Lung Cancer Detection Using Genomic Profiles and a Multilayer Perceptron Classifier

Julian Romero
Department of Computer Science
Rutgers University
New Brunswick, NJ 08901
jvr40@rutgers.edu

December 19, 2020

Abstract

This paper focuses on lung cancer detection using a mulilayer perceptron. There have been many other models that focus on cancer detection from image recognition in CT scans using deep residual learning to using linear classifier models with cancer patient's medical records. However, none that we know of take advantage of using genomic profiles to predict lung cancer among a dataset of real cancer patients. The model in this paper uses 200 cancer patients with a model that uses 568 input nodes based on distinct mutated genes with two hidden layers using a rectifier activation function and one discrete output layer that uses a logistic regression model. The model had a loss of 0.6781 for its first epoch and uses a logarithmic loss function. The model has a prediction accuracy of 78 percent, precision of 74 percent , and recall of 91 percent. The model shows promise in improving diagnoses of cancer before late stages.

1 Introduction

In the medical field, one of the most important steps a physician will take is a diagnosis. In the majority of cases, a diagnosis is not very quick when it comes to detecting a certain disease until symptoms appear. Cancer being one of the most prevalent diseases of our time is often always diagnosed late. For example, in figure 1 in the appendix section, lung cancer is diagnosed mostly at stage 3 and 4 with a percent diagnosis being at 65 percent compared to all lung cancer stages. In the related works, different methods are used to improve the accuracy and classification of a diagnosis, like deep residual learning, reducing misclassification, and a deep CNN. The ability to predict cancer before later stages occur is critical in saving patients from death. It is known that there are certain gene mutation signatures that are an indicator that cancer is present in certain cells. With this in mind, this paper aims to find patterns associated with these gene mutation signatures by analyzing different genetic sequences using Genomic Profiles and a Multilayer Perceptron in a dataset to predict lung cancer. This model will focus on predicting lung cancer and will consist of patients who have lung cancer and patients who have another form of cancer. From these gene mutations,Genomic Profiles and a Multilayer Perceptron model will be used to predict whether lung cancer is present in certain cells. Once the system is done training, testing on detecting lung cancer on another data set similar to the previous data set will occur. The testing is representative of patients in stage 1 cancer due to many mutated genes existing before later stages of cancer. From here, if a viable prediction accuracy above 70 percent is found this has the potential to be used to diagnose many patients with lung cancer before stage 3 and 4 cancer. We Will implement this as follows. The inputs will be the mutated gene sequences found from databases of cancer patients. The databases used will be from the national institute of cancers database called GDC. The outputs will be a discrete value of whether someone has lung cancer or not. The system will be encoded by trying to parse through different genetic sequences and using set mutagenic signatures as features. These signatures will be similar to the signatures found in figure 2. Many papers in the past have also focused on detection of cancer using deep learning. The following in related works addresses these.

2 Related Work

Detecting cancer with deep learning has the potential to save lives and provide a more efficient method of cancer detection. With a technique used by Bhatia [4], lung cancer was able to be detected from CT scans using deep residual learning. To achieve this, a pipeline of preprocessing techniques is delineated to highlight lung regions vulnerable to cancer and extract features using UNet and ResNet Models. This involves multiple classifiers which the feature set is fed into, using methods like XGBoost (Extreme Gradient Boosting) and Random Forest, and individual predictions to predict likelihood of a CT scan being cancerous. This outperformed previous methods, by an accuracy of 84 percent on LIDC-IRDI. In a paper by Shakeel [5], the improvement in the quality of lung image and diagnosis of lung cancer was addressed by reducing misclassification. This was done by removing noise from images and enhancing image quality, using improved profuse clustering technique (IPCT). By this, it resulted in a 98.42 percent accuracy rate with minimum classification error 0.038. In a study by Coudray [6], a deep CNN (inception v3) was trained on whole-slide images obtained from The Cancer Genome Atlas to automate classification to LUAD, LUSC, or normal lung tissue. This model was validated on independent datasets of frozen tissues, formalin-fixed paraffin-embedded tissues and biopsies. It was also trained to predict the ten most commonly mutated genes in LUAD. The AUC (average area under the curve) was between 0.733 to 0.856 among the six most common genes (STK11, EGFR, FAT1, SETBP1, KRAS and TP53). The distinction in this report will be using Genomic Profiles and a Multilayer Perceptron.

3 Main Contribution

The challenges we face is finding the right data that can be used to train a multilayer perceptron. As well as find the right model to find a significant accuracy and try to minimize loss when training. With these challenges our contributions that tackled these problems are as follows. Our paper uses a unique genetic profile dataset of 200 individual patients. We have made a multilayer perceptron with two hidden layers. Our model had a predicted accuracy of 78 percent for the tested model set. This is the only known work that uses genetic data to detect lung cancer that we know of. Our data focuses on mutated genes that are the most common among real patients from the national institute of cancer's GDC portal.

4 Experiment

4.1 Dataset Curation

The experiment's main design is as follows. We used 200 cancer patients from the GDC portal of the national institute of Cancer. A dataset of different patient's genetic profiles did not exist. So for us to have a workable dataset that can be used in our multilayer perceptron model we curated the dataset as follows. We first received data of a patient where we had a dataset of all the mutated genes known to cause cancer for that patient. Some patients shared common genes while others did not. We took the datasets of every individual patients genetic profiles and combined them into a unique dataset. We labeled each input discretely by labeling a "1" if a gene was present in that patient and a "0" otherwise. From this we were able to figure out that our input layer would have 568 nodes representing each distinct mutated gene. We labeled our output layer discretely with either "1" if the patient was a lung cancer patient, or "0" if the patient was another type of cancer patient. From the 200 cancer patients, 100 had lung cancer while the other 100 had some other form of cancer. The dataset can be seen in figure 3 of the appendix section.

4.2 Training model description

Once our dataset was made we trained our dataset using a multilayer perceptron model.Our model had an input layer of 568 Nodes based on distinct mutated genes, two hidden layers with 100 nodes each. The hidden layers used a rectifier activation function. Our weights were randomly set to low values using the dense function in keras library. Our output used a logistic regression model as it gives the output as a probability which we needed when classifying. The weights were updated using a stochastic gradient descent algorithm. Our loss function was determined using a logarithmic loss function.The example model can be seen in figure 4 of the appendix section. The model was trained with 200 epochs and batch sizes of 10. Our training data comprised 160 patients out of the 200 cancer patients.The loss at epoch 1 of our training model was 0.6871. Our training data showed that our loss had decreased with each epoch which is good for having an accurate result. As well as our accuracy in training increased with the increase in epochs. The training simulation can be seen in figure 5 of the appendix section.

4.3 Evaluation

In the model 10 percent of the 200 patients were used as the testing data. All predicted scores were found to be true if the score was above a 70 percent threshold. Once our training was done it was evaluated by comparing with our testing data. It was found to have an accuracy of 78 percent, a precision of 74 percent, and a recall of 91 percent. This shows the efficiency of the model. Our confusion matrix shows that we correctly predicted 20 patients with lung cancer based on genetic profiles. We predicted 11 patients do not have lung cancer and have a different type of cancer. However our system is not perfect as we had 7 false negatives and 2 false positives. The confusion matrix can be seen in figure 6 of the appendix section.

5 Conclusion and Broader Impact

Our model shows that a lung cancer detector is possible using genomic profiles. With an accuracy of 78 percent, it shows promise for other future models that want to predict cancer using genomic profiles. The model is just the beginning of what can be done with genomic profiles. This model showed evidence that there is a strong correlation with genes and cancer. Now that it is known that a classification model can work with genomic profiles next would be to introduce a noncancer patient dataset and really test if the model can predict lung cancer in patients with no cancer currently diagnosed. In the future to expand on this model one could incorporate medical records as features such as age, weight, smoker/ non smoker, etc. As well as expand on detecting any kind of cancer not just lung cancer by changing the model into a multiclass classification model.In regards to our model itself, one could speculate that overfitting of the dataset could of occurred. This could be fixed by acquiring a larger dataset of genomic profiles. However, this does come with its challenges as this dataset had to be made and it was extensive to do. Perhaps in the future with more available datasets of genomic profiles a more efficient model can be made. Hopefully our paper can offer more tools for physicians to use in diagnoses and contribute to the field of medicine.

6 References

[1] Alexander, J.A. Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro,D.S. Touretzky and T.K. Leen (eds.),Advances in Neural Information Processing Systems 7, pp. 609–616. Cambridge,MA: MIT Press.

[2] Bower, J.M. Beeman, D. (1995) The Book of GENESIS: Exploring Realistic Neural Models with the GEneral-NEural SImulation System.New York: TELOS/Springer–Verlag.

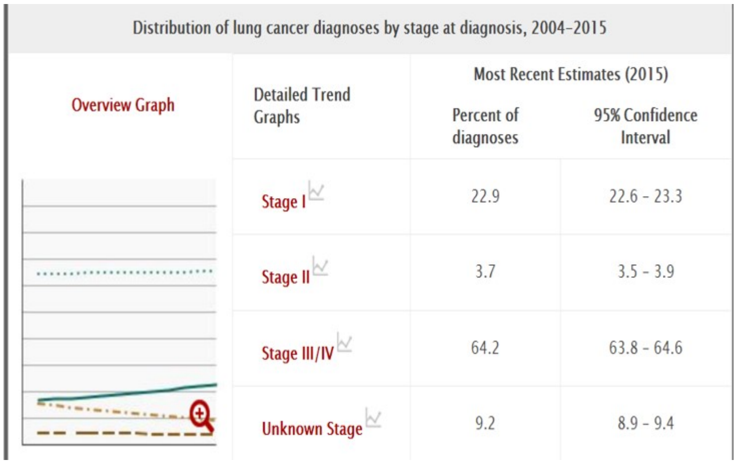
[3] Hasselmo, M.E., Schnell, E. Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3.Journal of Neuroscience15(7):5249-5262.

[4] Bhatia, S., Sinha,Y., Goel, L. (2019) Lung Cancer Detection: A DeepLearning Approach Department of Computer Science and Information Systems, BITS,

[5] Mohamed Shakeel, P., Burhanuddin, M.A, Ishak Desa, M. (2019) Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia Advanced Manufacturing Centre, Universiti Teknikal Malaysia Melaka, Malaysia

[6] Coudray, N., Ocampo, P.S., Sakellaropoulos, T. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. Nat Med 24, 1559–1567 (2018). https://doi.org/10.1038/s41591-018-0177-5

7 Appendix



Distribution of lung cancer diagnoses by stage at diagnosis, 2004–2015

Overview Graph	Detailed Trend Graphs	Most Recent Estimates (2015)	
		Percent of diagnoses	95% Confidence Interval
	Stage I	22.9	22.6 – 23.3
	Stage II	3.7	3.5 – 3.9
	Stage III/IV	64.2	63.8 – 64.6
	Unknown Stage	9.2	8.9 – 9.4

Figure 1: Diagnosis of Lung Cancer National Institute of Cancer.

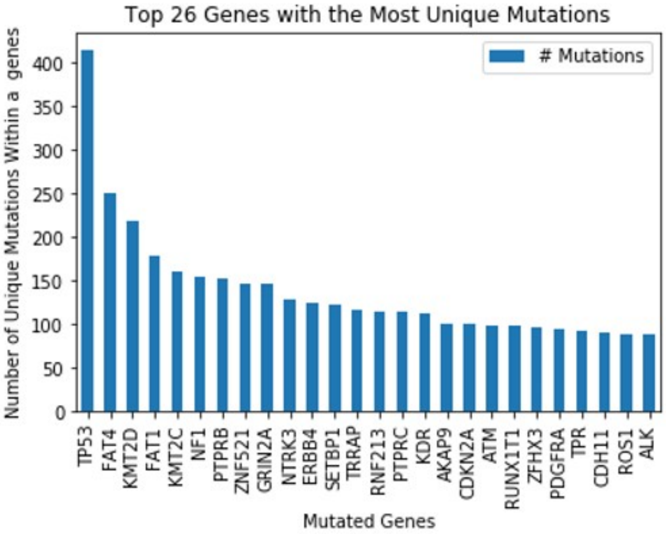
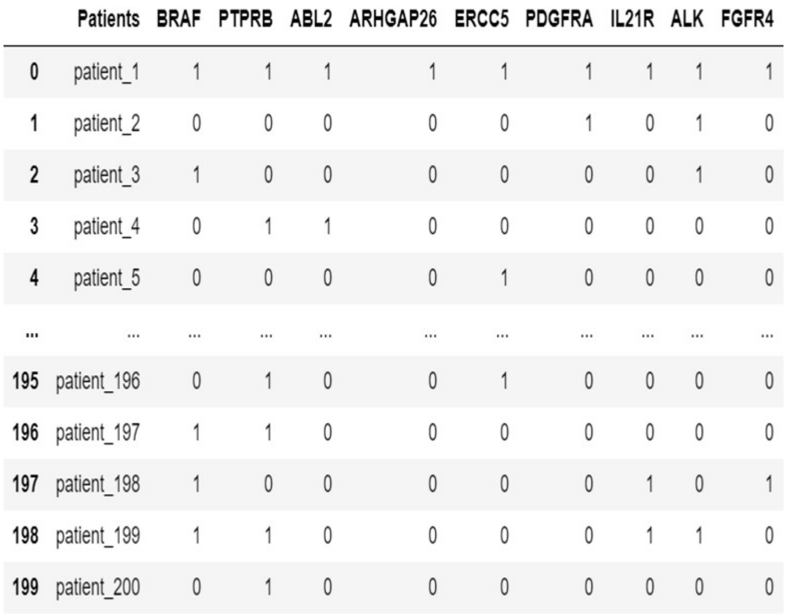


Figure 2: Top 26 mutated Genes in lung cancer patients



	Patients	BRAF	PTPRB	ABL2	ARHGAP26	ERCC5	PDGFRA	IL21R	ALK	FGFR4
0	patient_1	1	1	1	1	1	1	1	1	1
1	patient_2	0	0	0	0	0	1	0	1	0
2	patient_3	1	0	0	0	0	0	0	1	0
3	patient_4	0	1	1	0	0	0	0	0	0
4	patient_5	0	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
195	patient_196	0	1	0	0	1	0	0	0	0
196	patient_197	1	1	0	0	0	0	0	0	0
197	patient_198	1	0	0	0	0	0	1	0	1
198	patient_199	1	1	0	0	0	0	1	1	0
199	patient_200	0	1	0	0	0	0	0	0	0

Figure 3: Dataset of 200 patients.

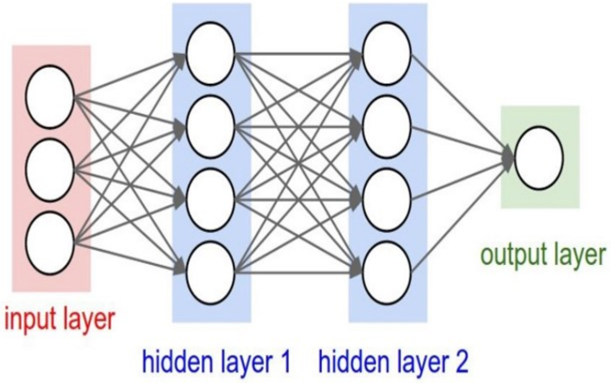
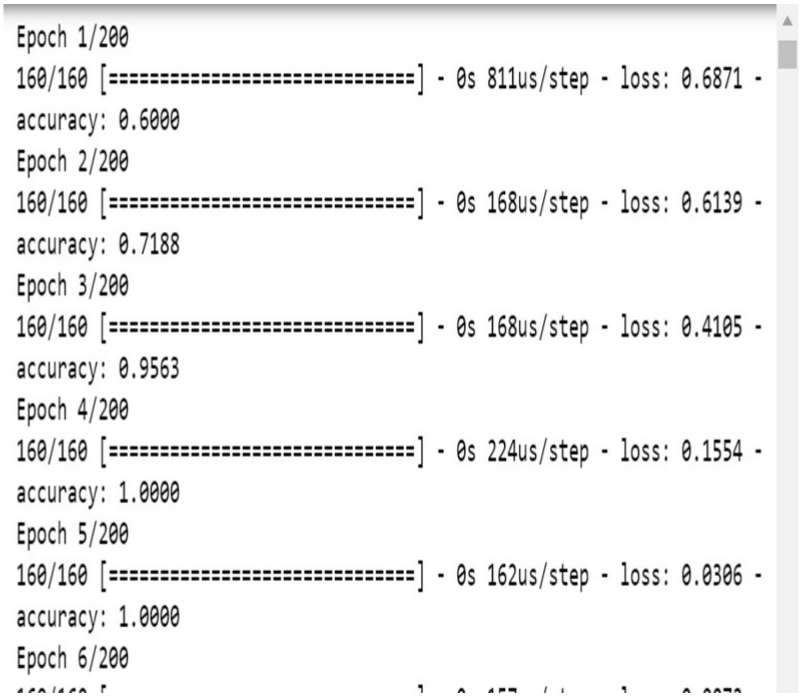
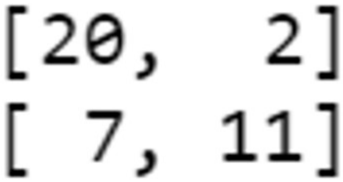


Figure 4: Example Model of Multilayer Perceptron



```
Epoch 1/200
160/160 [==============================] - 0s 811us/step - loss: 0.6871 -
accuracy: 0.6000
Epoch 2/200
160/160 [==============================] - 0s 168us/step - loss: 0.6139 -
accuracy: 0.7188
Epoch 3/200
160/160 [==============================] - 0s 168us/step - loss: 0.4105 -
accuracy: 0.9563
Epoch 4/200
160/160 [==============================] - 0s 224us/step - loss: 0.1554 -
accuracy: 1.0000
Epoch 5/200
160/160 [==============================] - 0s 162us/step - loss: 0.0306 -
accuracy: 1.0000
Epoch 6/200
```

Figure 5: Training of dataset for 200 epochs.



```
[20,  2]
[ 7, 11]
```

Figure 6: Confusion Matrix.