# *Classify and Predict Diseases Using Gene Data*

A comparison between classical machine learning and deep learning methods

Pritom Kumar Mondal
School of Electronic,Information and
Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
pritom@sjtu.edu.cn

***Abstract*— The discovery of human genes that contribute to the appearance and growth of hereditary diseases is an important problem in bioinformatics research. Many techniques have been devised for classifying disease based on gene sequence. Some popular traditional machine learning techniques and deep learning has shown significant improvement in this case. Currently there are many works going on gene based prediction and classification. In this paper, we will use classical and modern machine learning techniques to classify and predict the probability of occurring some diseases based on the genetic data.**

***Keywords—machine leaning, deep learning, gene***

## I. INTRODUCTION

The identification of the genetic basis of human diseases is essential for prenatal and postnatal diagnosis, prognosis and a proper counselling of affected families, and may in some cases even lead to the development of therapeutic strategies.[1]
Machine learning is an exciting field of research in computer science and engineering. It is considered a branch of artificial intelligence because it enables the extraction of meaningful patterns from examples, which is a component of human intelligence. The appeal of having a computer that performs repetitive and well-defined tasks is clear: computers will perform a given task consistently and tirelessly; however, this is less true for humans. More recently, machines have demonstrated the capability to learn and even master tasks that were thought to be too complex for machines, showing that machine learning algorithms are potentially useful components of computer-aided diagnosis and decision support systems.[2]

In this paper, we will use the power of machine learning and deer neural network to classify and predict diseases based on the human gene data.

## II. DATA PROCESSING

Our main dataset is a huge file of about 2.1GB. The file contains almost 6000 observations (human). Our dataset is a P x N matrix where P is the 23,000 genes and N is the number of observations so the matrix is about 23000 x 6000. Since the dataset is huge, we had to do some preprocessing. First performed PCA to reduce the dimensionality. After the reduction of the dimension our P almost become 4000. After the reduction of the dimension, we had to choose the diseases so that we perform the classification and prediction task using logistic regression and deep learning. Frist we tried to do multiclass classification, which needed longer time for training, and accuracy was not satisfactory since some of the disease were rare and those diseases had very few data to train and test. Eventually we choose 5 commonly seen disease for our model. For our experiment we took:

- Breast cancer/tumor: 893 samples
- Huntington's disease: 227 samples
- Leukaemia: 755 samples
- Brain tumor: 103 samples
- Lymphoma: 247 samples

for our model. Most of these disease had different levels but did not dive then into any other subcategories instead we merged all the different levels or sub categories into in big category. After that, we separately performed binary classification for each of the diseases. Therefore, our input 'X' consisted of almost 4000 features and output 'Y' was the binary digits 0 and 1, where 0 means the observation does not have the disease and 1 means the observation has the disease.

## III. METHODS

For our experiment, we mainly chose two different approaches the classical Logistic regression and deep learning approach. After that, we show the comparison of the accuracy of these two approaches. In addition, we also used another popular classical approach

### *A. Classical Approach*

There are several classical approaches that we can take to do the classification and prediction for our gene data. Popular approaches are – Support Vector Machine (SVM), Logistic regression, Principal Component Analysis (PCA) etc.

Since logistic regression is one the most popular and widely used classical method, therefore we choose it as our method. We choose "sklearn" as our framework. Since the gene data was huge in volume, therefore we decide to reduce the dimensionality. Using PCA, we performed the dimension reduction task. Our PCA had 99% variance. Therefore, the features reduced largely. This helped us to speed up our task. It should be mentioned in order to simply the training process we only used binary classification, we did perform some tests on multiclass classification which gave much worse result and took longer time to train. Therefore, we excluded multiclass classification part. For our simplified logistic regression function, we use "L2" norm as penalty and our model had a tolerance of 0.01. We use 30% of the data to perform the test accuracy. We choose "accuracy", "precision", "recall" and "F1 score" as our evaluation matrices.

We also implemented SVM to compare the result with logistic regression. We used the Support Vector Classification (SVC) framework provided by "sklearn".

### B. Deep Learning

We used the keras framework provided by tensorflow for training our model. We used the sequential model, which has one input and output layer and two hidden layer. The input layer contained all the features that we got after doing the PCA. We got altogether 3896 features and therefore the first layer has 3896 units. The first hidden layer has 256 units and second hidden layer has64 units. For both hidden layers we used one of the most popular activation function Relu. And for the output layer we used sigmoid function since our output was either 0 or 1. We tried to use Relu for the output layer but result did not improve rather the accuracy decreased significantly. We also used regularizations in each layer to make sure that the model does not over fit. For this model, we used L1 regularizations with the default value 0.01.

In order to optimize the model we used "Adam" optimizer with a learning rate of 0.001. We tried other learning rate but a higher learning rate overshoots and a lower learning takes much longer time to reach the global minima. Our model did 15 epochs for each disease and for each epochs, the batch size was 88. We chose 'categorical cross entropy' as the loss function for our model. 30% data was used for testing the accuracy. We also used "accuracy", "f1 score", "precision" and "recall" as evaluation matrices so that we can compare our results with the Logistic regression.

## IV. Results

In the Table 1, we can see the comparison of the accuracy of logistic regression and deep learning method. And as expected deep learning gave better accuracy than that of logistic regression. We can see that for "Brest tumor/cancer" both of them give more than or equal to 98% accuracy. In deep learning method, we can see that among all the diseases 'Lymphoma' give the lowest accuracy of 95%, which is still an acceptable result. On the other hand linear regression only gave 83% and 88% for "Huntington" and "Brain tumor" disease, which is much lower than that of deep learning method.

| Disease | Linear Regression (Accuracy) | Deep Learning (Accuracy) |
|---|---|---|
| Breast tumor/cancer | 0.98 | 0.99 |
| Huntington's disease | 0.83 | 0.96 |
| Leukaemia | 0.95 | 0.98 |
| Lymphoma | 0.93 | 0.95 |
| Brain tumor | 0.88 | 0.98 |

**Table 1**

| *Disease Name* | *class* | *Accuracy* | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|---|---|
| *Breast cancer/tumor* | 0 | *0.98* | *1.00* | *0.98* | *0.93* |
| | *1* | | *0.90* | *0.99* | *0.94* |
| *Huntington disease* | *0* | 0.83 | *1.00* | *0.82* | *0.90* |
| | *1* | | *0.17* | *1.00* | *0.29* |
| *Leukaemia* | *0* | *0.95* | *1.00* | *0.94* | *0.97* |
| | *1* | | *0.72* | *1.00* | *0.83* |
| *Brain cancer* | *0* | *0.88* | *1.00* | *0.87* | *0.93* |
| | *1* | | *0.13* | *1.00* | *0.23* |
| *Lymphoma* | *0* | *0.93* | *1.00* | *0.93* | *0.96* |
| | *1* | | *0.38* | *1.00* | *0.55* |

**Table 2**

Table 2. Shows "accuracy", "precision", "recall" and "F1 score" for logistic regression. If we take a look at the result then we can notice that in most of the cases "F1 score" is more than 0.5, only for "Huntinton" and "Brain cancer" has less score (for the disease case [1] ). It is also noticeable that these two disease has the lowest accuracy and lowest precision call among all other diseases. We assuame it happened because of the low data sample in the data set. If we can train the model with some more observations who has these disease then even this perticulr model will be able to give similar accuracy as other diseases. However, for all of these diseases has quite higher results for recall.

| *Disease Name* | *Accuracy* | *Precision* | *Recall* | *F1-score* |
|---|---|---|---|---|
| *Breast cancer/tumor* | *0.99* | *1.00* | *0.82* | *0.90* |
| *Huntington disease* | 0.96 | *0.96* | *0.96* | *0.96* |
| *Leukaemia* | *0.98* | *0.99* | *0.70* | *0.82* |
| *Brain cancer* | *0.98* | *0.98* | *0.98* | *0.98* |
| *Lymphoma* | *0.95* | *0.95* | *0.95* | *0.95* |

**Table 3**

Table 3. shows the result matrices of deep learning method. For all the diseases the results are quite satisfactory. Only the "recall" of "Leukaemia" is slightly lower than other diseases. But since it is 70% , we can say that the result is well inside the acceptation level.

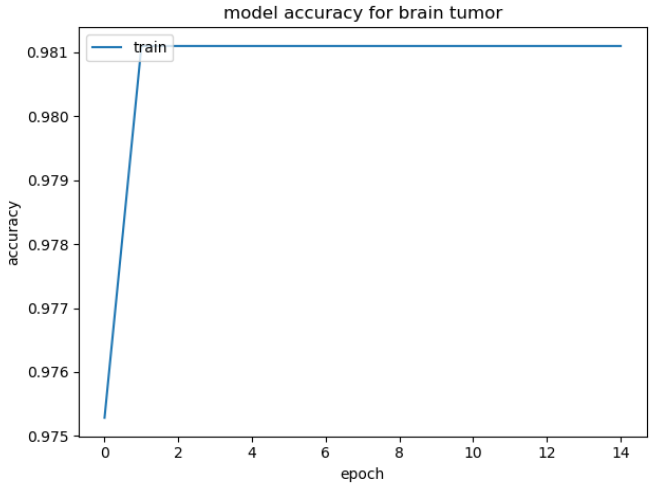


**Figure 1**

Figure 1 to Figure 10 are the graphs generated by deep learning model. We can see the accuracy of different diseases in the odd number figures. In most cases we can notice that the accuracy increases except the Figure 9 where for the "lymphoma" disease the accuracy went down however the accuracy is still 95% which is acceptable.

The even number figures indicates the loss function over epochs. In all the cases the loss decreased over time. We firstly tried with 50 epochs and noticed that after 10 epochs the loss function does not decrease much. Hence, we decided to use only 15 epochs. We also tried a higher regularization but that decreased the accuracy significantly moreover loss values of much more higher. Therefore we came into conclusion that for our model 15 epochs with a learning rate 0.001, regularization 0.01, 2 hidden layer with Relu activation function is a perfect configuration.

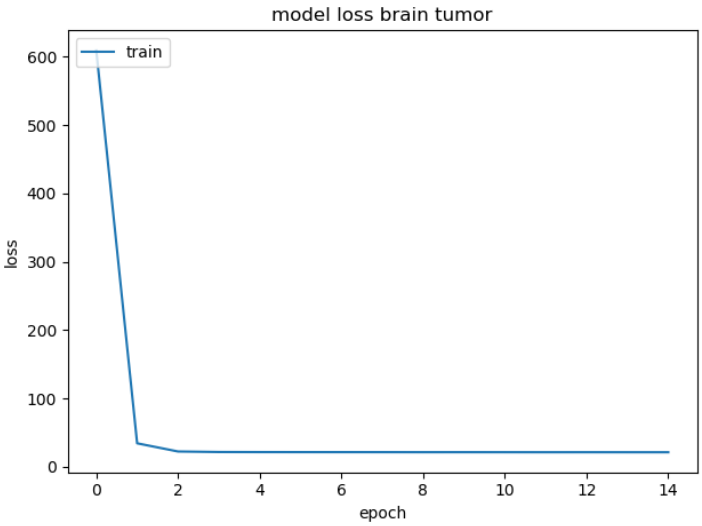


**Figure 2**

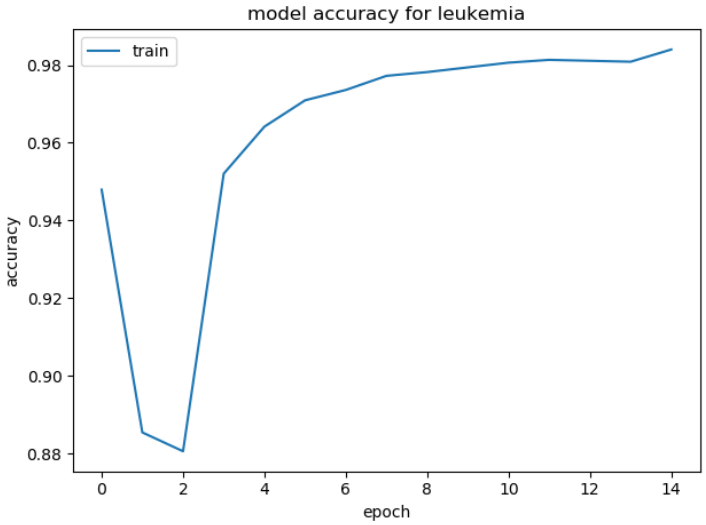


**Figure 3**

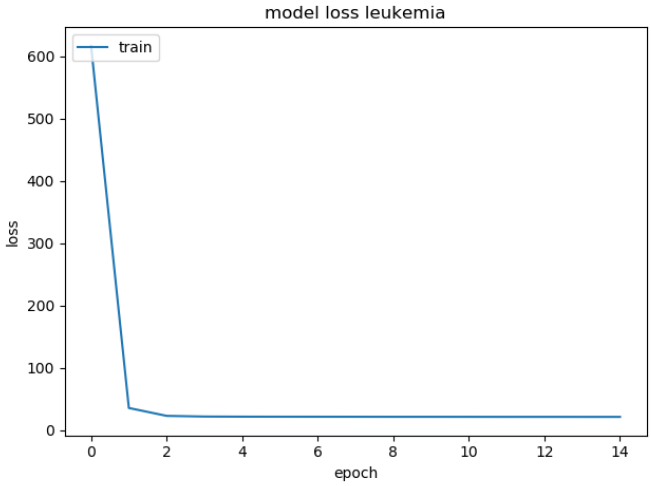


**Figure 4**

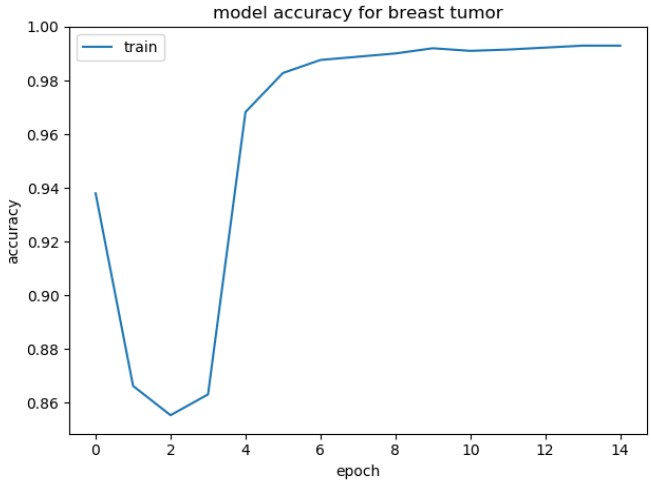


**Figure 5**

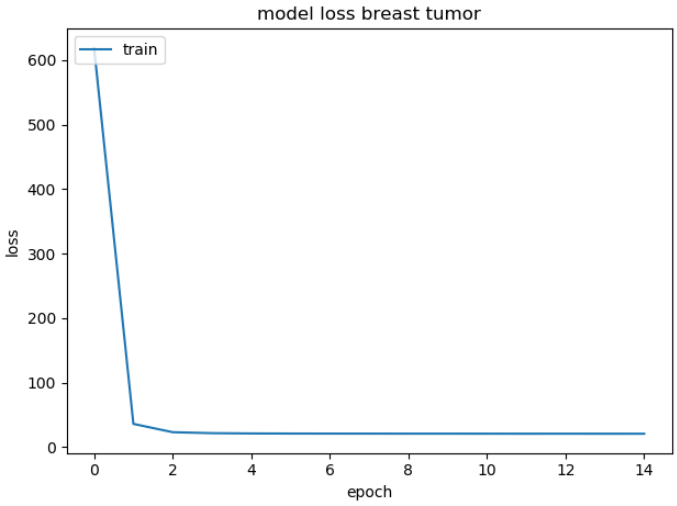


**Figure 6**

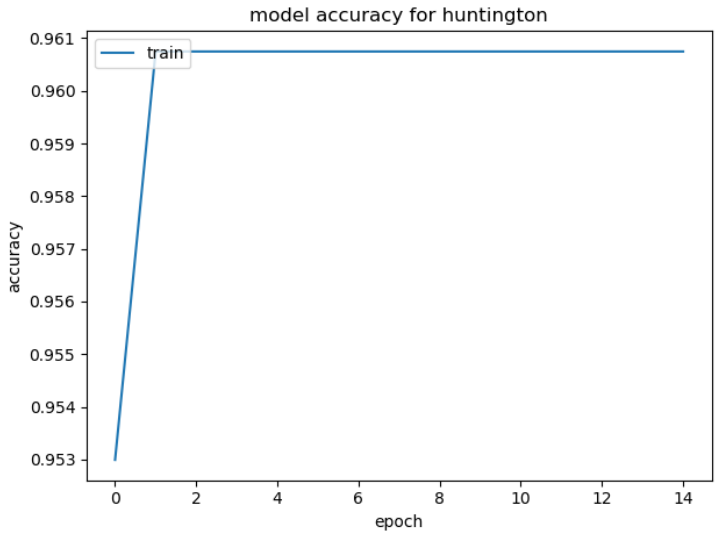


**Figure 7**

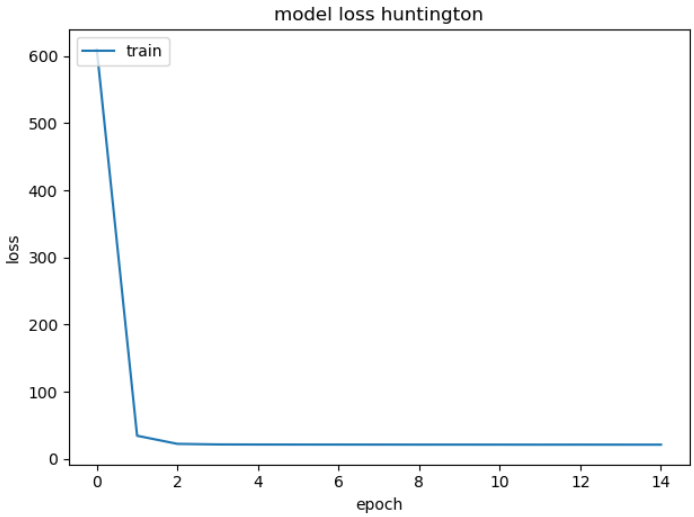


**Figure 8**

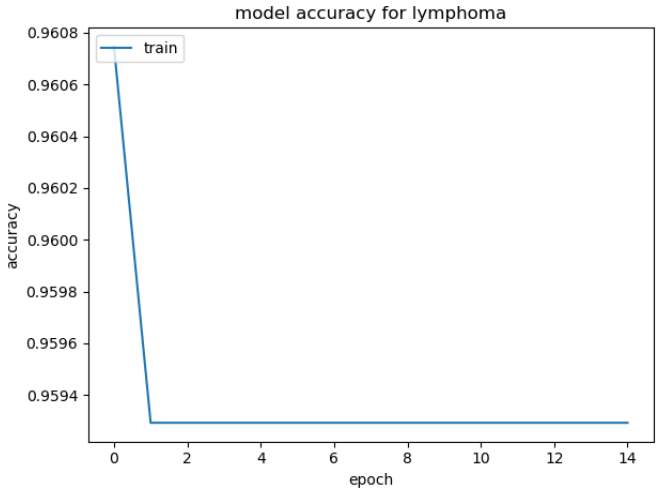


**Figure 9**

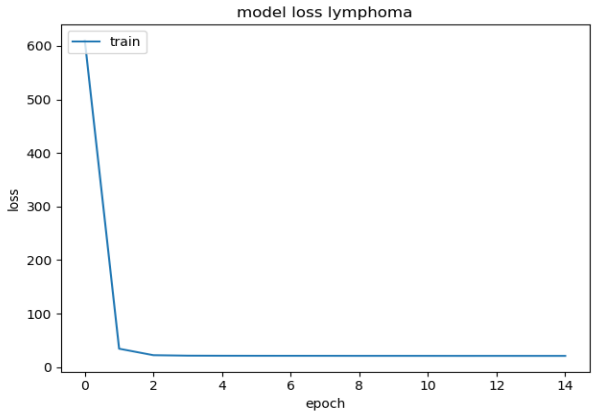


**Figure 10**

## V. Discussions

In this paper, we have shown classification and prediction results of two different approaches. It is clearly noticeable that the modern deep learning methods are much more effective while classifying and predicting the diseases based on the gene data. We also tried other popular classical approach SVM (Linear Support Vector Classification). That model gave almost 100% accuracy for all the cases. We assume that this happened because of overfitting then we used a higher regularization value but yet got 100% accuracy so we discard the result and continued with the Logistic regression model. Our classical logistic regression model gives more than 85% accuracy for any of the diseases and on the other hand, deep learning sequential model can give more than 90% accuracy for any diseases.

## Acknowledgment

We are greatly thankful to Prof. Yuan Bo who gave us a brief insight of biology and gene in his class. Moreover he helped us providing the valuable gene data which we are willing to use in future and do some more valuable works. Lastly thanks to Petros who worked closely with me in this project.

## VI. References

[1] M. P. a. P. K. P. Z. A. P. a. T. L. K. Bradley J. Erickson, "Machine Learning for Medical Imaging," 2017.

[2] R. M. P. F. D. Cunto, "Computational approaches to disease-gene prediction: rationale, classification and successes," *The FEBS Journal*, vol. 279, no. 5, 2012.