All Roads Lead to TinyML: The Rome of Efficient Machine Learning in Engineering

Dinuka Sahabandu[1], Nushara Wedasinghe[2], Madusha Weerasooriya[2], Asiri Gawesha Lindamulage[3*], Sanka Mohottala[2*]

1) University of Washington, USA 2) Sri Lanka Institute of Information Technology, Sri Lanka
3) Open University of Sri Lanka, Sri Lanka

*sanka.mo@sliit.lk

TinyML Website – All workshops in one place

https://tinyml-in-action.github.io

Organizing Committee

Dr. Dharshana Kasturirathna
SLIIT

Dr. Dinuka Sahabandu
University of Washington

Dr. Mahima Weerasinghe
SLIIT

Dr. Nushara Wedasinghe
SLIIT

Mr. Asiri Gawesha Lindamulage
Open University of Sri Lanka

Mr. Sanka Mohottala
SLIIT

Ms. Madusha Weerasooriya
SLIIT

Ms. Savini Kommalage
SLIIT

Organized by BrAINLabs Research Group, SLIIT and Faculty of Engineering, SLIIT
Funded by SLIIT Research & International (Grant No. PVC(R&I)/RG/2025/12)

Resource Personal

Dr. Dinuka Sahabandu
Department of Electrical and Computer Engineering
University of Washington

Dr. Nushara Wedasinghe
Department of Electrical and Electronic Engineering
Sri Lanka Institute of Information Technology

Mr. Asiri Gawesha Lindamulage
Department of Electrical and Computer Engineering
Open University of Sri Lanka

Mr. Sanka Mohottala
Department of Electrical and Electronic Engineering
Sri Lanka Institute of Information Technology

Ms. Madusha Weerasooriya
Department of Software Engineering
Sri Lanka Institute of Information Technology

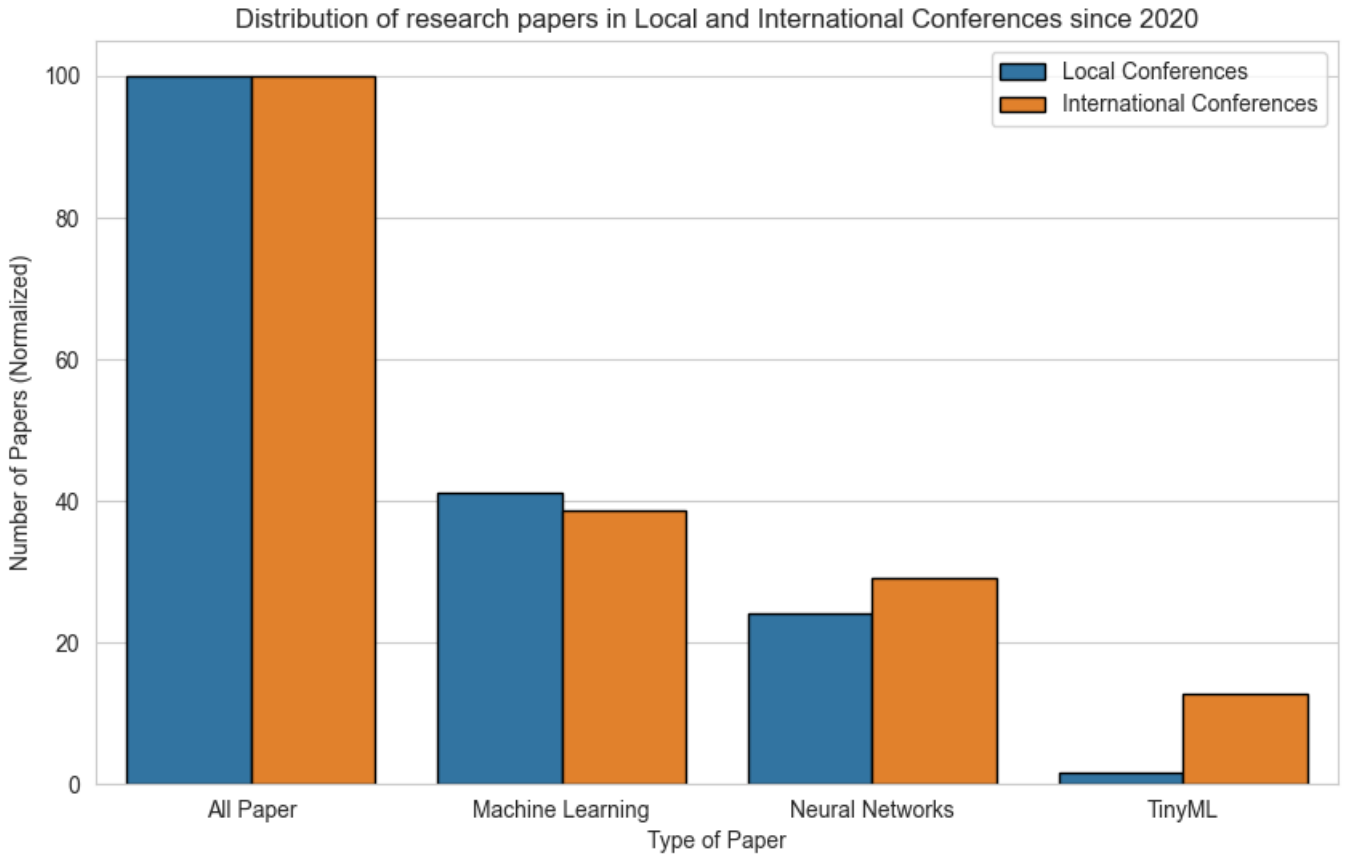
Motivation behind the workshop

In Sri Lanka awareness on TinyML is very lower than other areas.

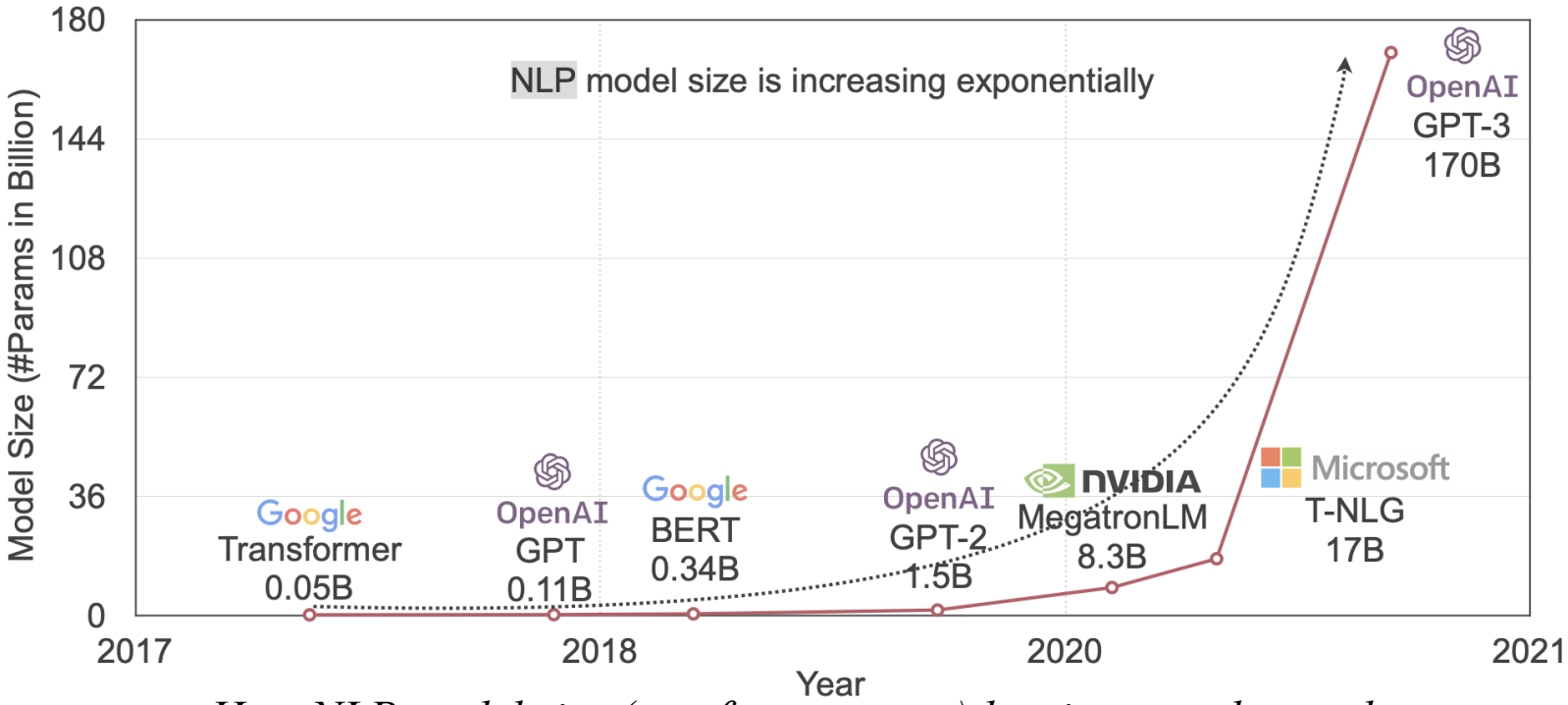
TinyML offers a novel, experimental research direction.

The workshop will equip participants to balance accuracy, latency, and memory usage, optimizing models for real-world deployment.

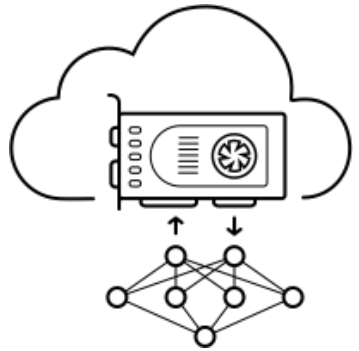
Workshop consists of theoretical sessions and coding sessions



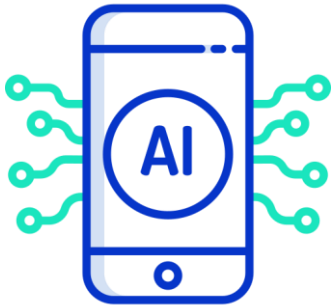
What is TinyML and why is it important



How NLP model size (no of parameters) has increased over the years



On Cloud



Edge Computing

- Billions of IoT devices around the world based on **microcontrollers** – CCTV, Smart wearable devices
- **Low-power**: green AI, reduce carbon

Challenge - Small memory (Around 1 MB), lower computational power

Tiny ML

TinyML Landscape

Converting large deep learning models into compact models for heavily resource constrained devices

- Reduce the number of parameters
- Lower the peak memory usage
- Reduce power consumption

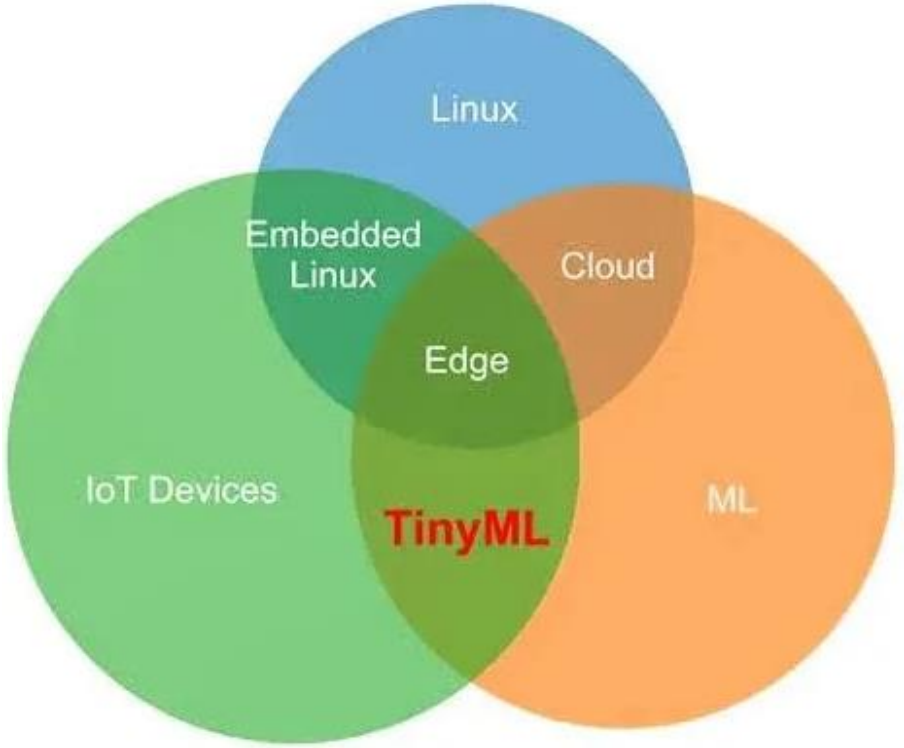
What are the techniques?

Model compression

- Pruning
- Quantization
- Weight clustering
- Knowledge Distillation

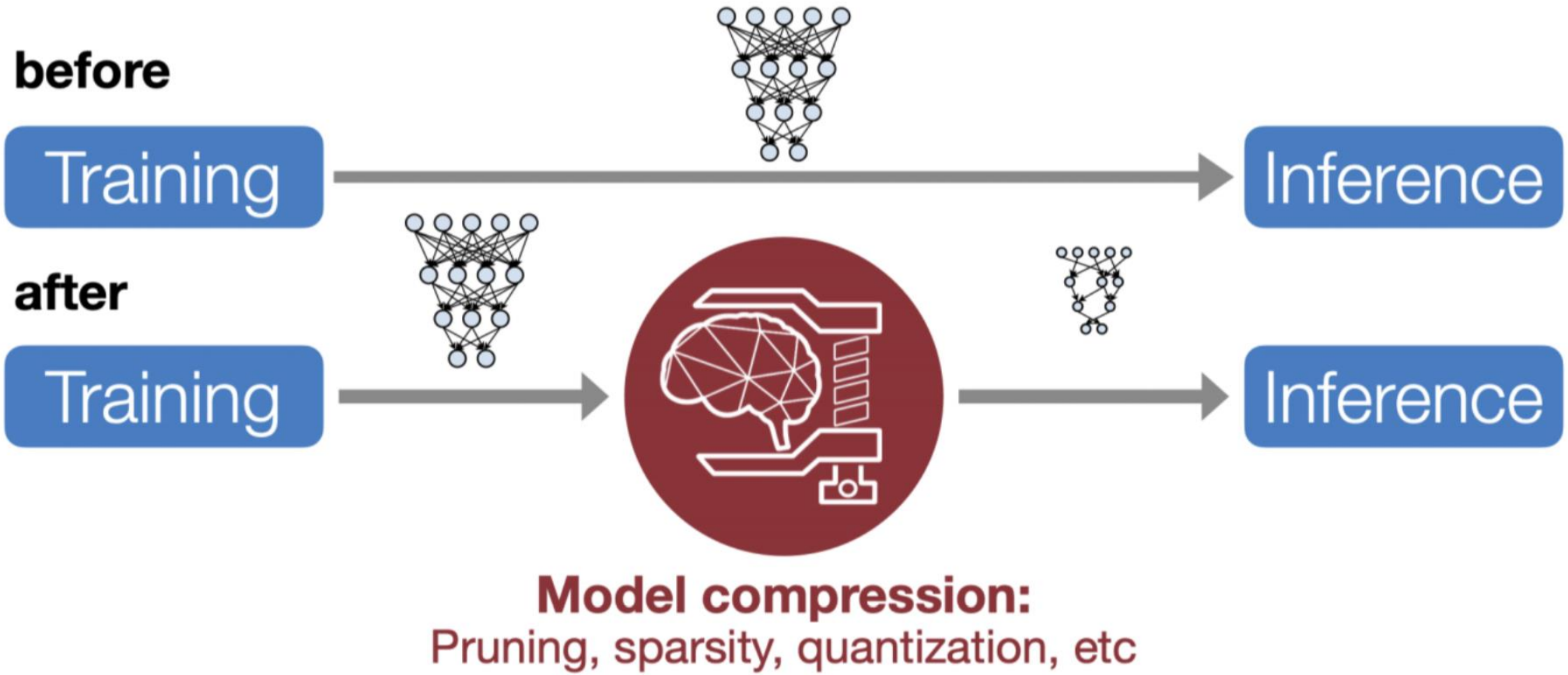
Tiny models

- MobileNet, EfficientNet
- Bio-inspired architectures
- Paramter efficient architecures

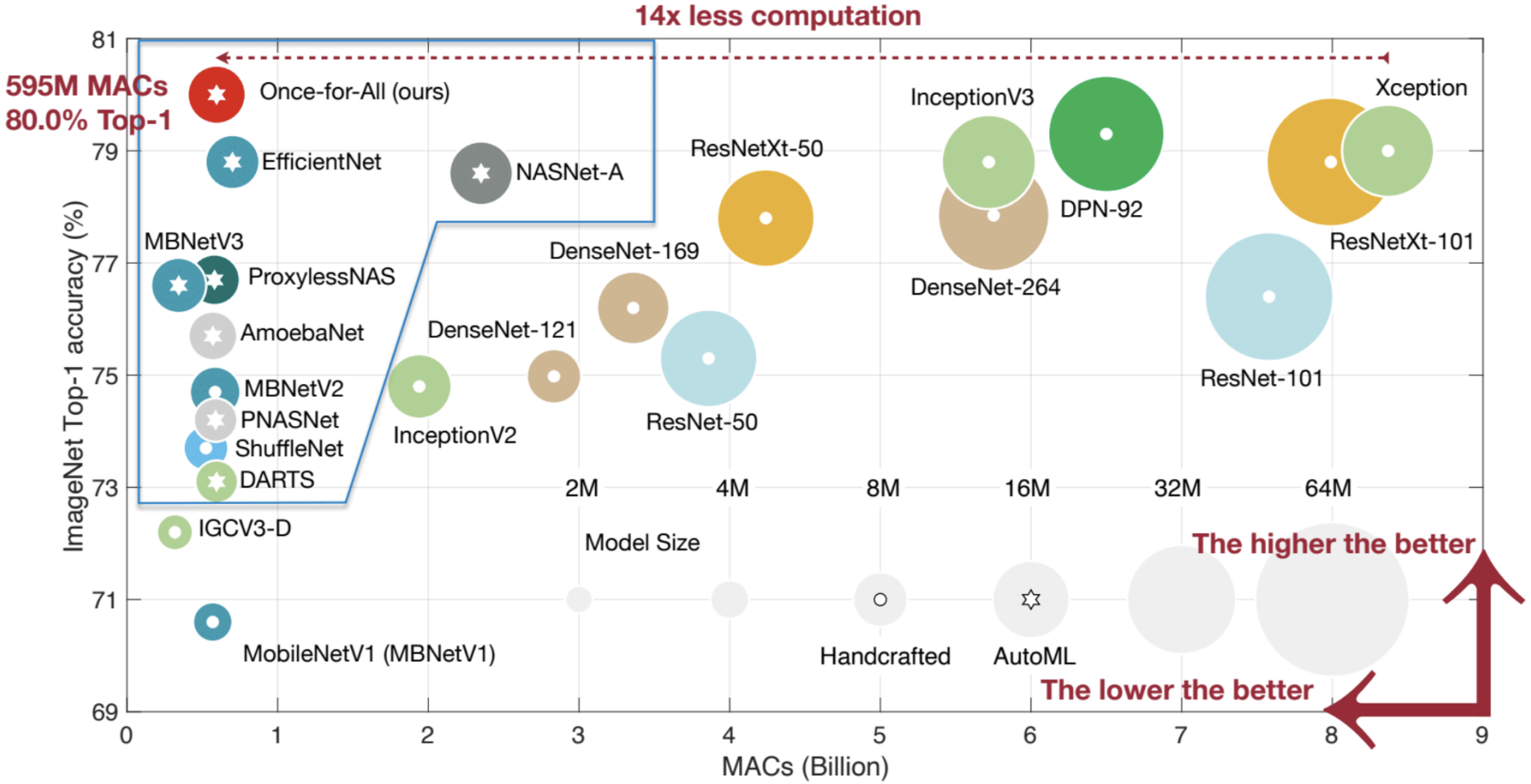


Model Compression

- Bridges the Gap between the Supply and Demand of AI Computing



Tiny Architectures



Once-for-All: Train One Network and Specialize it for Efficient Deployment [Cai *et al.*, ICLR 2020]

NAS has resulted a highly accurate model with less latency and and energy usage
MobileNet uses – Separable Convolution and Point-wise Convolution

MACs- Multiply–Accumulate operations

What you are going to learn

- Theoretical aspects
  - Model Compression
  - Bio-inspired TinyML architectures
  - Energy efficient and parameter efficient models
- Hands on sessions
  - End-to end model deployment on RaspberryPi with model compression
  - Model deployment on Arduino Nano BLE 33
- Technology stack
  - Languages – Python, Arduino
  - Frameworks- TensorFlow, TFLite, TfLiteMicro

All notebooks and slides will be shared through the website

By the end you will have an idea on..

- How model compression can be utilized effectively
- Model deployment process and pipeline
- Theoretical knowledge on Bio Inspired energy efficient architectures
- An example for Tiny architectures for head pose estimation
- How to develop a model, optimize it for Tinyml and deploy it

Lineup of the workshop

Starting time	End time	Duration (mins)	Topic	Presenter
1:00:00 PM	1:05:00 PM	0:05	Intro to panel	Dr Dharshana Kasthurirathna
1:05:00 PM	1:15:00 PM	0:10	Introdution to the session	Mr. Asiri Gawesha
1:15:00 PM	2:00:00 PM	0:45	Wet TinyML	Dr. Samitha Somathilaka
2:00:00 PM	3:30:00 PM	1:30	Model compression techniques	Dr. Dinuka Sahabandu
3:30:00 PM	3:40:00 PM	0:10	Interval	
3:40:00 PM	4:40:00 PM	1:00	Coding Session on Model Compression Techniques	Mr. Asiri Gawesha
4:40:00 PM	5:10:00 PM	0:30	Efficiency through architectural improvements	Mr. Sanka Mohottala
5:10:00 PM	5:30:00 PM	0:20	Energy Efficient Architectures	Dr. Mahima Weerasinghe