Results and discussion

1. **Performance of individual classifiers:**

   Accuracy is a good measure of performance of individual classifiers. The individual parameters were tested with the neural network as a classifier. The results are summarised in the following table :

Parameter	**Accuracy (%)**
PPI related parameters	76.83
Peptide length	77.42
Disorderness	77.50
Strand bias	77.50
Protein abundance	77.65
Expression Level	77.46
Codon usage bias	77.34
Rare codon usage	77.50
Complex number	83.7

Thus it is seen that the parameters are relevant for prediction of essentiality.

Further analysis with influence of individual parameters on the overall predictability was done with “attribute weights” operator.

The results can be summarised in the following table:

Attribute	**Weight**
Strand	0.1926
%dis	0.019779
Ax4	0.076166
dis_length	0.097052
Ax3	0.112364
Ax2	0.129
G3s	0.155629
Ax1	0.160791
Expression level	0.18762
pep_length	0.232864
Betweenness	0.336232
Fec value	0.351355
GC3s	0.353337
Nc	0.358917
Protein abundance	0.366428
Eigen	0.603723

k_nbor_all	0.665616
N-size2	0.687092
Closeness_all	0.780668
Degree_all	0.782022
N-size1	0.782022
Complex_number	1

Interpretation:

All the parameters we conceived as relevant were proven really influential to be able to predict essentiality. But the degree of influence on the prediction varied . The complex number parameter ranked highest meaning that Truly the proteins which are part of higher number of complexes are essential in nature. The network topological data were also found to be highly relevant for prediction of essentiality.The parameters like strand bias, disorderness fared less. This also could be due to the fact that strand bias is not very highly evident among essential and non-essential genes. We may need to look into the phenomenon of protein disorderness more thoroughly to extract the parameters which could be able to distinguish essential proteins in a better manner.

2. Holistic approach of prediction of essentiality

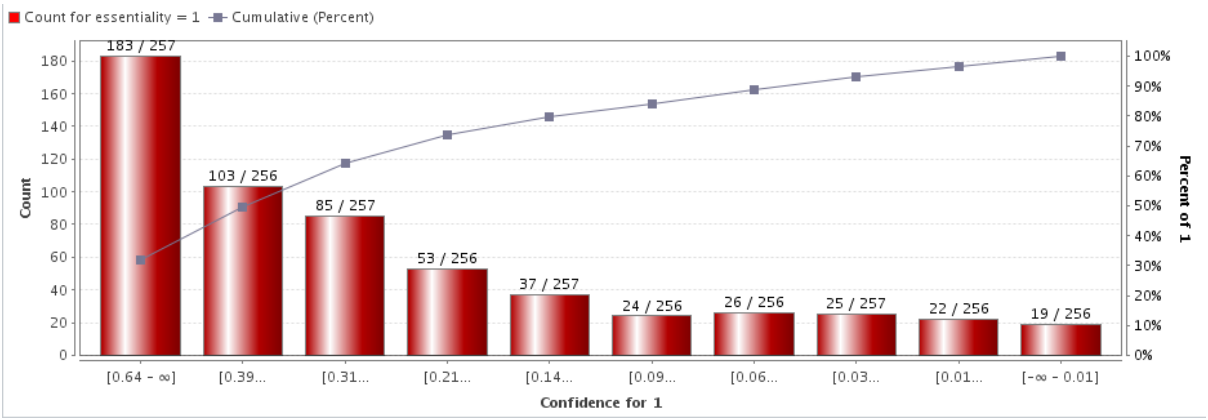
For the first time we integrated various kinds of features for prediction of essentiality and the result of the predictive modelling of he entire dataset taken together is given below.

measure	value
accuracy	72.50% +/- 3.09% (mikro: 72.50%)
classification_error	27.50% +/- 3.09% (mikro: 27.50%)
AUC (optimistic)	0.624 +/- 0.047 (mikro: 0.624) (positive class: 1)
AUC	0.624 +/- 0.047 (mikro: 0.624) (positive class: 1)
AUC (pessimistic)	0.624 +/- 0.047 (mikro: 0.624) (positive class: 1)
precision	36.29% +/- 6.80% (mikro: 36.15%) (positive class: 1)
recall	28.97% +/- 8.53% (mikro: 28.94%) (positive class: 1)
false_positive	29.500 +/- 9.124 (mikro: 295.000) (positive class: 1)
false_negative	41.000 +/- 7.510 (mikro: 410.000) (positive class: 1)
true_positive	16.700 +/- 5.158 (mikro: 167.000) (positive class: 1)
true_negative	169.200 +/- 11.241 (mikro: 1692.000) (positive class: 1)
specificity	85.16% +/- 4.45% (mikro: 85.15%) (positive class: 1)

The result denotes that the parameters that were considered in this research work successfully could distinguish between the essential and non-essential genes.

3. Lift chart:

The lift chart produced is given below



The high values of the chart clearly shows the efficiency of the predictability of the essentiality (coded as class 1). The chart proves that the predictive model considered here functions with much higher efficiency compared to the result obtainable without the model.

Future scopes

There is a huge scope of working with essential genes as for pathogens they are good drug targets and for humans they can lead insights into hereditary diseases. There is a huge scope of extending the work in many facets:

- The theories employed here is possible to be tested against various other microbial genomes.
- The parameters may be tested on Higher eukaryotes. It is not sure whether all parameters will function well, but the ones functioning well may denote the universal laws of biology. The ones not functioning well compared to the current study my throw useful insights into the evolution.
- When the input data when aggregated, there was no database of non-essential genes. So the current dataset assumed the candidates not included in the list of essential genes to be non-essential. This may not be true for all cases and thus the negative dataset may be mixed with false negatives. Now the database DEG hosts a list of non-essential genes. This system can now be made more consistent
- The Rapidminer system can be scaled to be put into a client server framework and hosted in the web. This may allow researchers to access the predictive modelling system online and perform their individual experiments.