Task 03 & Task 04 - Data Science Project

Kartika Waluyo & Vrinda Rajendar Rajanahally

1000555 & 1129446

Task Description

Task 03 aims at selecting appropriate sample size for training, validation and test sets defined for the chosen tissues i.e. Lung, Skin - Not sun exposed, and Nerve - Tibial.

Additionally, Task 04 deals with the selection of modelling methods that can be used to predict normally distributed input.

Task 03: Defining the sizes of training, validation and test sets

The first half of this report will cover the approach and strategy behind choosing the appropriate proportions for the different sets. The type of sets involved in this project are:

1. Training set: the samples belonging to this set will be used to create the model. It is typically the set with the largest sample size.
2. Validation set: this set will be used to tune the parameters of the model and to further optimise it.
3. Test set: the model is finally run on this set, which it has never seen. It is used to assess the performance of the model and the accuracy of results.

After referring various sources and published papers related to topics surrounding prediction of gene expression, our approach to choosing the correct proportions for the sets was to ensure that each one of them has an ample amount of donors. Choosing an arbitrary number of 30 samples, our main aim is to ensure that the test set must have a sample size of atleast 30 donors overlapping with Whole Blood.

By the method of trial and error, random sets of proportions were tried and tested on the chosen tissues. The proportions that work the best for modelling the three chosen tissues are:

1. Training set: 0.7 or 70%
2. Validation set: 0.2 or 20%
3. Test set: 0.1 or 10%

Figure 1 is a bar plot that visualises the count of each set for the chosen tissues i.e. Lung, Nerve - Tibial, Skin - (Not sun exposed). We can see that for each tissue, the test set comfortably exceeds the count of 30 donors. Hence, we can conclude that 0.7-0.2-0.1 makes for most suited and desired proportion for the sets.

Additionally, Figure 2 visualises the count of each of the sets for the other tissues. We observe that although there are many other tissues that have a high count of samples in the test set like Artery - Tibial, these tissues either probably have lesser count of overlapping donors with Whole Blood or probably do not share a strong correlation with Whole Blood. As discovered in previous tasks, tissues with an overall low count of donors and count of overlapping donors, like Fallopian Tube, Cervix - Endocervix and Kidney - Medulla, have the least count in each of the sets.

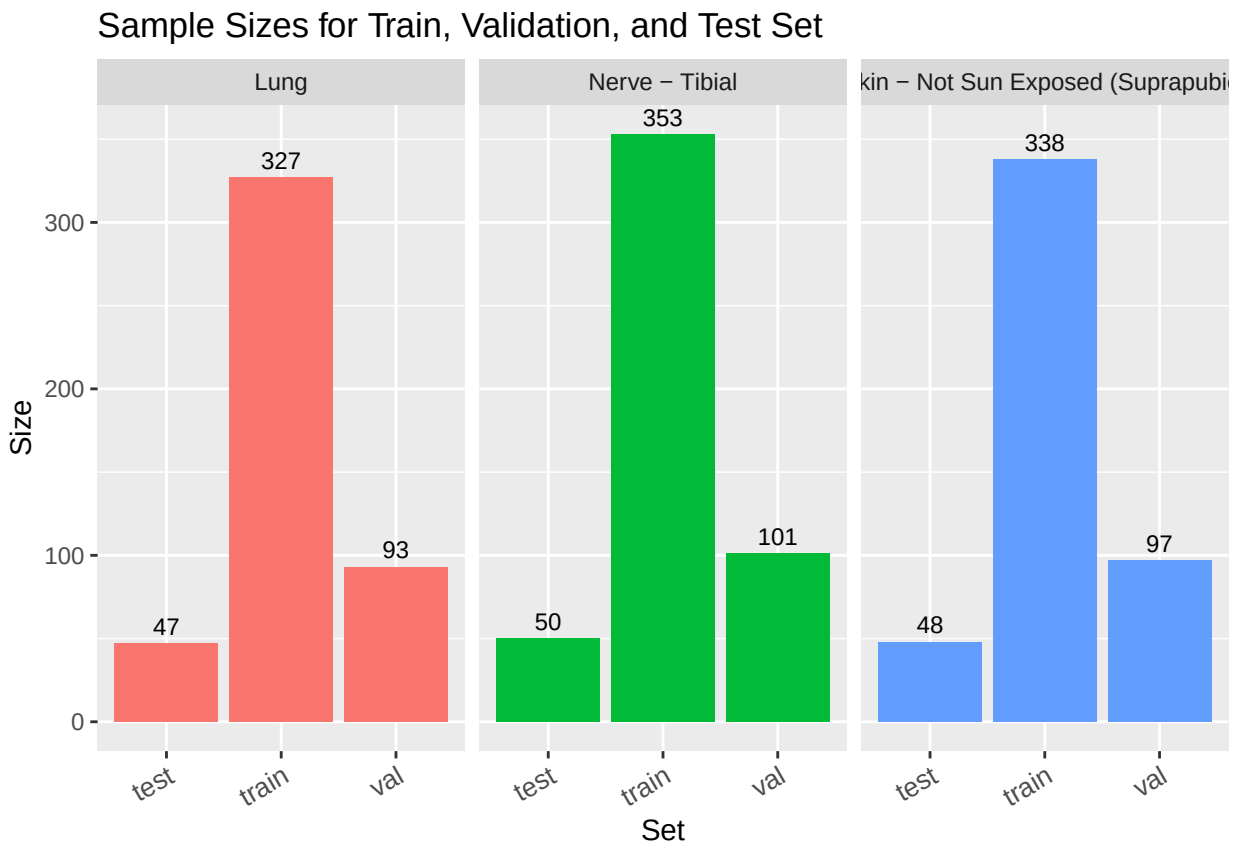

Figure 1: Count of samples per set for chosen tissues

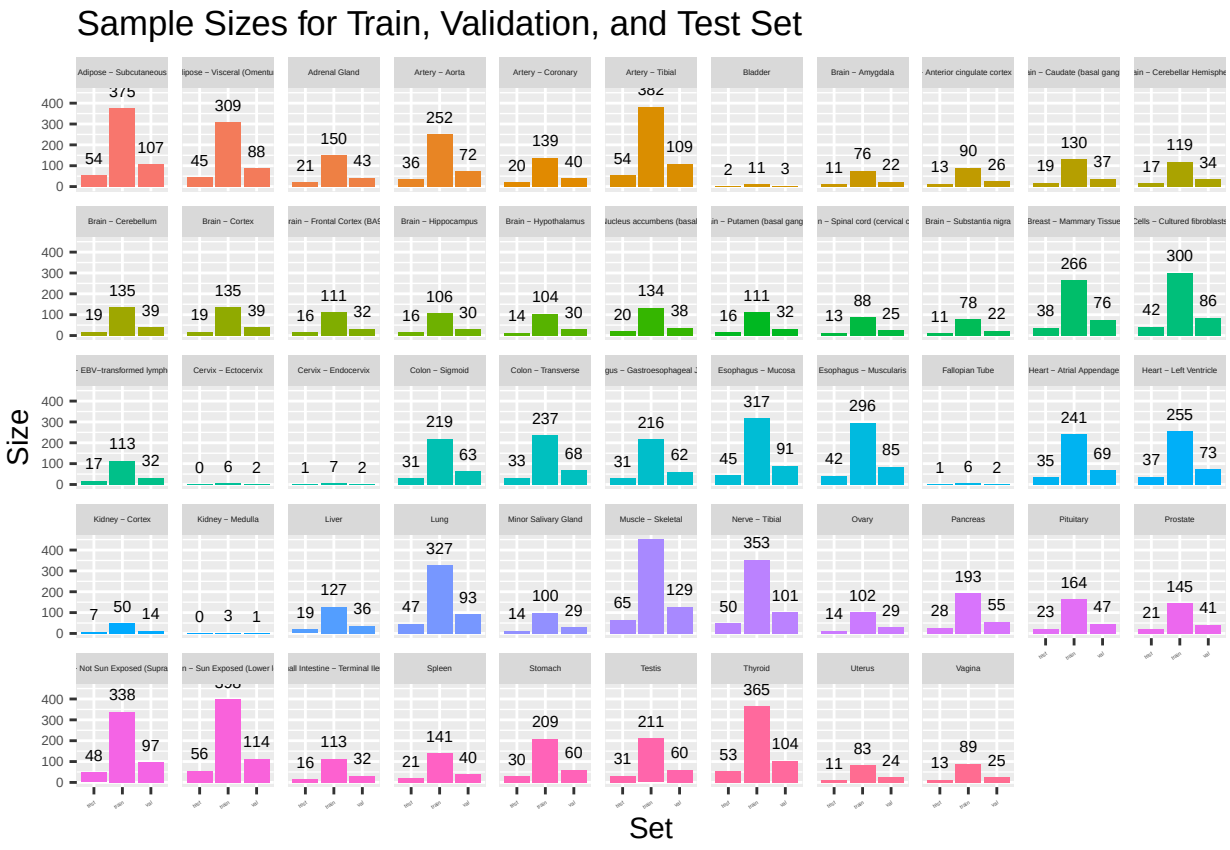

Figure 2: Count of samples per set for other tissues

Task 04: Choosing and comparing modelling methods

The the methods we wish to explore in this project are Neural Networks and Random Forests.

Neural Networks

Over the past few decades, neural networks modelling has been considered as one of the most powerful tools, and its ability to handle a huge amount of data made it very popular in the literature. Having deep hidden layers in the models has recently become an interest that has started to surpass classical methods performance in many fields, especially in pattern recognition.

Deep learning models have become popular in the bioinformatics field. Singh et al. (2016) used a unified CNN framework that automatically learns combinatorial interactions among histone modification marks to predict gene expression. Qi et al. (2012) used a deep multilayer perceptron (MLP) architecture with multitask learning to perform sequence-based protein structure prediction.

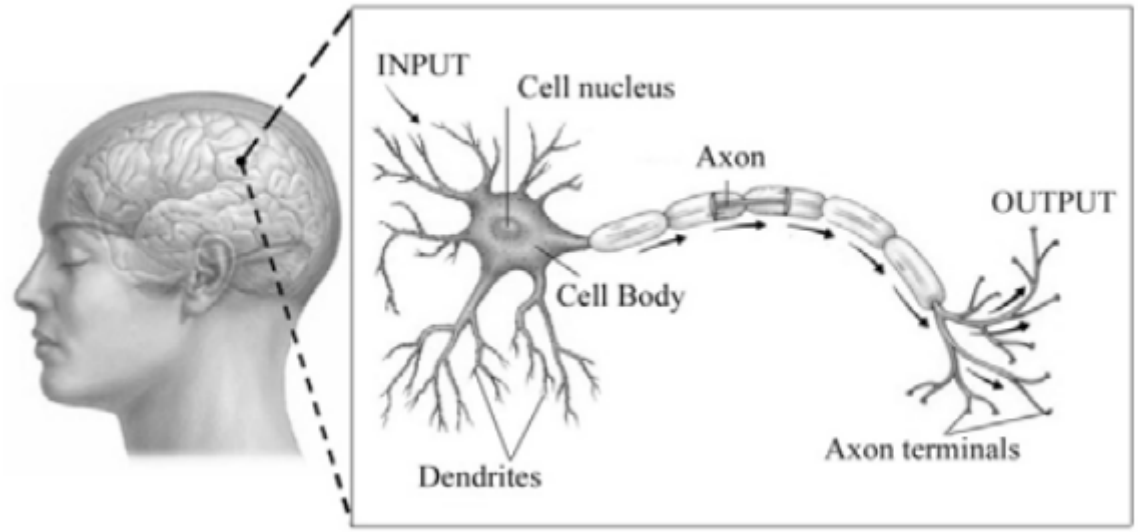

Figure 3: Neural networks in the human brain

The idea of neural network came from the most interesting organ in the human body, the brain. The human brain is made up of billions of basic units called neurons. Figure 3 illustrates the basic neuron unit. The neuron is made up of dendrites, a cell body and an axon connecting to axon terminals. The brain will receive information or inputs which are then transferred into the cell body through dendrites. The cell body works as the processing unit, where all the learnt information is then transferred into outputs and passed down by the axon. The muscles then receive the outputs from the axon terminals for actions. McCulloch and Pitts first studied this concept in 1943 to form a mathematical model.

Figure 4 shows a one hidden layer feed forward network with inputs $x_1, \ldots, x_i$, and output y_k. Each input has its own synaptic weight. The weights are then passed onto the hidden layer, which consists of several hidden neurons. A weighted summation of the inputs is performed by each neuron and then it passes a nonlinear activation function.

In our case, the input will be a matrix of the gene expression of Whole Blood and the output will be a matrix of the gene expression prediction of another tissue. The visualisation of the expected neural network model is shown in Figure 5. Note that the number of hidden layers is yet to be decided, and it is for visualisation

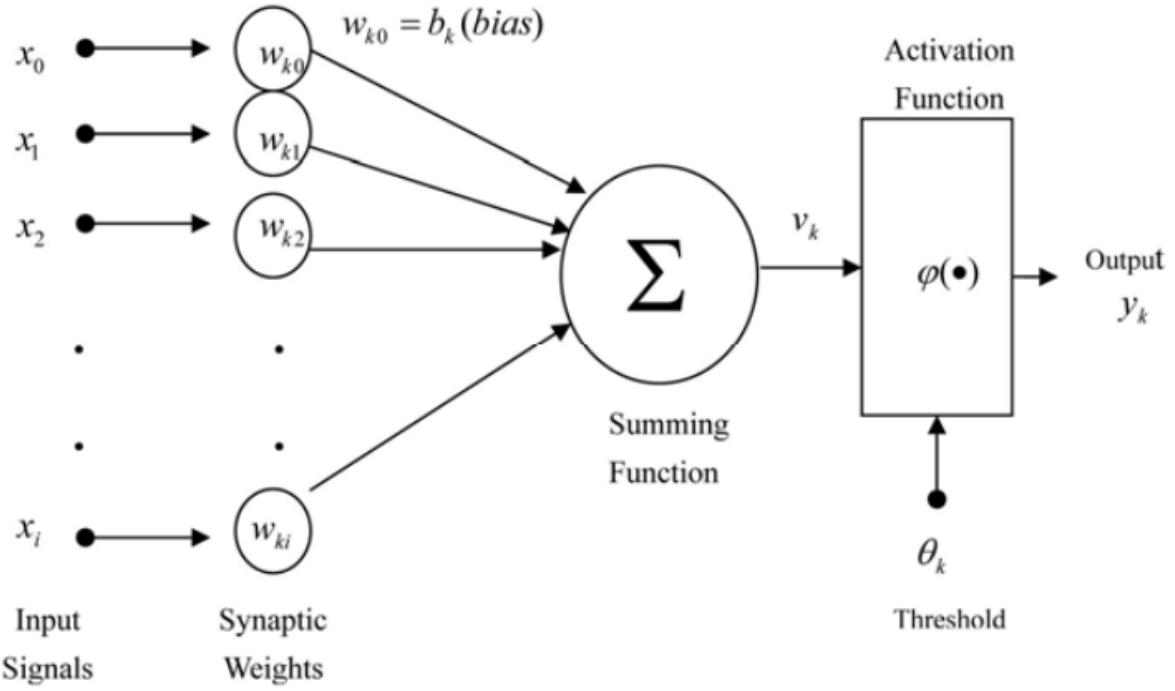

Figure 4: Hidden layers of a network

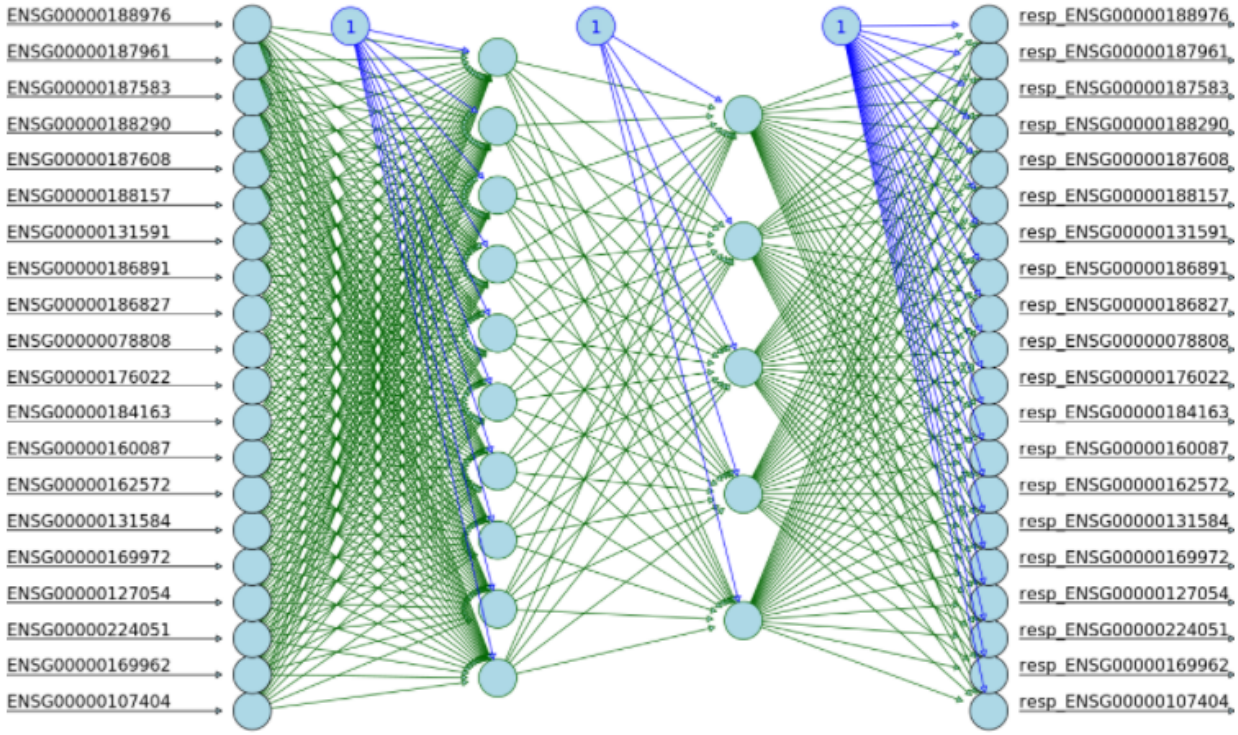

Figure 5: An snapshot of GTex data as a neural network

purpose only. Same thing applies for the number of input and output nodes. The number of input and output nodes in the real model will depend on the number of shared genes between Whole Blood and the other tissue.

Random Forests (RF)

Over the years, random forests has become a prominently used technique in the field of biology and bioinformatics. Some fields where random forests are used is in classifying different types of samples using gene expression, identifying diseases associated with particular genes and others.

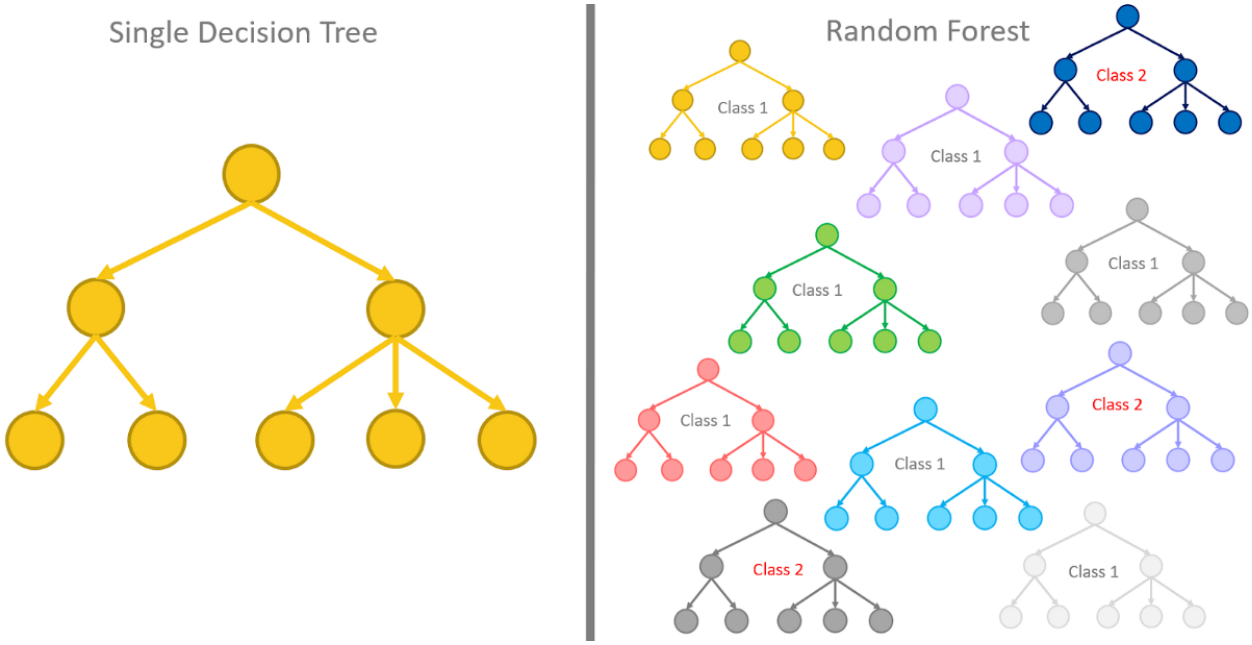

Figure 6: Single decision tree vs. Random Forest

A random forest is popularly known to be a classification algorithm, that selects features randomly. It also utilises the concept of bagging samples and majority voting scheme, that make it better than decision trees. Additionally, a random forest is a collection of many decision trees as seen in Figure 6, which is capable of classification and regression tasks.

Moving to the advantages of using the random forest technique, this machine learning algorithm works best in large and high dimensional data which is very well suited in the context of our project. A large random forest is preferable as it accounts for a robust model, with better accuracy and predictive capabilities. It focuses on selecting the best variable for prediction, and can easily help with identifying variables that aren't very significant to the model. Bootstrapping and ensemble schemes prevent the model from overfitting and hence, pruning trees is not required. Another advantage is that this algorithm accounts for missing values in the data and continues to maintain accuracy.

In the light of this project, the input will be a matrix of the gene expression of Whole Blood and the expected output is a random forest of the best predictors for gene expression in the other tissue. The model will be built using the training set and tuned for better results over the validation set.

Conclusion

On selecting the most appropriate proportions, the training, validation and test sets have an optimal sample size which can now be utilised in the modelling phase.

Neural networks and random forests are promising methods that can be used when the data is high dimensional and very large. It works out very well with the data being used in this project and has the scope to provide very promising results.

To conclude this task, our next approach is to use both modelling techniques, report on the findings in both and compare which modelling technique gives has better parameter selection, model accuracy and prediction power.