Predicting Survivability Based on Gene Expression of Cancer Patients

*Adeseye Adekunle*

*School of Computing Engineering and Digital Technology (SCEDT)*

*Teesside University*

*Middlesbrough, United Kingdom*

*B1081572@live.tees.ac.uk*

*Abstract*—The mortality rate associated with cancer as been a major global crisis within and out of the health sector. Methods and research have been initiated and are ongoing to analysis and understand the diseases, that ravages any organ of the human body. The most promising have been through the analysis of gene expression of cancer patients. By measuring the rate at with genes translate and transcribe to protein synthesis, drug administration pathways could be better implemented, and patient care would be better managed. Machine learning and Deep learning have proven to be the perfect tool for understanding large volume of data produced by gene expression matrix and in this experiment, an algorithm stack of 11 machine learning and 1 deep learning algorithms were used to create a model from the cancer patient data. Decision Tree, Support Vector Machine, Adaboost and Naïve Bayes algorithms gave the highest accuracy of 52% before tuning. By implementing a hyperparameter tuning, Decision Tree accuracy increase to 74%.

Keywords—Gene Expression, Ribonucleic Acid, Deoxyribonucleic Acid, Machine Learning, Deep Learning, Protein Synthesis, Adenine, Cytosine, Thiamine, Guanine, Uracil

# Introduction

According to the United Kingdom cancer research, cancer happens when genetic changes result in single or multiple cell creation [1]. These cells are continuously produced in an uncontrollable fashion leading to an abnormal growth of tissues(tumors) [1]. Autopsies in recent times shows cancer as a leading cause of death globally, which has spurned interest in research, to what could be the cause? How can we quickly detect cancerous cells in time? And what measures(medication) could be taken to avert deaths associated with cancer.

Gene expression data genomic data has been identified as a solution to solving the problem [2]. Gene expression is the processes of protein synthesis via RNA transcription and protein translation [3]. This makes use of the four-nucleotide chain base, combination (A, C, T, G or U) in the DNA-RNA protein synthesis [4]. The synthesized RNA and protein product act as regulatory mechanism for other gene expression.

Cancer being a terminal and debilitating disease can be managed through the survivability time of its patient, medical interventions via drug administration and proper care if genomic data about the patients are understood [2]. Machine learning and deep learning has been identified as a perfect tool for analyzing gene expression matrices, which involves combinations of nucleotide bases that produces gene expression microarrays ranging in thousands for each gene sample [4].

Multiple gene expressions producing huge amount of data seems perfect for machine learning models to interpret, with the downsize being high dimensionality data compared to the data points. This intention of this paper is to show the development of a machine learning or deep learning model that takes in the gene expression matrix and predict the survival of the patient.

# Literature Review

According to [4] gene expression microarray has made it possible to measure the translation rate of whole genes in cells or tissues simultaneously. Using machine learning, automated analysis can be carried out from the large volume of data produced by this gene expression data. The report by [2] believes that gene expression data will help predict the survival time of cancer patients, that will give a streamlined and specific treatment option and better care. They introduced a novel approach to solving the problem of high dimensionality to data points that plagues most gene expression models. On the METABRIC dataset, a discretized Latent Dirichlet Allocation (dLDA) followed by multitask logistic regression was used to reduce the dimensionality and this increased the accuracy of prediction greatly, even when a new Pan kidney was introduced to the approach and model.

# Data Source and description

Dataset source was from the gene expression data of patients with cancer and could be found at <https://mega.nz/folder/EvhkDaRb#6lQrn5h3oA67vsPgIAxwvg> . It consists of several columns with gene expression matrix each representing a single case. Physical inspection of the dataset shows inverted columns as rows that represent features and with the first 9 rows being clinical information about the patient. Gene expression matrix starts from the 10th row with prefix code ENSG0000000000X.XX. Addition data is also included, such as vital status, age and days to death, which could be used as features in the dataset.

## Target Variable

The vital status feature was leveraged to be used as the target variable for the prediction. This biopic samples makes it easy to classify the gene expression matrix and allows samples to be categorized as living or dead, even though it shows the mortality state from which the sample was obtained. The feature consists of only two labels that defines the mortality state of the patient. The labels are identified, Alive and Dead.

Chart, bar chart

Description automatically generated

*Figure 1* : *Label Distribution of the target feature*

The distribution of the labels or class within the target feature, is 23% Alive labels and 35% Dead labels.

## Features Distribution

Chart, histogram

Description automatically generated

*Figure 2:* *Mortality Status for Cancer Stages*

This shows the distribution of each sample at every cancer stage within the data. Majority of the patient in Stage I and Stage II show uncertainty or diagnosis that shows they might live and are in critical condition, while Stage III and Stage IV includes patients with high probability of dying.

Chart, histogram

Description automatically generated

*Figure 3:* *Mortality Status for patient initial weight distribution*

According to the statistical research from the American Cancer Research, being overweight increases the likelihood of having cancer, and this is linked to cancer detected in 11% of women and 5% of men [5]. From figure 4, it can be observed that there are spikes in death count as the distribution of patient initial weight goes above 200.

Chart, histogram

Description automatically generated

*Figure 4:* *Mortality Status for patient Age*

The National Cancer Institute believe that advancing age is the most important risk factor for all cancer types [6]. Detection only occurs in 25 cases for each 100000 individuals below the age of 20, to 350 cases per 100000 individuals within the range of 45-49 and 1000 cases per 100000 individuals above the age of 60. From figure 5 above, it could be observed that mortality occurs within the data sample as age increases above 50 with few isolated cases of live patients with age above 55.

# IV. METHOD AND EXPERIMENT

## Feature Description and Exploratory Data Analysis

The experiment was carried using the data mentioned above. Relevant for loading the dataset to model implementation was imported and used. The dataset was loaded into the jupyter lab environment, which shows each row as a feature within the dataset and clinical information included at the top of the data. Table 1 shows the description of the clinical information of patient and gene expression matrix.

|  |  |
| --- | --- |
| Features | Description |
| Initial Weight | The measured weight of the patient |
| Age At Diagnosis | Age of the patient at onset of cancer diagnosis |
| Vital Status | Shows if patient is alive or dead |
| Days to Birth | The number of days to patient birthday |
| Year of Birth | The year of birth of patient |
| Demographic ID | The unique identification for patient’s demographic. |
| Year of Death | Shows the year patient died, if sample was from dead patient. |
| Days to Death | Terminally ill patients at late stages of cancer |
| Paper Clinical Stage | Shows the severity of cancer in each patient or sample |
| ENSG0000000000X.XX | The gene expression matrix for each sample |

***Table 1. Feature Description***

## Missing Values

On inspection and using missing value function from the python (is null), there were no missing values present in the dataset.

A picture containing chart

Description automatically generated

*Figure 5:* *Visualizing the missing values on a plot*

We initiated a data preprocessing step by first dropping columns we deem not important, leaving only the initial weight, age at index and paper clinical stage from the clinical information of the patient. The dataset was transposed, and index column reset, and previous index removed from as the column header. On checking the data type of the dataset, each feature has the object datatype which was changed to numeric for the gene expression matrix, initial weight and age at glance. The values for the vital status were replaced with 1 for alive and 0 for dead with the datatype changed to int64. Paper clinical stage which consists of 4 labels was changed using one hot encoding to numerical value.

The dataset was normalized using the minmax scaler and a principal component analysis was implemented using 37 components to reduce the size of the dimensionality. Dataset was split into 33% testing set and 67% training set before implementation with model algorithms.

# MACHINE LEARNING AND DEEP LEARNING MODELS USED

A stack of 11 machine learning model and 1 deep learning models were implemented for easy comparison between each model, and to check the accuracy of each before hyperparameter tuning was added. Two of the models with the highest accuracy and the neural network are discussed below.

## Machine Learning Models

### Support Vector Machine (SVM)

This is a supervised ML algorithm that is used for classification and regression problems. This uses hyperplanes to differentiate two labels of a binary classification along distinct data points (support vectors). This algorithm was implemented using the radial bias and polynomial kernels. It gave a high accuracy when training and testing data was fitted to the model.

### Decision Tree (DT)

This is a supervised ML algorithm that can be used for both classification and regression problems but mostly used for classification problems. The internal nodes of each tree represent the dataset features while the branches represent decision rules. The outcome is the node of each leaf. Max depth was set at 5 to obtain a high accuracy be tuning.

## Neural Network

### Multilayer Perceptron (MLP)

This is a neural network used for classification regression problems. It implements this process by inputting weighted features which is done by a linear equation called activation, into a node, and outputs a block of code which has high value of a specific class in relation to other low values of other class, using a step function. At least a hidden layer exists in between the input and output layers, for approximating function. Using multilayer perceptron, mean probabilities received from the training model with 5-folds CV, was used to calculate the AUROC curve on the test data which produced 0.942. Tunning the hyperparameters, of hidden layer sizes, activation, leaning rate and solver produced an optimized value for the algorithm.

## Evaluation Metric For Model Performance

Model performance was evaluated using Precision, Recall and Accuracy.

\* Precision is the fraction of relevant examples (true positives) among all the examples which were predicted to belong in a certain class.

precision= true positives

true positives + false positives

\*Recall is the fraction of examples which were predicted to belong to a class with respect to all the examples that truly belong in the class.

Recall = true positives

true positives +false negatives

\*Accuracy refers to the percentage of correct predictions for the test data. It is calculated by dividing the number of correct predictions by the number of total predictions.

accuracy= correct predictions

all predictions

\* Precision and recall can be combined and used as a parameter for evaluating the performance of a metric this conjunction is known as the F-score.

Fβ= (1+β2 ) precision \* recall

(β2⋅precision) +recall

# Results

This are the results obtained when 12 algorithms were fitted to the dataset before hyperparameter tuning. SVM, DT, Adaboost and Naïve Bayes show high accuracy, while the neural network gave below optimal accuracy.

A screenshot of a computer

Description automatically generated with medium confidence

*Figure 6:* *Visualizing the accuracies of the model stack*

## Tuning and Result

By tuning the Decision Tree’s max depth to an optimal setting of 3 and max leaf node at 50, a 74% accuracy was obtained. Tuning the hyperparameter of SVM, which includes setting C uniform between a range of 1 to 10, a radial bias kernel of gamma, with cross validation of 3 all tuned in random search cv, the accuracy is reduced. C and gamma were used from the best param estimation. Tuning the MLP, we used a grid search CV and acquired the activation function of ‘relu’, learning rate set to constant, solver to adam, and a range of list values from 50 to 200, the accuracy increased to 47%.

# Discussion

Several machine and deep learning models were fitted to the gene expression dataset of cancer patients. While some models before tuning had an accuracy above 50, several also fall below the 50% accuracy value. The Decision Tree model performed well with an accuracy of 52% before tuning to 74% after hyperparameter tuning. Hence, it is the selected model for the implementation of this ML process.

# Reference

[1] WHAT IS CANCER? Cancer Research UK. 2022. What is cancer? [online] Available at: <https://www.cancerresearchuk.org/about-cancer/what-is-cancer> [Accessed 8 June 2022].

[2] KUMAR, L. AND GREINER, R. Gene expression-based survival prediction for cancer patients—A topic modeling approach Kumar, L. and Greiner, R., 2022. Gene expression-based survival prediction for cancer patients—A topic modeling approach. [online] Available at: <https://pubmed.ncbi.nlm.nih.gov/31730620/> [Accessed 8 June 2022].

[3] GENE EXPRESSION Genome.gov. 2022. Gene Expression. [online] Available at: <https://www.genome.gov/genetics-glossary/Gene-Expression> [Accessed 8 June 2022].

[4] Using Machine Learning to Design and Interpret Gene Expression Microarray Molla, M., Waddell, M., Page, D. and Shavlik, J., 2004. Using Machine Learning to Design and Interpret Gene Expression Microarray. AI Magazine, [online] Available at: <https://www.researchgate.net/publication/220604647\_Using\_Machine\_Learning\_to\_Design\_and\_Interpret\_Gene-Expression\_Microarrays> [Accessed 8 June 2022].

[5] DOES BODY WEIGHT AFFECT CANCER RISK? Cancer.org. 2020. Does Body Weight Affect Cancer Risk?. [online] Available at: <https://www.cancer.org/healthy/cancer-causes/diet-physical-activity/body-weight-and-cancer-risk/effects.html#:~:text=Being%20overweight%20or%20obese%20is,7%25%20of%20all%20cancer%20deaths.> [Accessed 9 June 2022].

[6] RISK FACTORS: AGE National Cancer Institute. 2021. Risk Factors: Age. [online] Available at: <https://www.cancer.gov/about-cancer/causes-prevention/risk/age#:~:text=Age%20and%20Cancer%20Risk&text=The%20incidence%20rates%20for%20cancer,groups%2060%20years%20and%20older.> [Accessed 9 June 2022]

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**