This is required for research projects and should be linked back to the project aim and objectives. It should describe the research methods that will be employed in the project and the research questions that will be investigated.

0.1 Research questions

The following are the research questions this research will seek to cover:

- Application of pruning algorithms on neural networks: how does the compression paradigm impact latency?
- Can we apply hyperparameter optimisation methods to neural network compression to minimize latency?
-

0.2 Preliminary Evaluation

To demonstrate the necessity of objective O0, this section presents findings from a series of preliminary benchmarks. According to the literature covered in section **??** course-grained pruning algorithms should provide a demonstrable improvement in latency during inference. Likewise quantisation is an even more consistent in its ability to reduce inference latency (see section **??**).

Table 1 presents findings when benchmarking inference of resnet20 with the CIFAR10 dataset on the NCS (table ref), these results show no real change between compression methods, this is an unexpected result.

Compression algorithm	Top 1 Accuracy	Top 5 Accuracy	Latency (ms)	Throughput (FPS)
N/A (baseline)	91.120	99.660	10.19	392.22
AGP filter, fine-grained, and row pruning	91.110	99.700	10.15	394.14
ssl channels removal	91.610	99.780	10.17	389.17

Table 1: Preliminary NCS inference results, Resnet20 trained and tested with the CIFAR10 dataset.

Investigation X in section Y will investigate this further with the aim of verifying proper application of the compression scheduler to the model.

0.3 Research methodology

0.3.1 Compressing

Distiller LR scheduling

0.3.2 Benchmarking

Openvino benchmarking tool

0.3.3 Optimisation

bayesian optimisation [1]