Project Idea:

Profiling and Optimising Energy Consumption for TF Lite Models.

Proposal.

There is a spurt of pretrained edge machine learning models that can be quickly deployed in microcontrollers, thanks to TensorFlow Lite. Reducing RAM usage and increasing battery life are a critical aspect in edge ML models. The core idea of our proposal is to prepare Tensorflow Lite libraries/examples that would reduce energy consumption based on machine learning architecture optimisations. As an example, we created a video that demonstrates how simple machine learning and microprocessor peripheral optimisation can reduce the current consumption of the *Wake Word detection* model in the Arduino Nano 33 BLE Sense microcontroller. These optimisations show that the *Wake word detection* model can run for upto 12 days with a 240 mAh coin cell battery.

Our quick experiments have shown that the *Wake Word Model* can consume up to 25% less flash memory with a minimal drop in accuracy. The experiments are summarised in **Table 1**. In the proposed work, we also aim to explore pruning measures that can reduce model size.

Table 1: Performance of wake word detection model on different quantisation techniques. Current consumption as observed in the Arduino Nano 33 BLE Sense board.

Techniques	**Accuracy**	**Model Size**	**Current Consumption**
Baseline (TensorFlow Float32)	89.0%	11.7 KB	?
Dynamic Range Compression (DRC)	94.7%	11.9 KB	?
Float16 Quantisation	94.7%	9.7 KB	9.7 mA (Classifier) 4.4 mA (Sampling)
Full Integer Quantisation*	88.6%	8.7 KB	6.5 mA (Classifier) 4.4 mA (Sampling)

*Implemented in Edge Impulse; **?** These models could not be loaded in the Arduino Nano 33 BLE Sense. The baseline model is not a TF Lite model. The TF Lite DRC model has some issues. The Arduino library throws incompatible architecture errors.

Our suitability for these projects.

Our group has worked on signal processing, applied machine learning and embedded systems to sense health metrics and compress neural networks that can run on edge devices. We have been working on our research named, *SpiroMask*. In *SpiroMask*, we have shown that a user can perform Spirometry using any consumer-grade mask. *SpiroMask* also makes it possible to monitor respiration rate continuously. Our group members have demonstrated how neural networks can be compressed with invariant accuracy.

Summary

In summary, we would like to propose helping via the GSoC for the following:

1. Optimise neural network architecture of the 'Wake Word Detection' model to reduce its energy consumption using quantisation and pruning.
    a. Given the popularity of *the* '*Wake word detectio*n' model as a first model of choice for beginners getting started with Tensorflow Lite in Microcontrollers, we would like to add energy-saving guides and libraries to this repository.