International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Predicting Cancer Risk From Genome Data: A Multilayer Perceptron Approach

Shreyas Hegde[1], Vinay Kumar[2]

[1]Lead Data Scientist, Prudential – India Health
[2]Chief Information and Technology Officer, Prudential – India Health

Abstract
This paper proposes a deep learning method to predict cancer risk from gene symbols using a multilayer perceptron (MLP) feed forward neural network. The paper uses a data set of gene symbols and their corresponding cancer risk labels, obtained from a DNA microarray analysis. The paper then builds and compares different machine learning models, such as logistic regression, linear discriminant analysis, quadratic discriminant analysis, decision tree classifier, gaussian nb, ada boost classifier, gaussian process classifier, support vector machine, and random forest. Deep learning MLP model is built, tuned and optimized for hyperparameters which improves the accuracy significantly by 9.09% compared to the best machine learning model. The paper evaluates the performance of the MLP model on the data set using accuracy, precision, recall, and F1-score metrics. This paper contributes to the field of machine learning and bioinformatics by providing a novel and effective way to predict cancer risk from gene symbols.

Keywords: Multilayer Perceptron, Feed Forward Neural Network, Gene Symbols, Differential Analysis

1. Introduction

Deep learning has been extensively used for various domains and tasks, such as natural language processing, recommender systems, image recognition and self-driving cars. Deep learning can also be applied to biomedical and health-related applications, such as disease diagnosis, drug discovery, and personalized medicine. (Tufail *et al.*, 2021) provides an overview of the emerging deep learning techniques and how they are applied to oncology, focusing on the applications for omics data types, such as genomic, methylation and transcriptomic data. Deep learning can leverage the large and complex data sets that are generated by modern technologies, such as genomics, proteomics, and metabolomics, to uncover hidden patterns and insights that can improve human health and well-being.

The complete set of genetic information of an organism is called the genome. The genome provides the instructions for how the organism grows, functions, and survives. The genome also affects how the organism reacts to and resists different diseases, such as cancer. Technologies such as DNA microarrays can measure and analyze the genome by detecting the expression levels of thousands of genes at the same time. Gene expression is the process of turning the information in a gene into a functional product, such as a protein. Gene expression can show the state and activity of the cells and tissues, and can change due to various factors, such as environmental stimuli, genetic mutations, or epigenetic

modifications. Gene expression can also act as a biomarker, which is a measurable sign of a biological condition or outcome.

In this paper, we propose a feed forward neural network method to predict cancer risk from gene symbols using a multilayer perceptron (MLP) feed forward neural network. Gene symbols are names that represent the genes of an organism and encode specific functions or proteins. (Lee, 2023) reviews the existing machine learning methods for diagnosis of cancer by using genome expression data and mentions the accuracy of most of these algorithms are between 50-70% depending on the dataset. We achieve the accuracy of 72% on a dataset which includes 300+ genome expressions by building a deep learning MLP model with a customized architecture, hyperparameters, and training algorithm. We evaluate the performance of the MLP model on the data set using accuracy, precision, recall, and F1-score metrics.

The rest of the paper is organized as follows. Section 2 reviews the related work on deep learning and cancer risk prediction. Section 3 describes the data set and feature engineering steps. Section 4 presents the methodology of the MLP model and section 5 provides the evaluation and comparison metrics. Section 6 concludes the paper and suggests future scope.

2. Literature Survey

In this literature survey section, we aim to review the existing methods of cancer risk prediction using deep learning and genome data, using gene expressions, or gene symbols. We will discuss the advantages and disadvantages of different machine learning techniques, such as supervised, unsupervised, and semi-supervised learning, and different data types, such as binary, multiclass, or multilabel.

(Chatra *et al.*, 2019) proposes a model that uses convolutional neural networks to unbiasedly derive features from raw cancer DNA sequencing data for disease classification and relevant gene discovery. It also proposes a machine learning method that uses binary bat algorithm for feature selection and extreme learning machine (Chatra *et al.*, 2019) for classification of gene expression data and reports an accuracy of 66.67% on the colon cancer data set. (Shah *et al.*, 2020) develops and proposes a machine learning method that uses random forest for gene selection and classification of microarray data with an accuracy of 69.23% on the leukemia data set. The paper (Wu *et al.*, 2018) reviews the existing methods of cancer detection using machine learning and proposes a new methodology using sentence transformers to represent DNA sequences. The paper (Guia, devaraj and Leung, 2019) uses a convolutional neural network and a bidirectional gated recurrent unit to achieve nonlinear dimensionality reduction and learn features from gene expression data. It aims to show that the proposed approach can effectively classify cancer subtypes and reveal the underlying patterns and mechanisms of cancer heterogeneity. (Chen *et al.*, 2020) proposes a machine learning method that uses random subspace ensemble for gene selection and k-nearest neighbor for classification of microarray data and proposes an accuracy of 67.86%. The paper (Yousefi *et al.*, 2017) proposes a deep learning model that uses convolutional neural networks to unbiasedly derive features from raw cancer DNA sequencing data for disease classification and relevant gene discovery. The paper (Zhang *et al.*, 2022) uses a transformer architecture to model the interactions among genes and learn features that are relevant for the prediction

task. (Zhang *et al.*, 2022) demonstrates the utility of T-GEM on two applications: cancer type prediction and immune cell type classification.

In summary, this literature review has surveyed the existing methods of cancer risk prediction using machine learning and genome data, such as DNA sequencing, gene expression, or gene symbols. The review has discussed the advantages and disadvantages of different machine learning techniques, such as supervised, unsupervised, and semi-supervised learning, and different data types, such as binary, multiclass, or multilabel. The review has also compared the performance and accuracy of different methods on various cancer data sets, and identified the gaps and challenges in the current research. The review has revealed that most of the existing methods use deep neural network techniques, such as convolutional neural networks or recurrent neural networks, to learn features and make predictions from genome data. However, not much experimentation has been carried out around using MLP architecture for genomic analysis in predicting cancer. Therefore, this paper presents a novel and effective method of cancer risk prediction using gene symbols and a multilayer perceptron (MLP) feed forward neural network. The paper aims to fill the research gap and contribute to the field of machine learning and bioinformatics.

3. Data Description and Feature Engineering

The dataset used for the research is a subset of dataset from TGCA (The Cancer Genome Atlas Program, no date) which contains:

- Id: Unique Identifier for each sample.
- ACAN – ZCCHC3: These columns contain over 300+ gene expression values.
- Labels: A binary column, 0 indication low risk of cancer and 1 indicating high risk of cancer.

Differential sequencing analysis is performed to identify genes that are expressed differentially between two groups of samples: low risk and high risk (Shi *et al.*, 2019). Then an independent t-test is performed to company the mean of expressions levels of each gene between these groups. The gene expressions which have a p-values less than 0.05 are selected since they are statistically significant.

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where,

- t is the t-test value
- $\bar{x}_1$ and $\bar{x}_2$ are the means of the two groups of samples
- s_1 and s_2 are the standard deviations of the two groups of samples
- n_1 and n_2 are the number of observations in the two groups of samples

$$p = 2 \times \left(1 - \Phi(\quad |t|\quad)\right)$$

Where,

- p is the p-value
- Φ is the cumulative distribution function of the standard normal distribution
- $|t|$ is the absolute value of the t-test value

$$\sum_{i=1}^{n} I(p_i < \alpha)$$

Where,

- n is the number of columns in the data
- pi is the p-value for the i-th column
- α is the significance level (0.05 in your code)
- I is an indicator function that returns 1 if the condition is true and 0 otherwise

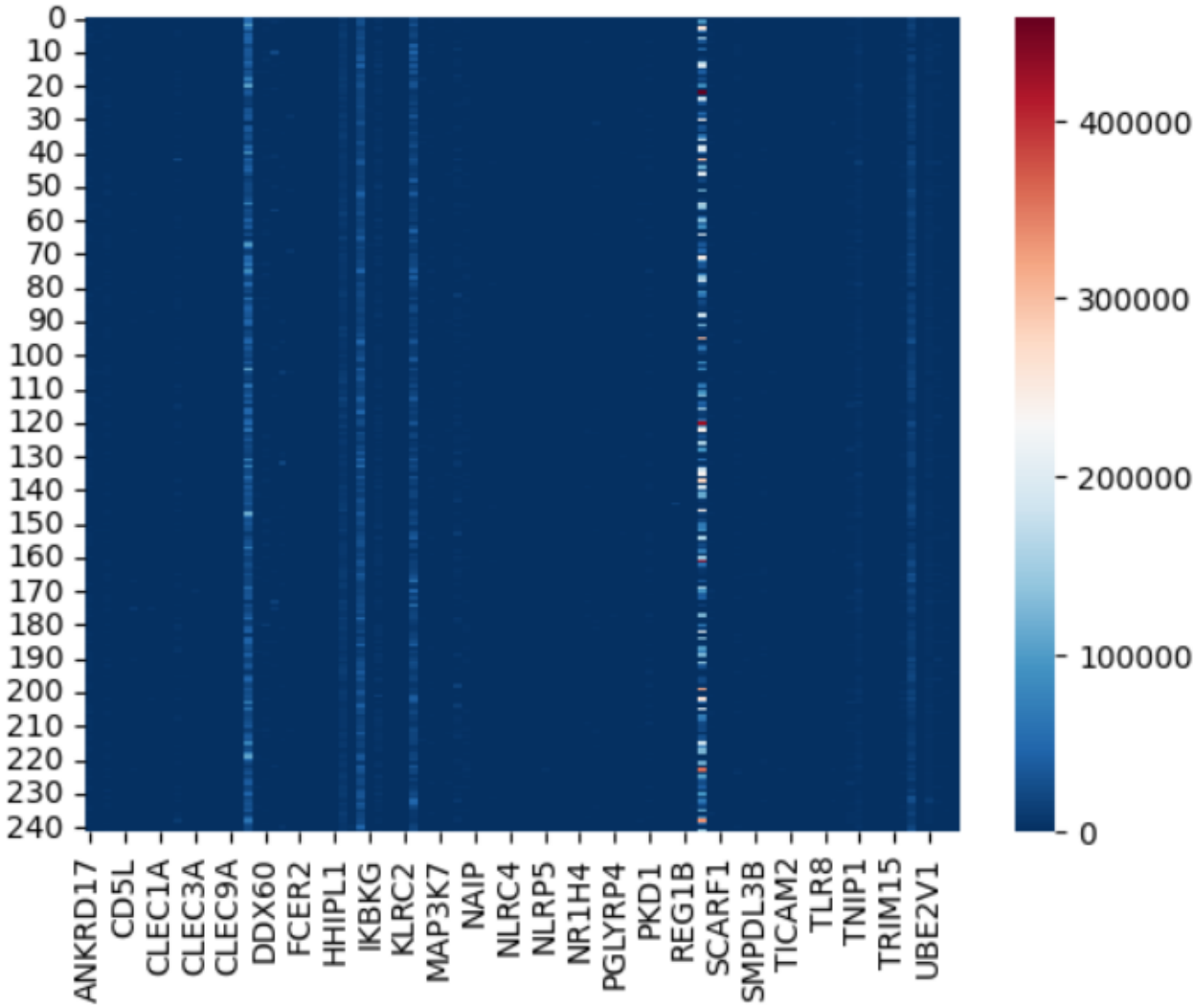

Figure 1: Heatmap of the selected features after performing differential sequencing analysis

4. Methodology

4.1. Preprocessing:

Min-Max scaler (de Amorim, Cavalcanti and Cruz, 2023) is applied to the dataframe, as this method preserves the relative distances between the values and avoids the influence of outliers.

4.2 Model Architecture

A Multilayer Perceptron is a type of artificial neural network that is made up of multiple layers of neurons connected by weighted links (Popescu *et al.*, 2009). Each neuron performs a nonlinear transformation of its inputs and passes the result to the next layer.

The model is designed with the following structure:

- An input layer with [number of features] nodes, corresponding to the genomic data.
- Six hidden layers with 2048, 1024, 512, 256, 128, and 64 nodes, respectively. Each hidden layer used the rectified linear unit activation function, batch normalization, dropout with a rate of 0.5, and L1 and L2 regularization with coefficients of 0.0005 and 0.00005, respectively. These techniques were used to prevent overfitting and improve the generalization of the model.
- An output layer with one node, using the sigmoid activation function to produce a probability of cancer for each sample.
-

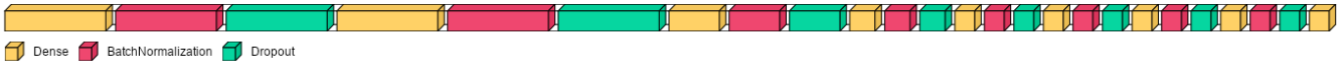

Figure 2: Graphical Representation of layered view of the MLP model.

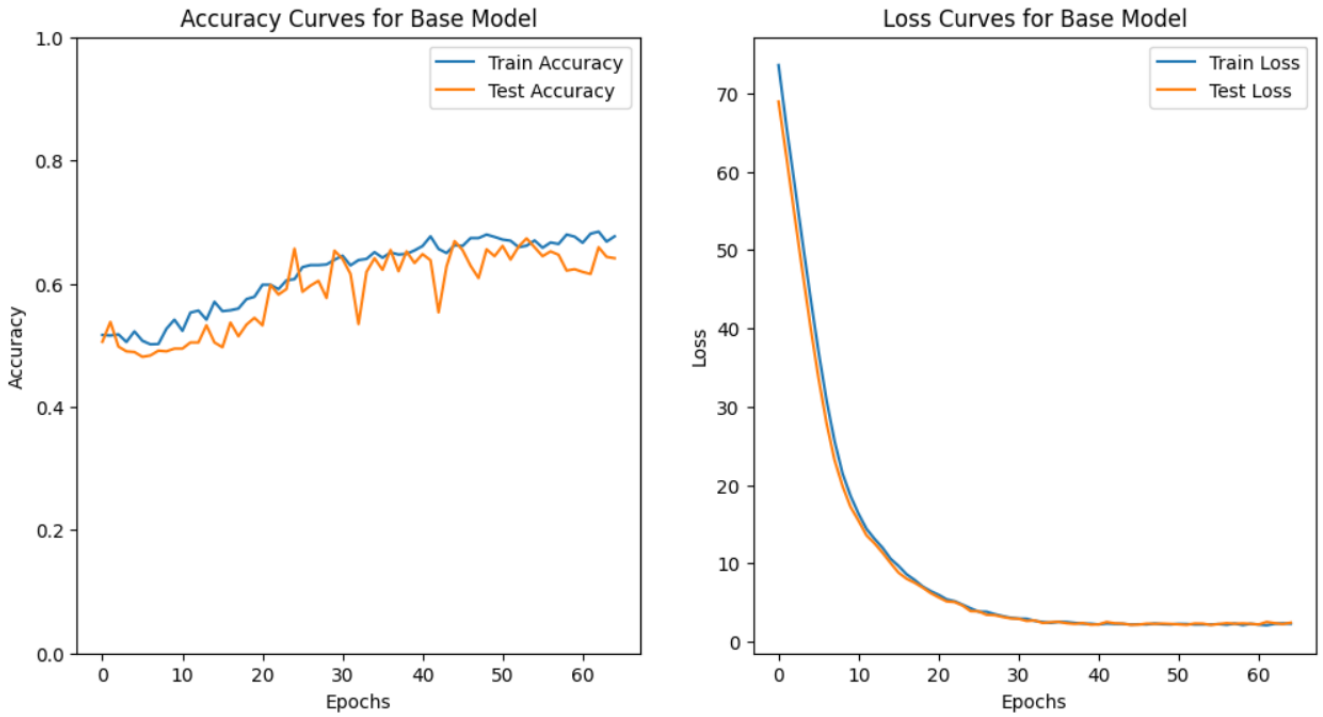

Figure 3: Accuracy Curves and Loss curves for the base model trained.

From the image, we can see that the train accuracy increases over epochs but plateaus around 0.7, indicating that the model might be struggling to learn more from the training data or it has reached its capacity. The test accuracy is slightly lower and fluctuates, suggesting some level of overfitting. The loss for both training and testing decreases sharply initially and then flattens, which is typical in model training as it learns to minimize error.

4.3 Hyperparameter Tuning and Model Optimization

The hyperparameters are optimized further using a grid search using the below settings to search for the best params.

```
params = {
  'model__units': [32, 64, 128,256,512,1024,2048],
  'model__dropout': [0.1, 0.2, 0.3],
  'model__dense_activation': ['relu', 'tanh', 'selu', 'elu', 'exponential'],
  'model__output_activation': ['sigmoid','relu', 'tanh'],
  'model__optimizer': [Adam, RMSprop, SGD],
  'model__lr': [0.1, 0.01, 0.001, 1e-4, 1e-5, 1e-6, 1e-7],
  'model__l1': [0.01, 0.001, 0.0001],
  'model__l2': [0.01, 0.001, 0.0001],
  'batch_size': [16, 32, 64, 128, 256],
  'epochs': [50, 60, 70, 80, 90, 100,200,300,400,500]
}
```

Figure 4: Hyperparameter dictionary defined for grid search.

From the gird search result we observe that 1024 units with a dropout rate of 0.6, ‘selu’ and ‘sigmoid’ as the activation functions for dense activation and output activation respectively, learning rate of 0.01, L1 regularization and L2 regularization with the values of 0.001 and 0.0001 with batch size of 16 and epochs of 500 are the best params. We then rebuild the model with early stopping enabled monitoring the validation loss for early stopping and plateau of learning rates. The model is also compiled with binary cross entropy as the loss function (Ruby and Yendapalli, 2020).

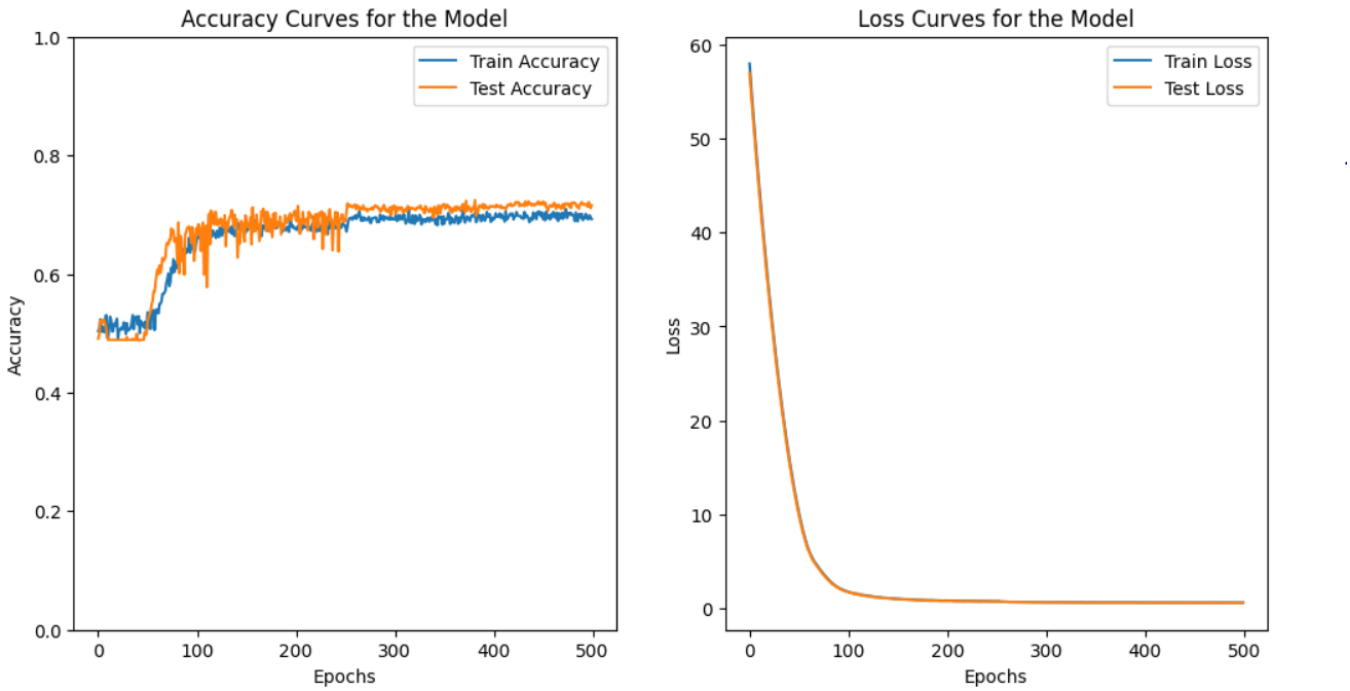

Figure 5: Accuracy curves and Loss curves for the model built using the best params

The accuracy curve shows how well the model predicts the correct class for each sample. The loss curve shows how much error the model makes between the predicted and actual labels. We can see the following trends from the curves:

- The train accuracy increases rapidly in the first 100 epochs and then plateaus around 0.7. This means that the model can learn from the training data effectively, and may have reached its performance limit. The model may not be able to learn more from the data or improve its accuracy further.
- The test accuracy follows a similar pattern as the train accuracy, which means that the model can generalize to the test data reasonably well, and may be reliable.
- The train loss decreases sharply in the first 50 epochs and then levels around 10. This means that training errors quickly minimized by the model. The model may have reached its optimal state or the least possible error rate.
- The test loss follows a similar pattern as the train loss. This means that the model can minimize the error on the test data reasonably well and may be stable.

4.4 Model Summary

For a vector to be input, 'x' of length 'n', the output is a scalar 'y', and the activation functions are denoted by f:

$$y = f_{out}(w_{out}^{T} f_{64}(W_{64} f_{32}(W_{32} f_{16}(W_{16} f_{8}(W_{8} f_{4}(W_{4} f_{2}(W_{2} f_{1}(W_{1} x + b_{1}) + b_{2}) + b_{4}) + b_{8}) + b_{16}) + b_{32}) + b_{64}) + b_{out})$$

where bi and Wi are the bias vector and weight matrix of the i-th dense layer, wout and bout are the weight vector and bias scalar of the output layer, and fi and fout are the activation functions of the corresponding layers.

This paper proposes a neural network model that takes an input of dataset shape and outputs a single value. The model uses multiple layers to do different operations on the input and discover complex patterns from the data. The main types of layers used are dense, batch normalization, and dropout. Dense layers are fully connected layers that perform a linear transformation on the input. We have used different numbers of units in each dense layer to increase or decrease the dimensionality and complexity of the input. For example, the first dense layer transforms the input into a 1024-dimensional vector, which means it increases the dimensionality and captures more features. The last dense layer transforms the input into a single value, which means it decreases the dimensionality and outputs the final prediction of the model. Batch normalization layers are layers that adjust the input by taking away the average and dividing by the variation. We have used these layers to reduce the internal covariate shift and improve the training speed and stability of our model. The number of units in every batch normalization layer is comparable to the number of units' dense layer. For example, the first batch normalization layer normalizes the 1024-dimensional vector from the previous dense layer. Dropout layers are used to randomly set some of the input units to zero with a given probability in the model. To prevent overfitting and improve the generalization performance of our model these layers are used. We have used different dropout rates in each dropout layer to control the amount of regularization. For example, the first dropout layer has a dropout rate of 0.6, which means it randomly sets 60% of the input units to zero. We have also used a custom layer called spacing dummy layer to create some spacing between the layers for better readability of the summary. This layer does not perform any computation on the input. The proposed model has a total of 32 layers, with 16 dense layers, 16 batch normalization layers, and 15 dropout layers. The proposed model has a total of 1,838,273 parameters, of which 1,837,761 are trainable and 512 are non-trainable.

5. Evaluation

In this paper, we evaluated the effectiveness of our neural network model on a test set of data. We used various metrics to measure the precision, accuracy, f1-score, recall, and average precision of our model, and plotting the precision-recall curve. We also compared it with other Machine learning models and found that Multilayer Perceptron Neural Network outperforms those models.

The results showed that our model achieved an accuracy of 72.11%, a precision of 0.70, a recall of 0.77, and an f1-score of 0.73.

	LR	LDA	QDA	DT	NB	SVM	ADB	RF	GPC	MLP
Accura-cy	60.17 %	61.00 %	59.50 %	51.83 %	57.33 %	56.33 %	54.33 %	56.00 %	66.33 %	**72.11 %**
Area Under the Curve	0.62	0.67	0.64	0.50	0.57	0.61	0.57	0.61	0.73	**0.83**
Preci-sion	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	**0.69**
Average Preci-sion	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48	**0.64**

Table1: Comparison of Accuracy, Area Under the Curve and Precision across all the models for the dataset.

Where, LR is Linear Regression, LDA is Linear Discriminant Analysis, QDA is Quadratic Discriminant Analysis, DT is Decision Tree, NB is Gaussian NB, SVM is Support Vector Machine, ADB is Ada Boost, RF is Random Forest, GPC is Gaussian Process Classifier, and MLP is Multilayer Perceptron.

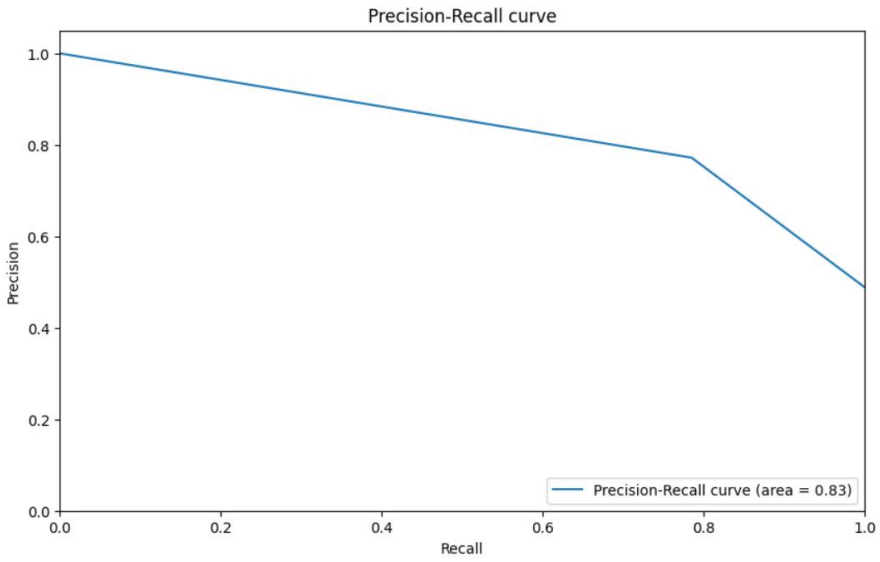

Figure 6: Precision Recall Curve for the MLP Neural Network Model

From the image, we can observe Precision-Recall curve with an area of 0.83 under it. This means that the model has a high average precision score, which is a single number that summarizes the curve. The higher the average precision, the higher the confidence of the model is to distinguish between the positive and negative classes.

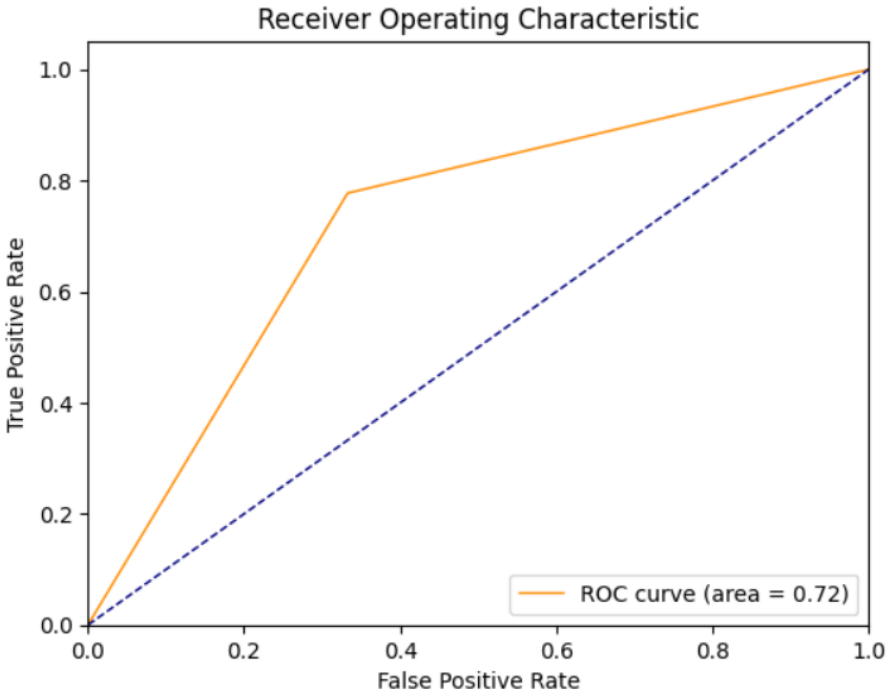

Figure 7: Receiver Operating Characteristic for the MLP Neural Network Model

From this image, we can observe, the area under the curve (AUC) is 0.72, indicating a good level of prediction accuracy.

6. Conclusion and Future Scope

In this paper, we proposed a deep learning method to predict cancer risk from gene symbols using a multilayer perceptron (MLP) feed forward neural network. We used a data set of gene symbols and their corresponding cancer risk labels, obtained from a DNA microarray analysis. We compared our MLP model with various machine learning models, such as logistic regression, linear discriminant analysis, quadratic discriminant analysis, decision tree classifier, gaussian nb, ada boost classifier, gaussian process classifier, support vector machine, and random forest. We found that our MLP model outperformed all the other models, achieving an accuracy of 72%. We also evaluated our MLP model using precision, recall, and F1-score metrics, and showed that it had a high performance in predicting cancer risk from gene symbols.

However, our work has some limitations and directions for future research. First, our accuracy score of 72% can be further improved by using more data, more features, or more complex models. Second, we can add attention mechanism to our MLP model, which can help the model focus on the most relevant gene symbols for each cancer risk label. Third, we can conduct more experiments on hyperparameter tuning and optimization, which can enhance the performance and robustness of our MLP model. We

hope that our work can inspire more research on applying deep learning methods to bioinformatics problems, and ultimately help in the diagnosis and treatment of cancer.

7. References

1. de Amorim, L. B. V., Cavalcanti, G. D. C. and Cruz, R. M. O. (2023) 'The choice of scaling technique matters for classification performance', *Applied Soft Computing*, 133, pp. 1–37. doi: 10.1016/j.asoc.2022.109924.
2. Chatra, K. *et al.* (2019) 'Cancer data classification using binary bat optimization and extreme learning machine with a novel fitness function.', *Medical & biological engineering & computing*, 57(12), pp. 2673–2682. doi: 10.1007/s11517-019-02043-5.
3. Chen, R. *et al.* (2020) 'Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data', *Bioinformatics*, 36(5), pp. 1476–1483. doi: 10.1093/bioinformatics/btz769.
4. Guia, J., devaraj, M. and Leung, C. (2019) 'DeepGx: deep learning using gene expression for cancer classification', in, pp. 913–920. doi: 10.1145/3341161.3343516.
5. Lee, M. (2023) 'Deep Learning Techniques with Genomic Data in Cancer Prognosis: A Comprehensive Review of the 2021-2023 Literature.', *Biology*, 12(7). doi: 10.3390/biology12070893.
6. Popescu, M.-C. *et al.* (2009) 'Multilayer perceptron and neural networks', *WSEAS Transactions on Circuits and Systems*, 8.
7. Ruby, U. and Yendapalli, V. (2020) 'Binary cross entropy with deep learning technique for Image classification', *International Journal of Advanced Trends in Computer Science and Engineering*, 9. doi: 10.30534/ijatcse/2020/175942020.
8. Shah, S. *et al.* (2020) 'Optimized gene selection and classification of cancer from microarray gene expression data using deep learning', *Neural Computing and Applications*, pp. 1–12. doi: 10.1007/s00521-020-05367-8.
9. Shi, Y. *et al.* (2019) 'Accurate and efficient estimation of small P-values with the cross-entropy method: Applications in genomic data analysis', *Bioinformatics*, 35(14), pp. 2441–2448. doi: 10.1093/bioinformatics/bty1005.
10. The Cancer Genome Atlas Program (no date) *No Title, 2019*.
11. Tufail, A. Bin *et al.* (2021) 'Deep Learning in Cancer Diagnosis and Prognosis Prediction: A Minireview on Challenges, Recent Trends, and Future Directions', *Computational and Mathematical Methods in Medicine*. Edited by I. Y. Liao, 2021, p. 9025470. doi: 10.1155/2021/9025470.
12. Wu, Q. *et al.* (2018) 'Deep Learning Methods for Predicting Disease Status Using Genomic Data', *Journal of biometrics & biostatistics*, 9.
13. Yousefi, S. *et al.* (2017) 'Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models', *Scientific Reports*, 7. doi: 10.1038/s41598-017-11817-6.
14. Zhang, T.-H. *et al.* (2022) 'Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions', *Cancers*, 14, p. 4763. doi: 10.3390/cancers14194763.