graphed in Figure 4 that highlighted age at diagnosis as the most important factor in determining lung cancer stage in terms of the average GINI decrease. Factors that has to do with the patient's exposure and consumption of cigarettes were among the top 3 factors, mainly because exposure to cigarette smoke is the leading cause of lung cancer (Alberg et al., 2016). Gender and vital status has the least effect on the model's prediction.

The mean GINI decrease per variable and the number of times each was used in the final Random Forest model is shown in Table 3. The age at diagnosis is the most reoccurring factor as it was used 2,031 times.

Table 3. Variable importance in the random forest model.

FACTORS	MEAN GINI DECREASE	NUMBER OF TIMES USED IN THE MODEL
Age at diagnosis	39.23	2,031
Cigarettes smoked per day	33.83	1,887
Years smoked	32.08	1,816
Year of birth	30.60	1,792
Tissue/organ of origin	15.18	871
Primary diagnosis	9.73	655
Gender	7.66	529
Vital status	6.78	533

Artificial Neural Network

The assessment of the first model with 1 hidden layer was first done by predicting the tumor stages of the cases included in the training data. Results show that 76.69% of the stages in training data set were correctly identified by the model. The optimal number of neurons in the hidden layer is said to be any number between the number of neurons in the input layer (17) and the number of neurons in the output layer (4). The results are the same for the models with 4 to 17 neurons as hidden layer. Accuracy obtained from confusion matrix did not change as well and stayed moderately good.

The error given by the cross entropy function shows that the invariant in the error means the neural network no longer learns from proceeding iterations and that the weights already saturated which is a problem as it slows the networks learning from new data afterwards. Thus, network regularizations was applied yet no observable performance improvement was gained.

As all the candidate models performed well in terms of prediction accuracy, and produced similar results, selecting the best one would be on the aspect that they vary, thus, the model with least number of neurons on the hidden layer was selected. The final artificial neural network model for lung tumor stage classification was a 17-1-4 network.

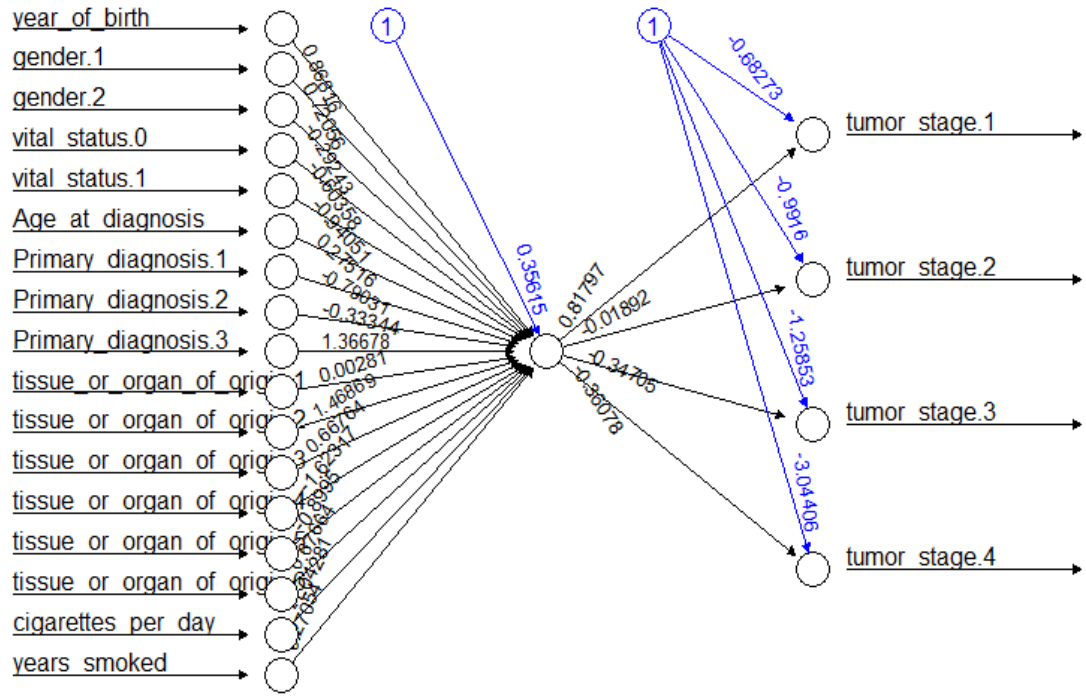

Figure 5. Final 17-1-4 Artificial Neural Network for lung tumor stage

The final ANN model as illustrated in Figure 5 is done through 1000 iterations. Softmax was used as activation function and was made up of 22 neurons, 17 neurons in the input layer, 1 in hidden layer and 4 in the output layer. The input process for the dummy variables will either be 0 or 1 while it requires standardization or normalizations of continuous inputs before feeding it to the model. Each neurons in the input layer is connected by an arrow to the neurons in the hidden layers

with corresponding weights computed through iterations on training the model, back-propagated to the network from each iterations that adjusted the weights. Values attached to the blue arrows are biases.

The output layer, composed of four tumor stages, will be returning probabilities being classified as each tumor stage. This was made possible by the softmax activation function with values from 0 to 1. The probabilities will then be analyzed compared to each other the one with the highest probability will be taken as the final output.

The final ANN model prediction on evaluation data was shown by a confusion matrix below (Table 4).

Table 4. Confusion matrix of ANN model's prediction on testing data.

		ACTUAL		TOTAL
		Negative	Positive	
PREDICTED	Negative	180	36	216
	Positive	36	36	72
	Total	216	72	288

Artificial neural network model's sensitivity and specificity were not obtained from classes as it treated multiclass factors differently, output was in vectors of probabilities per class that was further transformed into 1's and 0's to build the confusion matrix. Alternatively, the artificial neural network has a high specificity of 83.33%, fair sensitivity and balanced accuracy obtained through confusion matrix method (see Table 5).

Table 5. Neural network model prediction statistics.

STATISTICS	VALUE
Sensitivity	0.5000
Specificity	0.8333
Balanced Accuracy	0.6667

Model Comparison

To come up with the best lung tumor stage classifier, comparison of the models was done using accuracy metrics obtained from confusion matrix method.

In multiclass classification and in using unbalanced classes, the accuracy alone cannot provide the complete picture of how models perform hence, Cohen's Kappa statistic was also used.

Table 6 shows that the good accuracy of the neural network model on training and evaluation set with a fair Cohen's Kappa statistic made it outperform the Random Forest model.

Table 6. Model comparison in terms of accuracy.

MODEL	ACCURACY		KAPPA	
	Training	Evaluation	Training	Evaluation
Random Forest (100 trees)	1	0.49	1	0.006
Neural Network (17-1-4)	0.7669	0.75	0.3783	0.3333

6. Recommendations

The results of both models may be accounted to the scarcity of the number of cases for both training and evaluation due to prevalence of unreported observations, small number of predictors, and lack of highly relevant lung cancer stage determining variables.

The use of complete and structured big clinical data is highly recommended as good data results to excellent predictions. A balanced data is also preferable when the output variable has a multinomial classes as evaluation metrics such as total accuracy, sensitivity and specificity are dependent on how good and balanced the data is. The researcher also observed that in the premise of studying medical subjects, precise measures are desirable factors to be considered.

Moreover, it is indorsed to add more informative/subject-correlated variables such as tumor grade/tumor size/tumor load and BMI, as well as the use of local data and consideration of other machine learning methods that will improve lung cancer stage classification. The multiplicity of available machine learning methods that can handle classification offer endless ways to improve lung cancer stage classification.

LITERATURES CITED

Alberg, A. J., Brock, M. V., & Samet, J. M. (2016). Epidemiology of Lung Cancer. Murray and Nadels Textbook of Respiratory Medicine. doi:10.1016/b978-1-4557-3383-5.00052-x

Chen, J. H., & Asch, S. M. (2018). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. New England Journal of Medicine, 376(26), 2507-2509. doi:10.1056/nejmp1702071

Eldridge, L., & Hughes, G. (2019, February 27). What's the Survival Rate for Each Stage of Lung Cancer? Retrieved March 07, 2019, from https://www.verywellhealth.com/lung-cancer-survival-rates-by-type-and-stage-2249401

Leray, P., & Gallinari, P. (1999). Feature Selection With Neural Networks. Behaviormetrika,. 26(1), 145-166. doi:10.2333/bhmk.26.145

Lu C, Onn A, Vaporciyan AA, et al. (2010). "Chapter 78: Cancer of the Lung". Holland-Frei Cancer Medicine (8th ed.). People's Medical Publishing House. ISBN 978-1-60795-0141

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine, 375(13), 1216-1219. doi:10.1056/nejmp1606181

Wolpert, D. H. (1996). The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation, 8(7), 1341-1390. doi:10.1162/neco.1996.8.7.1341

World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 5.1. ISBN 978-92-832-0429-9.