# Intro

# Background

## Our ANN

There were a variety of factors that contributed to the type of ANN that we chose to use, it had to be repeatable, reliable, and simple enough to quantify the training difficulty in the time frame of this project. Due to these reasons we were unable to use pre-built libraries or ANNs for our testing as it was not possible to accurately quantify training difficulty, nor possible to ensure that it was not changing variables without us knowing.

The ANN we used was a stripped down version of what is commonly used today, it lacked features like a bias and the ability to set the eta due to the complexity of accounting for these things in quantifying the training difficulty. This meant that our network is more difficult to train and will have a lower accuracy compared to new networks, while this means that our results may not be valid for other type of ANNs and it is unlikely that they would be valid for other types of AI, due to time constraints further testing was out of scope of this project.

Our ANN was a multi-layer perceptron network in which could have zero to many hidden layers with any number of nodes in each hidden layer, we chose to test with a range of one to three hidden layers, and a variety of nodes on each layer to give a wide range of training results whilst being able to automate the training process.

The actual ANN we used is appended.

## Test data

The test data we used was opensource data that is freely available from the UCI Machine Learning Repository **Invalid source specified.**, we settled on 7 datasets: Card, Cancer, Diabetes, Iris, Abalone, Heart, and Dermatology.

When choosing our datasets we ensured we would have enough to make a preliminary analysis and we chose ones in which had a wide range in standard deviation to prevent un-usable data from being obtained at the end of the analysis.

## Training cost

Calculating the training cost was done using the following equations

where is the number of epochs and is the number of elements in the training data set.

Where is the number of nodes in hidden layer , is the number of hidden layers, is the number of inputs, and is the number of outputs.

## Standard deviations

Seven different types of standard deviations were calculated per dataset for use in our analysis, they are as follows

1. The standard deviation of the original data set as a whole (SD)
2. The standard deviation of the dataset of the normalised dataset (normalised to have values from 0 to 1) (SD of norm)
3. The standard deviation of the standard deviation of each column (SD of SD)
4. The standard deviation of the average of each column (SD or Ave)
5. The max standard deviation of the standard deviation of each column (max SD)
6. The min standard deviation of the standard deviation of each column (min SD)
7. The range of 5 and 6 (SD range)

# Qualitative results

Our fist analysis of the results was through qualitative means, essentially the data was put in the below table and graph, and it became clear that there was no obvious correlation.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | SD of norm | SD of SD | SD of Ave | max SD | min SD | SD range | TC |
| Card | 0.35470243 | 0.149396 | 0.228065227 | 0.499824331 | 0 | 0.499824 | 78660 |
| Cancer | 0.289513801 | 0.055529 | 0.074977122 | 0.364385716 | 0.173267415 | 0.191118 | 7231928 |
| Diabetes | 0.245370876 | 0.027912 | 0.19841826 | 0.198210469 | 0.117498658 | 0.080712 | 7403520 |
| Iris | 0.262508334 | 0.063358 | 0.017593727 | 0.317983642 | 0.180664296 | 0.137319 | 51412200 |
| Abalone | 0.266829983 | 0.064212 | 0.225434806 | 0.274080139 | 0.037015094 | 0.237065 | 772046352 |
| Heart | 0.395743582 | 0.152798 | 0.211655676 | 0.49947221 | 0 | 0.499472 | 3351786320 |
| Dermatology | 0.347119146 | 0.062894 | 0.191929185 | 0.386473535 | 0.151726736 | 0.234747 | 6200018704 |
|  |  |  |  |  |  |  |  |