1. How can each of these parameters be fine-tuned?

• Number of hidden layers: we can start with a small number of hidden layers and can gradually increase it by monitoring the accuracy of the model. We should also look for overfitting such as high training accuracy and low validation accuracy.

• Network architecture (network depth): the same method of optimizing the network architecture as optimizing the hidden layers can be applied. We start with a small number of hidden layer and gradually increase the number of hidden layer by monitoring the train and validation accuracy.

• Each layer's number of neurons (layer width): start with a small number and gradually increase it and ensure there is no overfitting in the model. We can use regularization technique such as dropout.

• Form of activation: different activation functions such as sigmoid, Tanh, Relu, Leaky Relu, PRelu can be used. Sigmoid and Tanh functions are not commonly used due to vanishing gradient problem where the weights stop updating after a certain point as the derivative of the activation function gives the value in a small range. ReLU activation function can be used, but it can also be replaced with its variants such as LeakyRelu or Prelu, if the ReLU doesn’t reach the convergence point due to dying ReLU.

• Optimization and learning:

Different optimization algorithms can be used such as gradient descent, SGD, mini-batch SGD or SGD with momentum or Adam optimizer. Adam Optimiser is proven to be the best one so far as it smoothens the path to reach global minima and the learning rate can be controlled in an attempt to reach the global minima faster.

• Learning rate and decay schedule

Learning rate: it is used to decide the step size the optimizer takes to decide the updated weight value. Adam optimizer allows you to control the learning rate in a way so that global minima can be found faster.

• Mini batch size: it is used in SGD where a certain number of inputs are fed into the network together as a batch so that the computational complexity can be decreased and convergence becomes faster as compared to SGD.

• Algorithms for optimization:

Different optimization algorithms can be used such as gradient descent, SGD, mini-batch SGD or SGD with momentum or Adam optimizer. Adam Optimiser is proven to be the best one so far as it smoothens the path to reach global minima and the learning rate can be controlled in an attempt to reach the global minima faster.

• The number of epochs (and early stopping criteria):

Early stopping as a callback can be used to find the correct epoch value as it stops training the model where the validation set accuracy starts to decrease.

• Overfitting that be avoided by using regularization techniques.

Early stopping or Dropout methods can be used to avoid overfitting in ANN.

• L2 normalization:

This is used with the loss function to avoid overfitting. It is a hyperparameter that controls the strength of the penality.

• Drop out layers:

It randomly drops some of the neurons to prevent overfitting. We can define the percentage of the neurons to be dropped.

• Data augmentation:

Data augmentation is a technique used to artificially increase the size of the training dataset by creating modified versions of the original data. It is a common technique used in deep learning to improve model performance and prevent overfitting.

Data size can be increased by adding some noise or rotating the images.