Feature Selection of Non-Small Cell Lung Cancer Nodules

Axel Masquelin
The University of Vermont
axel.masquelin@uvm.edu

Andrea Elhajj
The University of Vermont
andrea.elhajj@uvm.edu

Samantha Connolly
The University of Vermont
samantha.connolly@uvm.edu

Thayer Alshaabi
The University of Vermont
thayer.alshaabi@uvm.edu

ABSTRACT

Through the use of data-characterization algorithms, the field of radiomics provides a wealth of mineable, quantitative data that is extracted from medical images. The data from these radiomic features unveil disease characteristics that cannot be seen by simple visual inspection of these medical scans. Feature selection algorithms may be used to identify the features with the greatest relevance and predictive power to ultimately create a model for disease diagnosis and outcome prediction. This research investigates a data-driven feature selection method using symbolic genetic programming in tandem with support vector machines (SVM) and random forests (RF) to build a predictive model for non-small cell lung cancer diagnosis. This study compared the effectiveness of using genetic programming alone, genetic programming as a feature selector followed by a SVM or RF, SVM alone, and RF alone to create to build a predictive model for non-small cell lung cancer diagnosis from radiomics data. Genetic programming as a feature selector followed by SVM or RF performed similarly well as SVM or RF alone, but many fewer features were used. This suggests that selecting features using genetic programming before an SVM or RF leads to more generalized results.

ACM Reference Format:
Axel Masquelin, Andrea Elhajj, Samantha Connolly, and Thayer Alshaabi. 2019. Feature Selection of Non-Small Cell Lung Cancer Nodules. In *Proceedings of the Genetic and Evolutionary Computation Conference 2019 (GECCO '19).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Non-small cell lung cancer (NSCLC) is a heterogeneous disease with a variety of bio-markers, some reflecting a point mutation, while others are associated with a range of protein expressions. Though the most widely associated cause is smoking, factors such as radon, asbestos, mineral and metal dust exposure, air pollution, radiation treatment to the chest or breast area, HIV/AIDS, as well as genetic predisposition, increase the likelihood of incidence. Immunosenescence (the gradual deterioration of the immune system brought on by age) contributes to a diagnosis rate of 50% in patients 70 years of age and older [4].

According to the severity of the disease, NSCLC is categorized on a scale of I to IV. Stage I represents a localized tumor that is characterized by a size of $< 5cm$ and has yet to spread to the lymph nodes. Stage II NSCLC is $5-7cm$, localized, and has possibly spread to the lymph nodes adjacent to the affected lung. Stage III is $> 7cm$, localized, and spread to a major structure within the chest. Lastly, Stage IV is in both lungs or has metastasized to another part of the body [1]. NSCLC metastases have the potential to spread to nearby lymph nodes, as well as through the bloodstream to the brain, bones, and adrenal glands.

Initial stages of NSCLC are largely asymptomatic. Squamous cell carcinoma is most commonly detected in a patient once it reaches a size of $30mm$, at which point symptoms finally manifest [3]. This takes approximately 8 years, and thus the risk for metastasis by this time is substantially high. In fact, 80% of NSCLC diagnoses are found to be at stage III or IV, and are not suitable candidates for surgery [3]. Additionally, a recent population-based analysis found that lung cancer patients have the highest malignancy-associated suicide rate in the United States [12].

Over the past decade, the field of medical imaging has both improved and expanded tools for pattern recognition [7]. In tandem with these advancements, the traditional practice of simple visual inspection and interpretation has evolved into the practice of radiomics. Radiomics is the emerging field involving the process of converting medical images into mineable, quantitative data for the betterment of clinical decision-making in the care of patients with cancer. This quantitative data describes features of the imaged tumor, including location, intensity, shape, size or volume, and texture that cannot be gleaned from laboratory test results or other noninvasive assessments [7]. The tumor phenotype information is then combined with other relevant patient data to create models that seek to improve the accuracy of both disease diagnoses and outcome prediction [7].

Feature selection (FS), also known as variable or attribute selection, is the selection of a subset of the most relevant attributes from the greater set in order to accurately model a particular problem. FS and dimensionality reduction are two methods that aim to remove the irrelevant and redundant attributes do not improve predictive power of a model. However, dimensionality reduction does this through the creation of new attribute combinations, while feature selection maintains or removes data attributes without changing them [8].

The overarching objective of FS is to identify the features that provide as good or better accuracy while utilizing less data, thereby aiding in the creation of a faster and more cost-effective model [8]. Fewer attributes in a predictor model are a desirable outcome of FS as

GECCO '19, July 15–19, 2018, Prague, Czech Republic

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

this reduces redundancy and complexity [8]. Although FS algorithms have many applications in data mining and machine learning, it is especially useful in prediction of disease and many other medical applications. Disease diagnosis is an expensive process in terms of both testing time and cost [5]. FS methods play a significant role in the diagnosis and prediction of chronic diseases [9]. Recent research has even utilized FS combined with support vector machines (SVM) for breast cancer diagnosis [11].

Nevertheless, radiomics data contains common information that may be scattered across multiple features. As a result, most if not all data-driven predictive models are vulnerable to what is commonly known as overfitting in which a given model is likely to perform well on the current set of samples but will not be able to generalize well on unseen data samples. In this work, we investigate a data-driven feature selection method using symbolic genetic programming combined with SVM and Random Forests to build a predictive model for cancer diagnosis.

2 METHODS

2.1 Dataset

This study utilizes data collected through the National Lung Screening Trial (NLST) [2]. The NLST was a randomized controlled trial designed to determine if NSCLC mortality in high-risk individuals could be reduced by low-dose helical computed tomography (LDHCT) screening compared to single-view chest radiography. The study, a joint effort between the National Cancer Institute's Division of Cancer Prevention (DCP) and the Division of Cancer Treatment and Diagnosis (DCTD), was conducted between August 2002 and April 2004. This trial involved 54,000 participants with eligibility based on the following criteria: 55-74 years-old, 30 or greater pack-years (defined as the product of the number of packs of cigarette smoked per year and the number of years of smoking), and a maximum of 15 years since quitting for previous smokers [2].

Individuals fitting these specifications and not in conflict with the exclusion criteria were randomized to one of the two study groups in equal proportions, with one group receiving the LDHCT scan, and the other receiving the single-view chest radiography. The participants were offered an examination at one-year intervals over the study, with the initial exam performed either immediately at the time of randomization or shortly thereafter. It is also worth noting that the patient withdrawal rate was exceptionally low, and 96.6% of these participants were tracked until death or until the end of the study [2].

2.2 Data Preprocessing

The results from the LDHCT screening study arm were utilized for this feature selection investigation. The LDHCT data was further trimmed to only include the information of tumors that were between 4mm to 20mm in size, resulting in a dataset with 3,702 patients and 273 features. These relevant features include factors such as entropy, maximum intensity, ventilation heterogeneity, uniformity, kurtosis.

The data was then cleaned by dropping out the feature columns with over 90% missing data as well as removing the patients with over 50% of their data fields missing. Next, any patients missing a data entry in the diagnosis target attribute were also dropped from the dataset. Patients with 50% or less missing information had their field entries filled with the mean of the attribute from this cleaned patient dataset. This processing resulted in a transformation from 273 features to 268 features and 3,702 patients to 3,534 patients.

Moreover, we investigated different techniques for up-sampling and down-sampling the number of cases in our dataset as a way to avoid the trivial solution of predicting one class all the time and still getting good accuracy due to the class imbalance issue in our dataset. We found that down-sampling the number of negative cases to match up with the number of positive cases in our dataset worked the best. Ultimately, we ended up with a total of 672 patients divided into 336 positive cases and 336 negative cases.

2.3 Trimmed Dataset

Additionally, we used a feature selection technique called Recursive Feature Elimination to rank features and filter out uncorrelated features based on their direct contribution to the overall prediction score. Recursive Feature Elimination is based on the basic idea of repeatedly creating a predictive model (e.g. SVM or random forest) to evaluate a given set of features in the dataset. We start with all of the features in the dataset. In every iteration, we take out a feature from the dataset and re-evaluate the model to check the impact of that given feature on the prediction accuracy. We repeat that process till all features in the dataset are tested. Utilizing this technique we identified a set of 57 important features in our dataset as shown in Figure 2.

Furthermore, previously identified features using lasso regression from Dr. Charles M. Kinsey and Dr. Raul San Jose Estepar's work were also included to generate a trimmed radiomics dataset of 77 features. The radiomics features describe statistical representation of the tumor at various levels, such as $0i$ (nodule region only), $15c$ (15 mm from centroid of lesion), and $10b$ (10 mm from boundary of lesion).

2.4 Experiment

In this work, we utilize a widely used evolutionary computation framework known as the Distributed Evolutionary Algorithms in Python `DEAP` [6] along with `Scikit-learn` an open-source library for machine learning algorithms [10].

The operators we used were: and, or, not, add, subtract, multiply, divide, exponent, sin, cos, log, greater than, less than, equal to, and if else. The division operator returned 1 if the denominator was 0. The ifelse operator returned the negated first feature if the first feature was less than the second feature. If the first feature was not less than the second feature, the first feature was returned as shown in Algorithm 1. The terminals we used were: ephemeral constants, False, True, and the radiomics data features.

Algorithm 1: ifelse Operator

Data: Feature 1 (f1), Feature 2 (f2)
if $f1 < f2$ **then**
  return $-f1$
else
  return $f1$

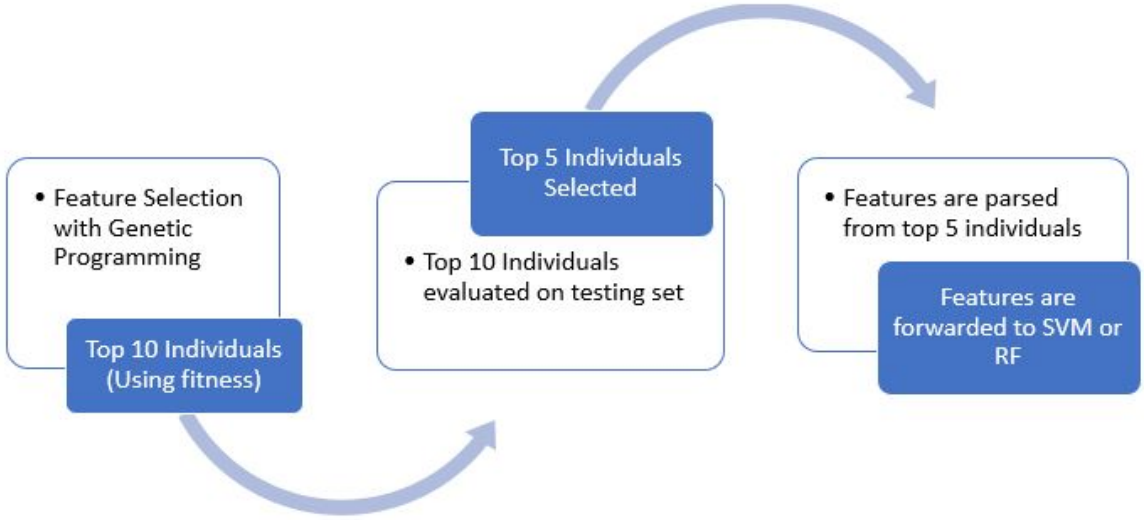


Figure 1: A pipeline approach where we use genetic programming as a feature selector followed by a classifier (e.g. SVM or Random Forest) to evaluate the accuracy of the features selected.

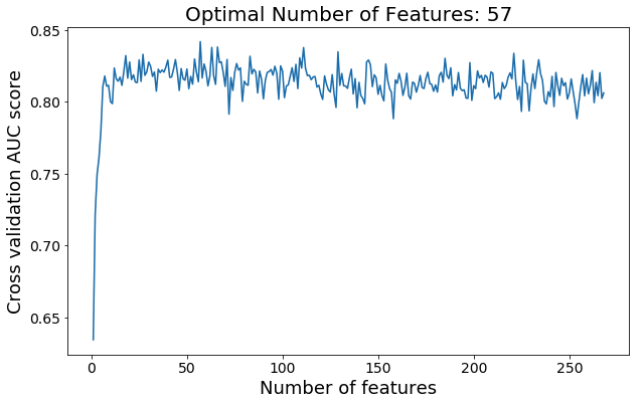


Figure 2: Feature selection using recursive feature elimination to select the optimal subset of features that yields the best performance.

We used a similar fitness function as the one from [13]. The fitness function that we used more strongly weights incorrect predictions than correct predictions. This fitness function is:

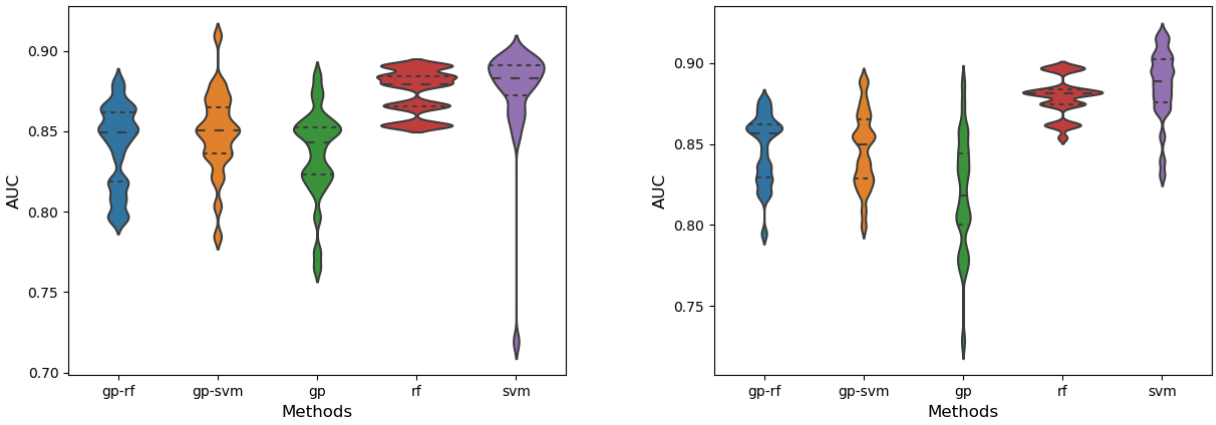
$$f_i = CorrectRate_i \times (1 - FalseRate_i)^2 \quad (1)$$

Additionally, two main tools were used to control bloat: double tournament selection and a static depth limit of 20. Double tournament selection was used for parent selection with a tournament size of 7 individuals, and a parsimony size of 1.8. Only fitness is considered in the first round of tournament selection. The parsimony size of 1.8 means that there is a 90% chance of selecting individuals with the best fitness scores in the first round. Similarly, there is a 90% chance of selecting the smaller tree on the second round of tournament selection. In the crossover and mutation operators, the static limit decorator was added. This prevents new trees from having a depth greater than 20.

In order to determine the most appropriate mutation function and crossover along with their rates, we compared the average maximum fitness at generation 250 for various combinations of parameters. We tested the mutEphemeral, mutUniform, and mutNodeReplacement mutation functions with mutation rates of 0.2, 0.4, and 0.6. We tested the cxOnePoint crossover function with crossover rates of 0.2, 0.4, and 0.6. We determined that the most effective combination of these parameters was the mutNodeReplacement mutation function with a mutation rate of 0.4 and a crossover rate of 0.6. A full list of the GP parameters used in our experiments is shown in Table 1.

3 RESULTS

A total of 25 trials were performed with each approach (5 repetition, 5 cross-validation) using the methodology describes previously: genetic programing (GP), Support Vector Machine (SVM), and Random Forest (RF). After 250 generations, the genetic programming approach was terminated and evaluated by examining an individual tree's ability to predict cases' categories as malignant tumors or benign nodules. For the pipeline approach GP was evaluated as a feature selection mechanism followed by a SVM or RF as classifiers. For this method, we obtained the top 10 most fit individuals identified through GP. We then evaluated these 10 individuals to find the best five individuals on a testing set. The top 5 individuals' features were then parsed and sent forward to either a SVM or RF to be evaluated, as seen in figure 1. This experiment was conducted on both the trimmed and original dataset in order to evaluate the stand-alone performance of genetic programming and gain insights on the effect decreasing features on genetic programming performance. Overall, GP as a stand alone feature selector and classifier proved to be significantly different from SVM and RF with p-values < 0.001. This trend can be similarly observed when utilizing the



(a) Trimmed Dataset **(b) Original Dataset**

Figure 3: Violin plot showing the AUC scores distribution across 5 repetitions for gp-rf, gp-svm, gp, rf, and svm with 5folds cross-validation on both the original and trimmed datasets.

Table 1: GP Parameter Optimization; We ran several experiments to find the best parameters suited for our dataset. We used the values that yielded the best performance.

Parameter	Value
Dataset split	[75%] Training, [25%] Testing
Cross-validation	5 folds
Number of runs	5
Max Generations	250
RandomNum Generator Seed	2018
Initialization Function	genHalfAndHalf
Population Size	500
Min Initial Tree Size	5
Max Initial Tree Size	20
Mutation Function	mutNodeReplacement
Max Mutation Tree Depth	20
Mutation Rate	.4
Crossover Function	cxOnePoint
Max Crossover Tree Depth	20
Crossover Rate	.6
Selection Function	selDoubleTournament
Tournament fitness size	7
Tournament parsimony size	1.8

trimmed dataset, as GP is again significantly different from SVM and RF with a p-value < 0.001.

Using the pipeline methodology, one can see from Figure 3 that in GP-RF and GP-SVM performed significantly better than a stand-alone GP when utilizing the original dataset. However, this trend does not hold true on the trimmed dataset where a stand-alone GP performed similarly to that of a GP-SVM and GP-RF, with p-values of 0.54 and 0.09, respectively, at a 95% confidence level. Comparing the baseline SVM and RF to the pipeline GP-SVM and GP-RF shows that the pipeline approach is significantly different from baseline, p-values of 0.006 and 0.00014, respectively for both the trimmed and original dataset.

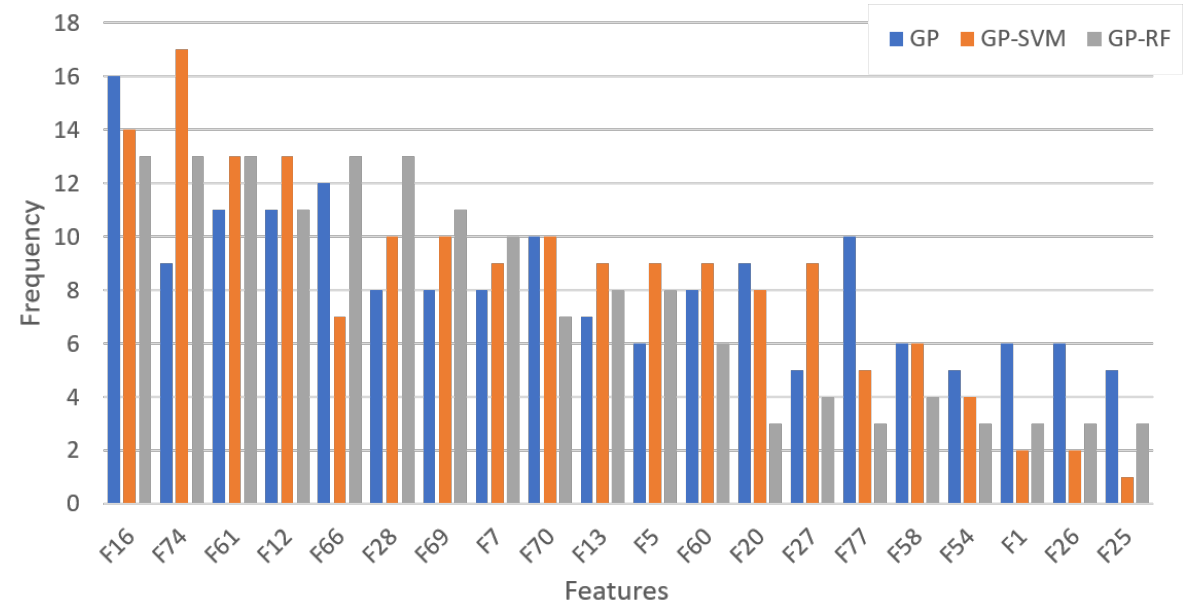
In the trimmed data set, the twenty most commonly selected features can be seen in Figure 4, and include features such as: variance of the nodule, mean deviation 15 mm from centroid of the nodule, auto correlation 10mm from boundary of nodule, and the median intensity of the nodule. These specific features were similarly found within the original dataset, as shown in Table 2.

4 DISCUSSION

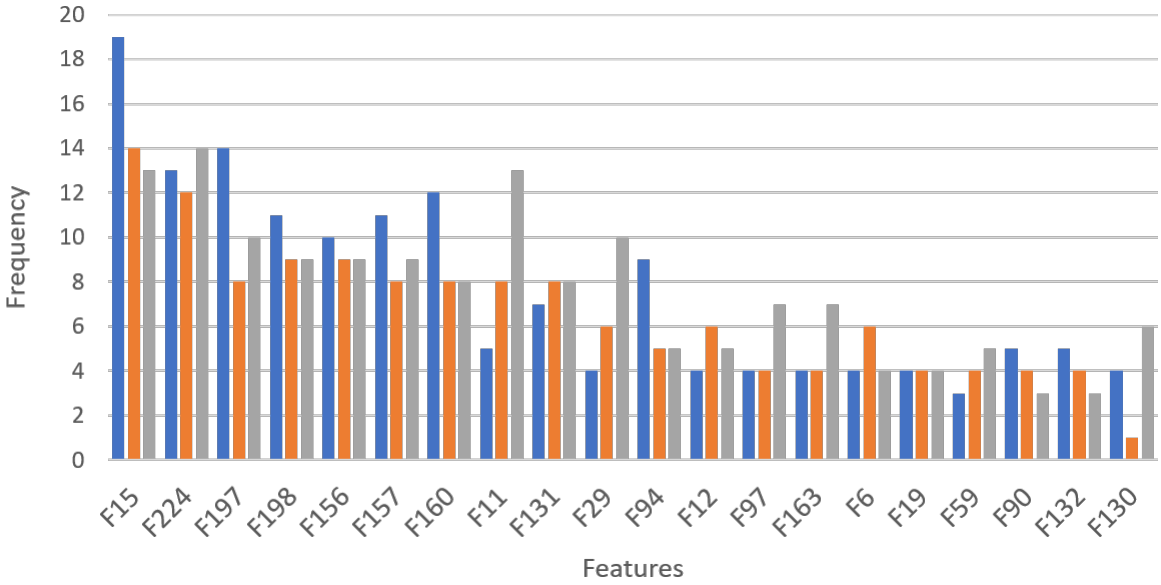
We were surprised to see that using GP for feature selection followed by either a random forest or a SVM did not yield any significant improvement in terms of AUC score when compared to either a random forest or a SVM alone. We had expected to see that feature selection using genetic programming would improve the overall performance.

However, the AUC patterns were very similar between the original and trimmed datasets. This suggests that trimming the dataset did not strongly affect the accuracy of the solutions formed by these methods. For genetic programming alone, trimming the dataset seemed to have reduced the variation of its AUC; for SVM alone, trimming the dataset seemed to have increased the variation of its AUC.

More importantly, the predictive model utilized all features given in the dataset using the standalone RF or SVM approach. However, in the pipeline approach the classifier was only looking at 10 up to 15 features on average while maintain similar performance. This suggests that our pipeline approach can generalize better by filtering out unnecessary features that do not contribute much to the overall accuracy. This will also reduce the likelihood of overfitting the current set of cases in our dataset.



(a) Trimmed Dataset



(b) Original Dataset

Figure 4: frequency of feature selection for GP, GP-SVM, and GP-RF based on the (a) trimmed dataset, and (b) original dataset.

Furthermore, Figure 4 shows a histogram of the top 20 selected features by all tested models on the trimmed and original datasets separately. Interestingly, the features selected by GP differed greatly between the trimmed and the original datasets (see Table 2). We were surprised to see such a strong difference between the features chosen, but suspect it could be due to features interacting with each other in a higher dimensional space.

For the trimmed dataset, three of the most common twenty features were various measurements of the compactness (see Table 2). This suggests that compactness could be a very important feature when determining whether or not a nodule is cancerous. For the original dataset, four of the most common twenty features were the percent of lung with Low Attenuation Areas with density values below either -856 or -910 HU ("laa856perc" or "laa910perc"). Another four of the most common twnety features were the Long Run High Gray-Level Emphasis (lrhgle). (Table 2). This suggests that low attenuation areas and the long run high gray-level emphasis could be very important features when determining whether a

Table 2: Top 20 selected features from all tested models on the trimmed and original dataset.

TRIMMED DATASET				ORIGINAL DATASET			
ID	Feature	ID	Feature	ID	Feature	ID	Feature
F1	xNorm	F28	compactness20i	F6	energy0i	F130	lrlgle10b
F5	voxelcount0i	F54	srlgle0i	F11	medianintensity0i	F131	lrhgle10b
F7	energy0i	F58	ventilationheterogeneity10b	F12	range0i	F132	extrudedsurfacearea10b
F12	medianintensity0i	F60	compactness110b	F15	standarddeviation0i	F156	laa910perc15b
F13	range0i	F61	autocorrelation10b	F19	variance0i	F157	laa856perc15b
F16	variance0i	F66	ventilationheterogeneity15b	F29	compactness20i	F160	surfaceareamm215b
F20	uniformity0i	F69	sphericaldisproportion15b	F59	lglre0i	F163	compactness215b
F25	surfaceareamm20i	F70	lre15b	F90	laa856perc10b	F197	lrlgle15b
F26	surfacevolumeratio0i	F74	meandeviation15c	F94	surfacevolumeratio10b	F198	lrhgle15b
F27	compactness10i	F77	extrudedsurfacearea15c	F97	maximum3ddiameter10b	F224	laa856perc15c

nodule is cancerous or benign. Although laa and lrhgle were more often selected, compactness was maintained as important factors in nodule categorization. In both the trimmed and the original dataset, the energy of the nodule, the median intensity of the nodule, the range of the nodule, and the variance of the nodule were in the top seven features (Table 2). Interestingly enough, factors such as long run low grey-level emphasis seemed to become less important in the trimmed dataset; differences like this demonstrate why it is important to analyze data using different methods.

5 CONCLUSIONS

Genetic programming as a stand alone feature extractor and classifier proved to do generally worse than SVM, RF, and the pipeline combinations as measured by AUC score. The pipeline approaches performed similarly well to the SVM and RF alone, but did so with many fewer features, suggesting that using genetic programming for feature selection followed by either an SVM or RF could be a more generalized approach. As determined by genetic programming as a feature selector, the most important features when predicting whether a nodule is cancerous or benign are compactness, low attenuation areas with density values below either -856 or -910 HU, long run high gray-level emphasis, the energy of the nodule, the median intensity of the nodule, the range of the nodule, and the variance of the nodule.

REFERENCES

[1] [n. d.]. Non-Small Cell Lung Cancer Treatment. ([n. d.]). https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq#section/_134

[2] [n. d.]. Trial Summary - Learn - NLST - The Cancer Data Access System. ([n. d.]). https://biometry.nci.nih.gov/cdas/learn/nlst/trial-summary/

[3] S S Birring. 2005. Symptoms and the early diagnosis of lung cancer. *Thorax* 60, 4 (Jan 2005), 268âĂŞ269. https://doi.org/10.1136/thx.2004.032698

[4] Francesca Casaluce, Assunta Sgambato, Paolo Maione, Alessia Spagnuolo, and Cesare Gridelli. 2018. Lung cancer, elderly and immune checkpoint inhibitors. *Journal of Thoracic Disease* 10, S13 (2018). https://doi.org/10.21037/jtd.2018.05.90

[5] Tsang-Hsiang Cheng, Chih-Ping Wei, and V.s. Tseng. 2006. Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches. *19th IEEE Symposium on Computer-Based Medical Systems (CBMS06)* (2006). https://doi.org/10.1109/cbms.2006.87

[6] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. 2012. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research* 13 (jul 2012), 2171–2175.

[7] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. 2016. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 278, 2 (2016), 563–577. https://doi.org/10.1148/radiol.2015151169 PMID: 26579733.

[8] Isabelle Guyon. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3 (2003), 1157–1182.

[9] Divya Jain and Vijendra Singh. 2018. Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Informatics Journal* 19, 3 (2018), 179âĂŞ189. https://doi.org/10.1016/j.eij.2018.03.002

[10] F. 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. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[11] Santi Wulan Purnami, S.p. Rahayu, and Abdullah Embong. 2008. Feature selection and classification of breast cancer diagnosis based on support vector machines. *2008 International Symposium on Information Technology* (2008). https://doi.org/10.1109/itsim.2008.4631603

[12] Mohamed Rahouma, Mohamed Kamel, Ahmed Abouarab, Ihab Eldessouki, Abu Nasar, Sebron Harrison, Benjamin Lee, Eugene Shostak, John Morris, Brendon Stiles, and et al. 2018. Lung cancer patients have the highest malignancy-associated suicide rate in USA: a population-based analysis. *ecancermedicalscience* 12 (2018). https://doi.org/10.3332/ecancer.2018.859

[13] Shelly Xiaonan Wu and Wolfgang Banzhaf. 2011. Rethinking Multilevel Selection in Genetic Programming. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation (GECCO '11)*. ACM, New York, NY, USA, 1403–1410. https://doi.org/10.1145/2001576.2001765