

tinyML[®] Summit

Miniature dreams can come true...

March 28-30, 2022 | San Francisco Bay Area



www.tinyML.org

Optimizing AutoML for the tinyML Future

Elias Fallon



T I N Y

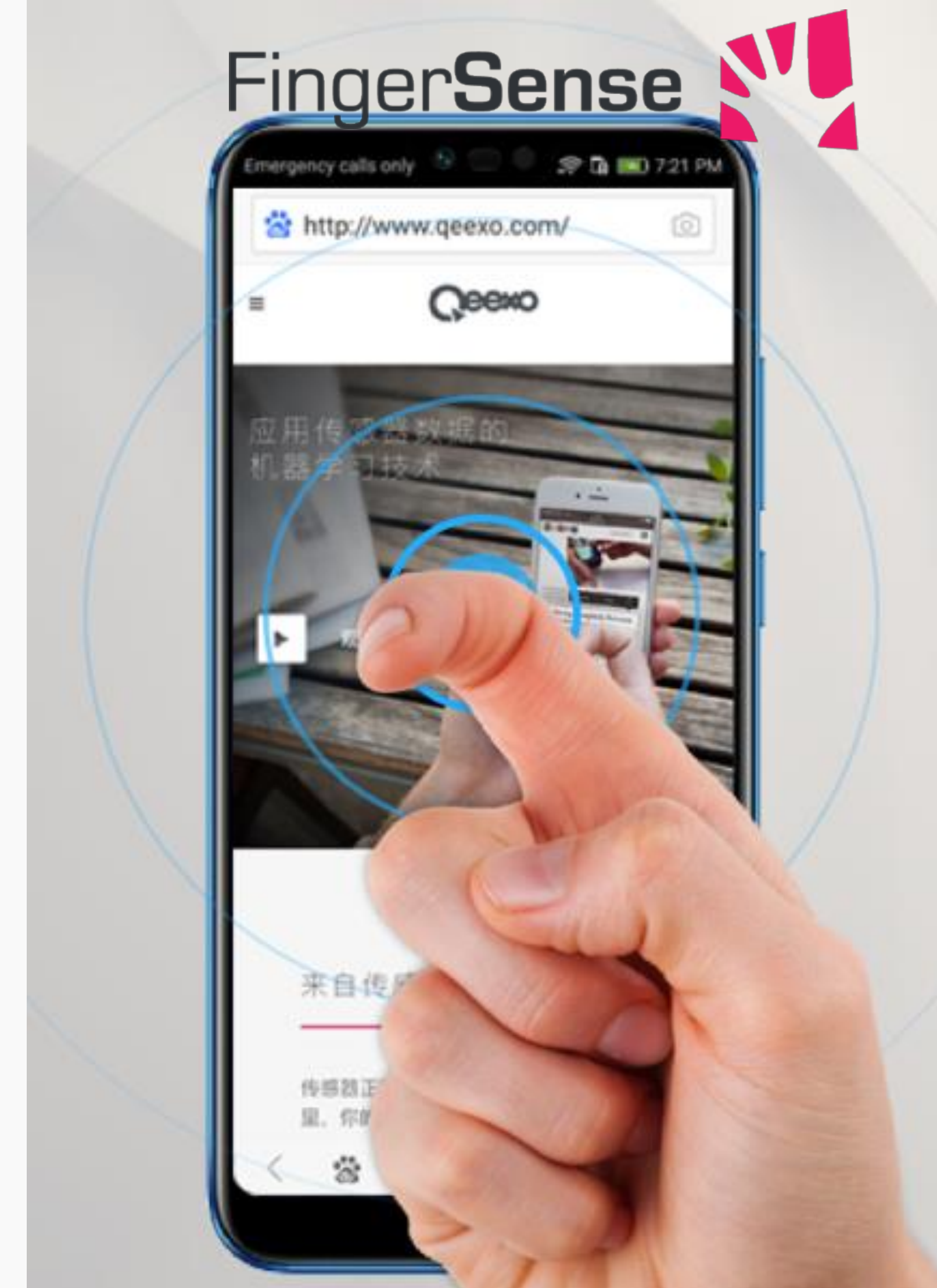


Summit 2022



Introduction

- What is AutoML for tinyML?
- How can AutoML unlock the unique power of tinyML hardware?
- Demonstrate the power of no-code AutoML for tinyML applications



- FingerSense has been in market since 2015
- Shipped on 400+ million devices
- Deployed on 260+ different smartphone variants



What is AutoML?

“Automated machine learning (AutoML) is the process of automating the tasks of applying machine learning to real-world problems. Each of these tasks may be challenging, resulting in significant hurdles to using machine learning. AutoML aims to simplify these steps for non-experts.”

Wikipedia contributors. (2022, January 29). Automated machine learning. In Wikipedia, The Free Encyclopedia. Retrieved 19:28, March 1, 2022, from https://en.wikipedia.org/w/index.php?title=Automated_machine_learning&oldid=1068650058

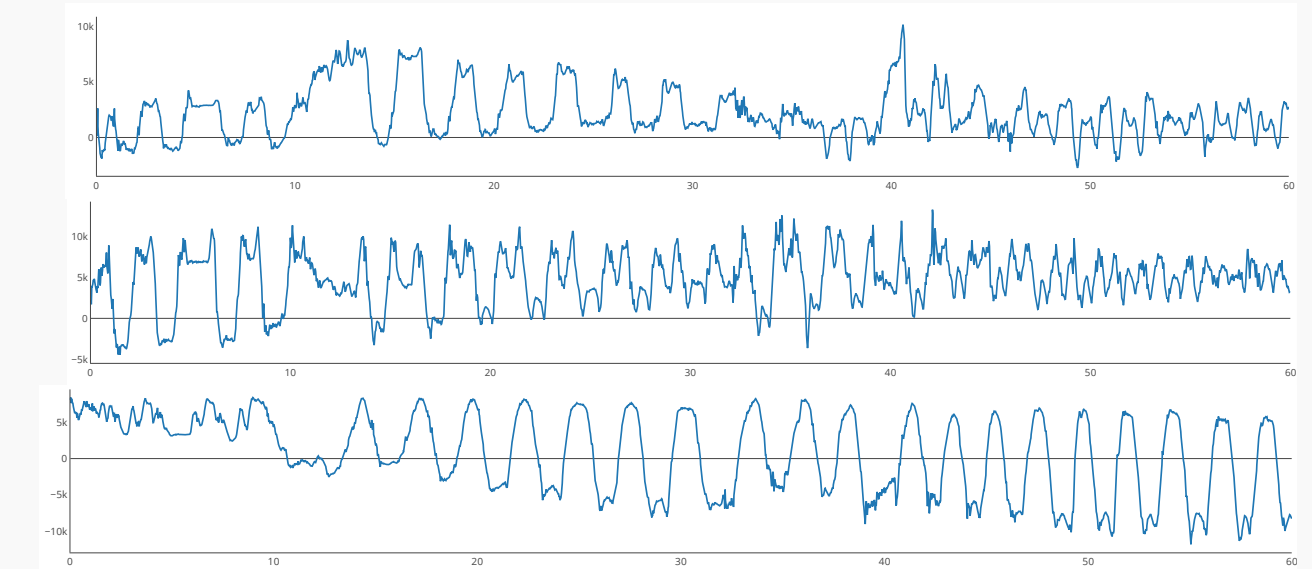
AutoML for tinyML

■ Basic AutoML Capabilities

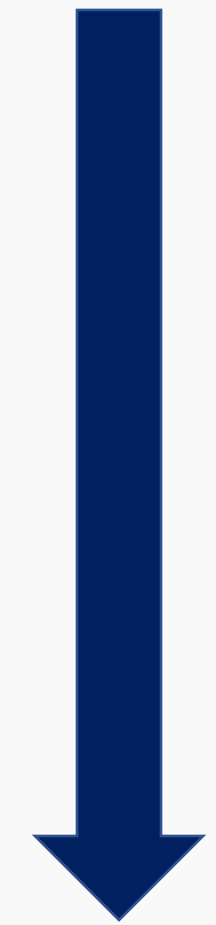
- Hyperparameter optimization: Find the best options for the training process to get the best accuracy out of models
- Model Options: Find the best options on the model (size, shape, activations) to get the best accuracy
- Model Architecture Search: What type of model best solves the problem to maximize accuracy

■ Data Pre-processing and Feature Extraction

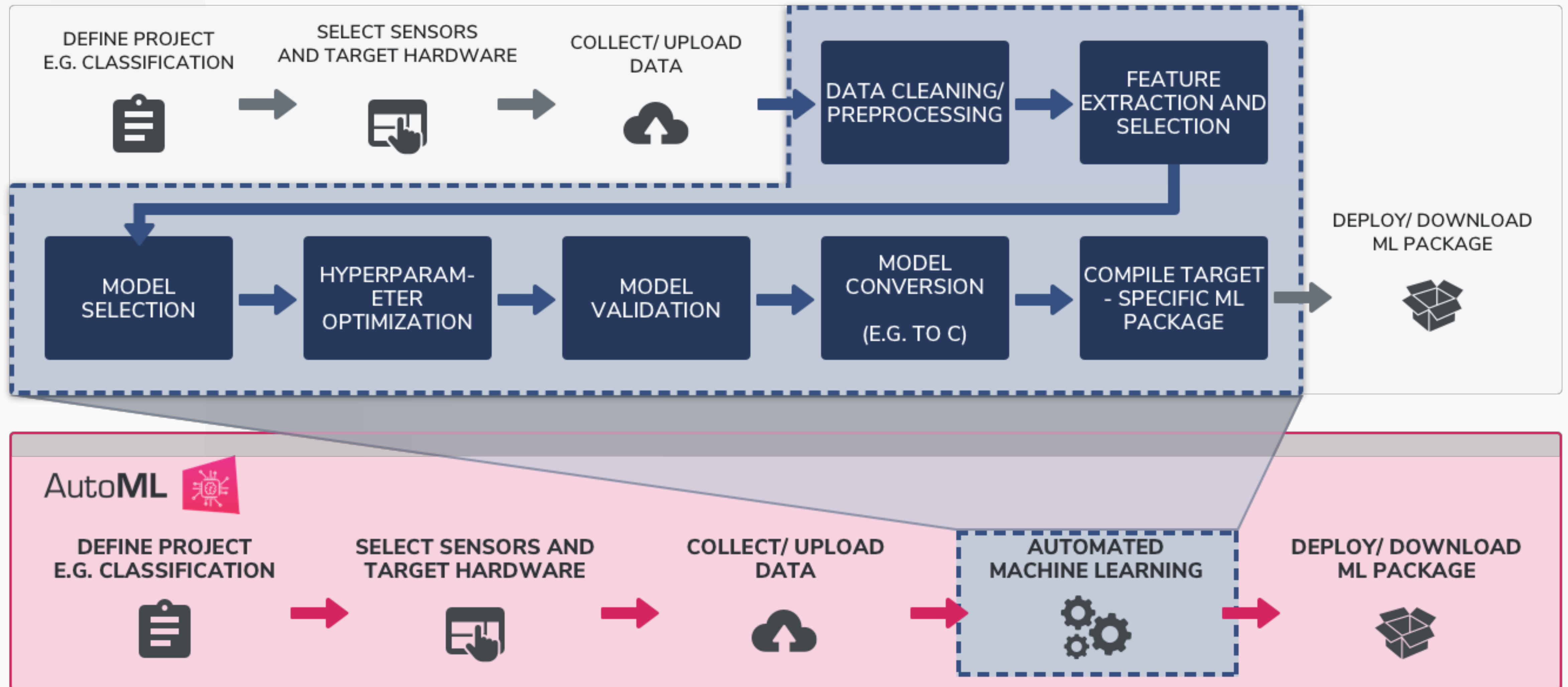
- Raw sensor data may need cleaning or assisted labeling
- Some models may need feature extraction to perform best. AutoML can select which feature extraction transformations result in the best model performance.
- Data and features will impact model performance and model selection may impact which pre-processing and feature extraction is needed. Having full flow of data collection through feature extraction and model selection and training in a common flow can greatly improve overall productivity and performance



Data

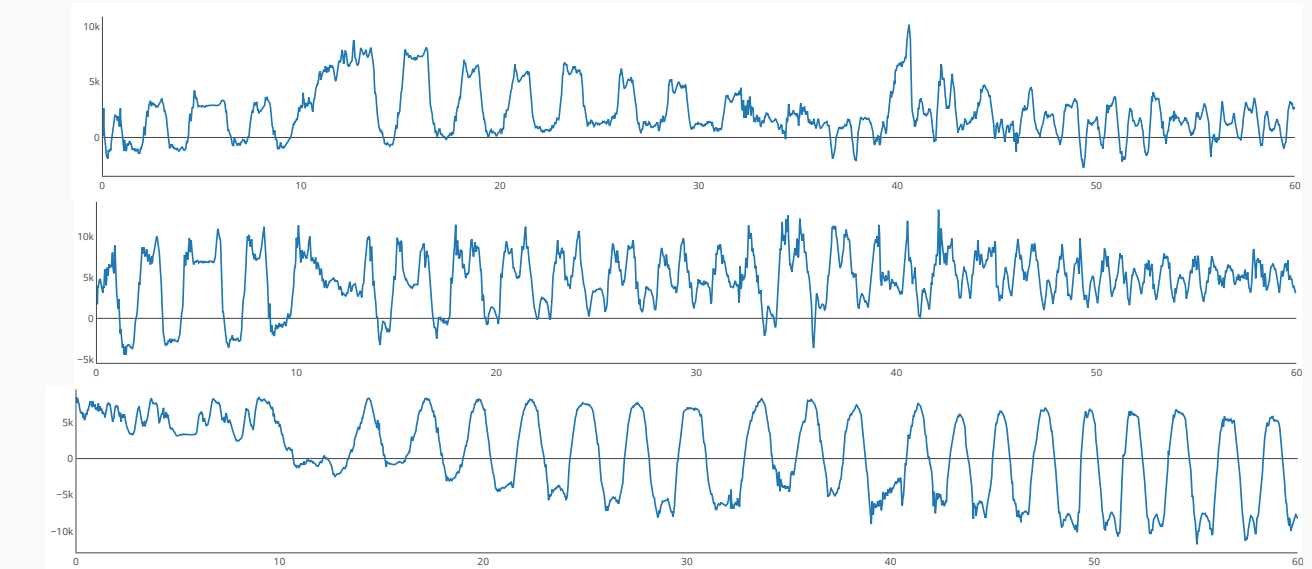


Fully Automated, Behind-the-Scenes Machine Learning

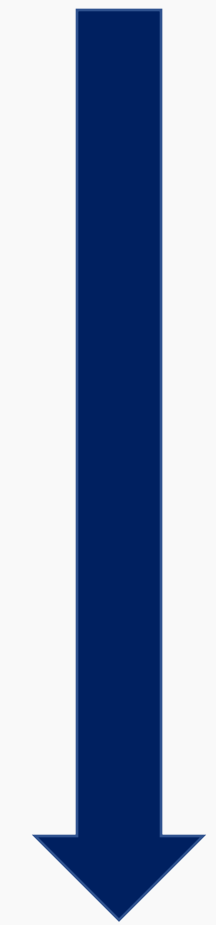


Hardware-aware AutoML

- **Model pruning, quantization, optimization**
 - Model Architecture Search: What type of model best solves the problem and minimizes on-device memory, latency, power and maximizes accuracy
 - Processors on tinyML hardware may not support all capabilities of desktop computers
 - E.g., no GPU, no floating-point unit, limited memory
 - Techniques like model pruning, quantization, and hardware-aware optimization can allow model to run well on target hardware
 - These techniques still typically assume a traditional “Von Neumann” compute architecture (CPU and memory)

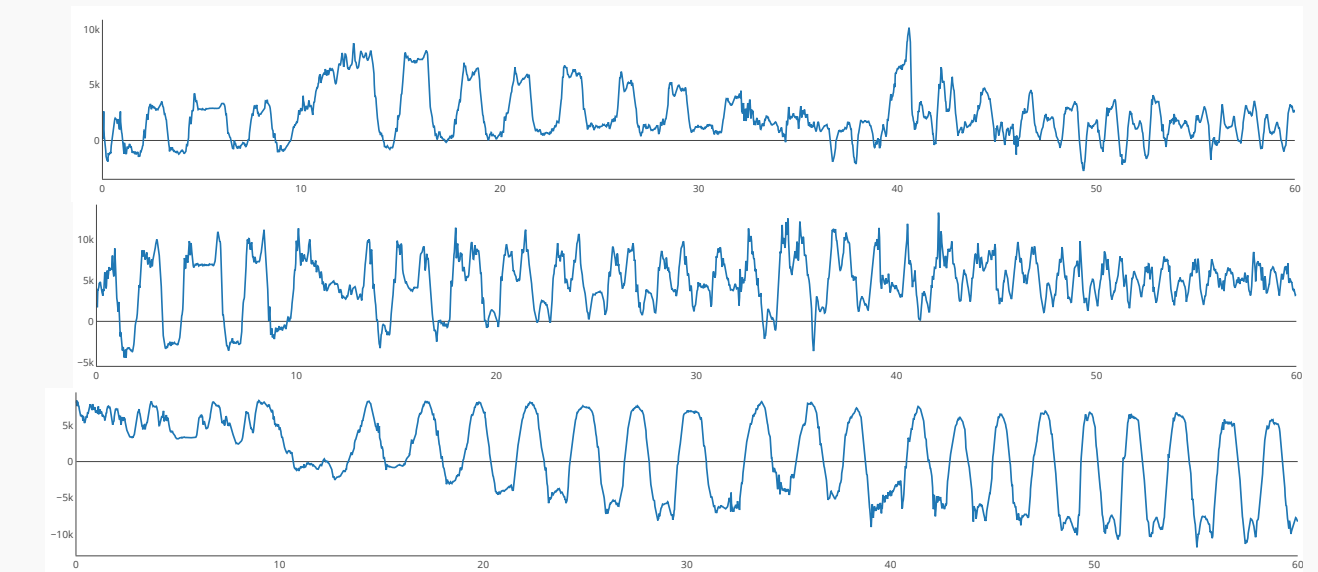


Data

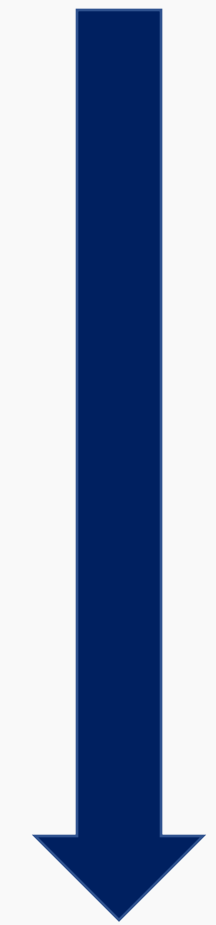


Hardware-enabling AutoML

- **Targeting specialized hardware**
 - More and more machine learning- and neural network- specialized chips are being developed and productized
 - Typically has a custom tool chain to go from standard representation (e.g., ONNX) to output ready to configure the hardware
 - *Goes from neural network accelerators...*
 - E.g., Arm's Ethos-U, requiring Tensorflow Lite Micro + Vela Compiler
 - *...to extremely low-power machine learning model accelerators adjacent to sensors*
 - E.g., ST Microelectronics Machine Learning Core [MLC] chip + Unico software
- **Full AutoML Solutions**
 - Need to be aware of hardware capabilities and limitations when making other AutoML decisions
 - Which model type/size works best is different depending on the target hardware

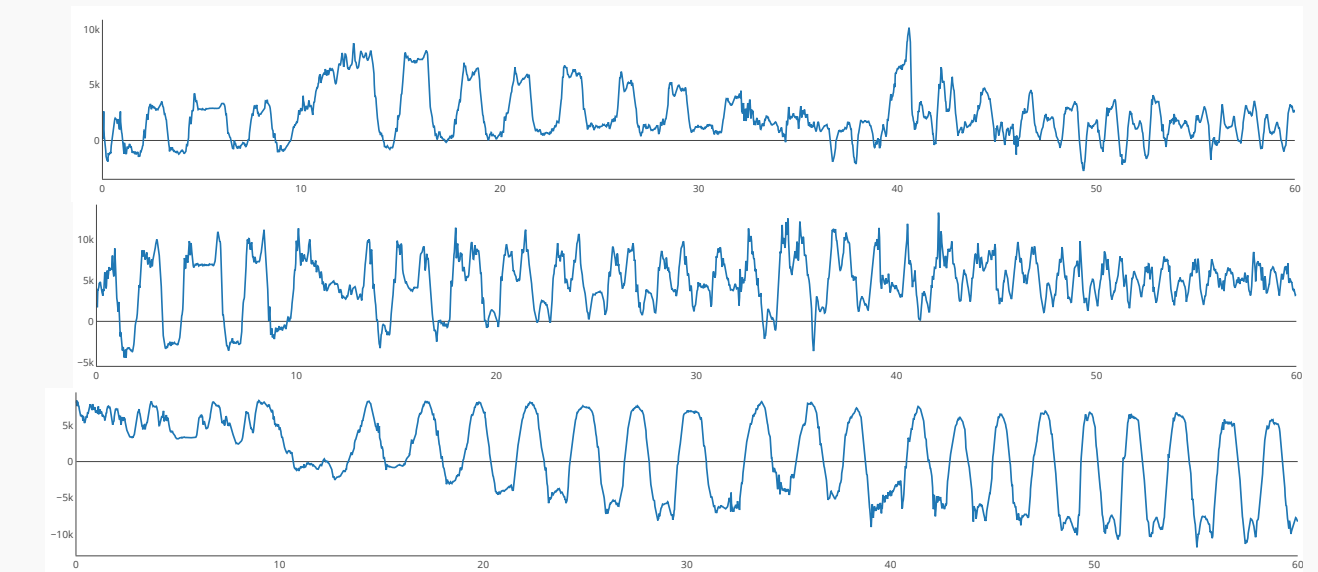


Data

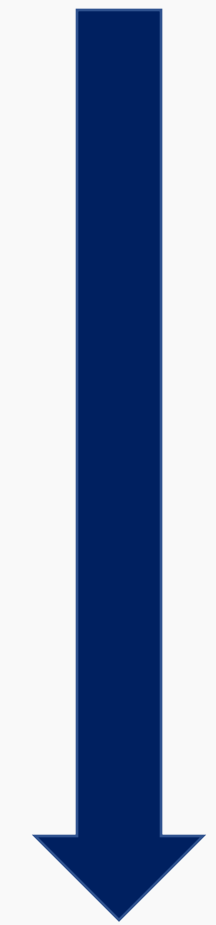


Target application

- **Raise-to-Wake Activity detection**
 - Use accelerometer and gyroscope to recognize specific “wrist raise to talk” action
 - Collected only 6.5 minutes worth of data to train the model
 - When “wake action” is detected by ML model, trigger additional application: e.g. microphone for recording
- **Use AutoML to find the best model**
 - Initially evaluate using STWIN MCU (Arm Cortex M4)
 - Evaluate tradeoff between models



Data



AutoML in action

Algorithm Selection

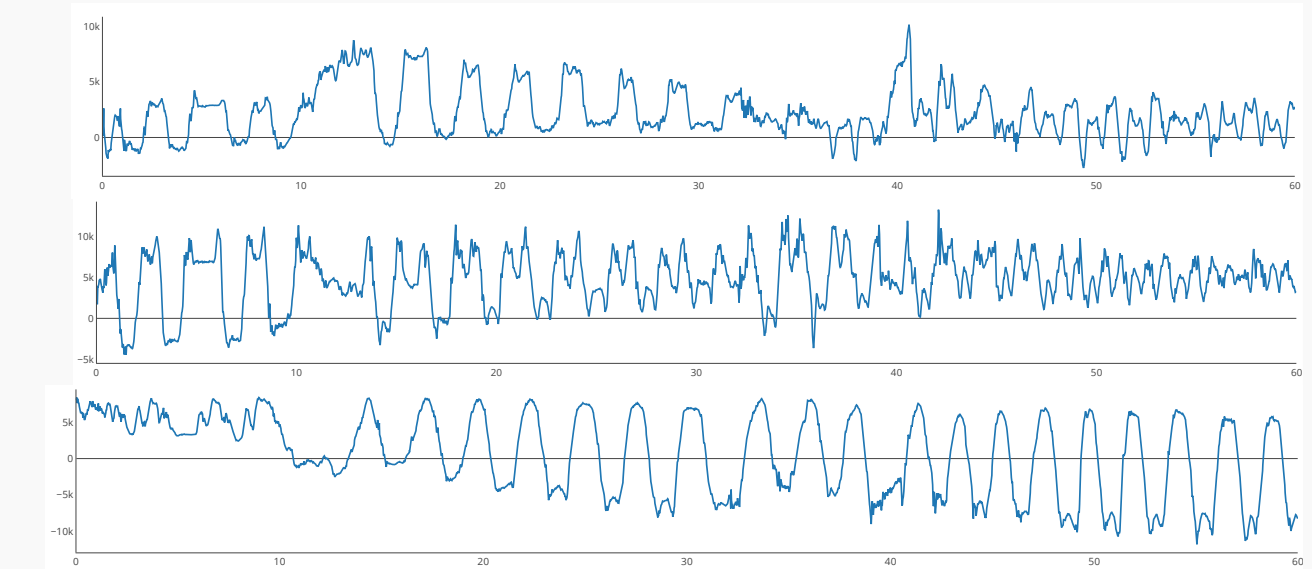
step 4 of 5

Model History

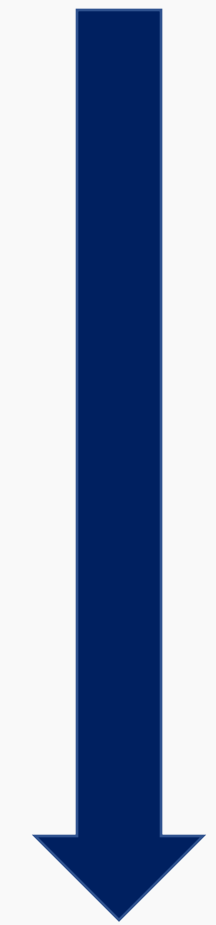
DATE	DATASETS	BUILD ID	FEATURE SELECTION	HYPERPARAMETER OPTIMIZATION	INSTANCE LENGTH (MS)	CLASSIFICATION INTERVAL (MS)	BUILD STATUS	TRAINING DETAILS	DELETE
3/15/2022	SLEEP (Group) WAKE (Group)	205978	Automatic	On	2163	200	READY		
ML MODEL	CROSS VALIDATION	TEST PERFORMANCE	LATENCY	SIZE	PERFORMANCE SUMMARY	SAVE	PUSH TO HARDWARE	LIVE CLASSIFICATION ANALYSIS	
Artificial Neural Network	0.99 +/- 0.01	0.99	< 1 ms	2.62 KB					LIVE TEST
Convolutional Neural Network	0.94 +/- 0.08	0.99	6.00 ms	14.18 KB					LIVE TEST
Decision Tree	0.988 +/- 0.010	0.97	< 1 ms	117 B					LIVE TEST
Gaussian Naive Bayes	0.96 +/- 0.04	0.99	2.00 ms	428 B					LIVE TEST
Gradient Boosting Machine	1.0 +/- 0.0	0.99	1.00 ms	3.87 KB					LIVE TEST
Logistic Regression	0.98 +/- 0.04	0.99	< 1 ms	240 B					LIVE TEST
Polynomial Support Vector Machine	0.98 +/- 0.04	0.98	1.00 ms	4.42 KB					LIVE TEST
Random Forest	1.0 +/- 0.0	0.99	1.00 ms	5.26 KB					LIVE TEST
RBF Support Vector Machine	0.96 +/- 0.07	0.98	1.00 ms	4.35 KB					LIVE TEST
Support Vector Machine	0.99 +/- 0.02	0.99	< 1 ms	304 B					LIVE TEST
XGBoost	0.999 +/- 0.005	0.99	1.00 ms	13.33 KB					LIVE TEST

Target application

- **Extremely Low-power Raise-to-Wake Activity detection**
 - Use MLC accelerometer and gyroscope to recognize specific “wrist raise to talk” action
 - Same data as initial case
 - Only wakeup MCU when “wake action” is detected
- **Power efficiency**
 - The MLC continuously monitoring motion is roughly 40x as power efficient at executing simple models compared to MCU (μA vs mA)
 - <https://qeexo.com/ai-in-the-industry/>
 - By having majority of use-time in MCU-sleep with only μA consumed by MLC for wake detection we enable significantly longer battery life in a wearable device



Data



Qeexo AutoML optimizes for unique chip capabilities

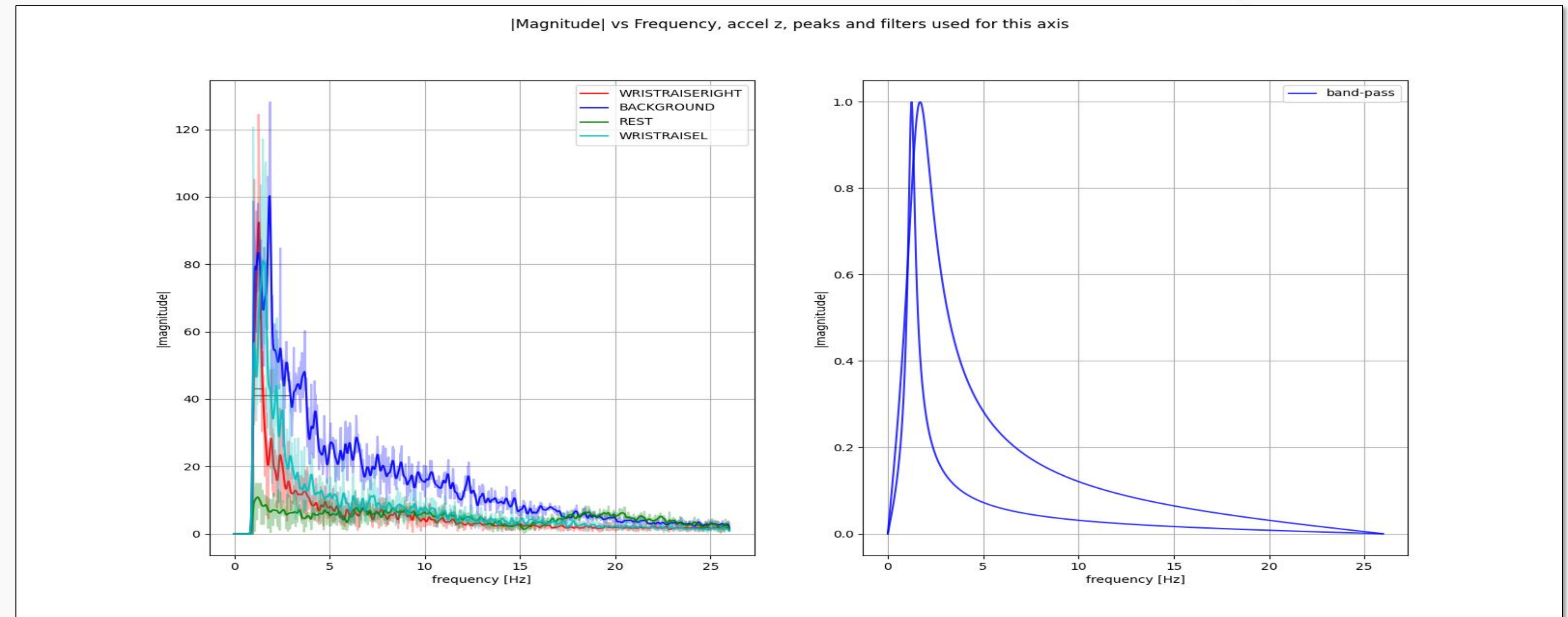
- **ST MLC** Chips support signal filters as inputs to the Decision tree: High-pass, band-pass, IIR1, IIR2
- **Qeexo AutoML** determines optimal filters in conjunction with training the decision tree: optimizing overall performance
 - Driven by training data, the filter automatically adapt to the activities being detected
- **Results** in models that are tuned to take maximum advantage of the full hardware capabilities available from **ST MLC-based** systems

Training Preference Step 2 of 5 ✕

☒ Automatic Filter and Feature Group Selection

☐ Manual Filter and Feature Selection

< NEXT CANCEL



Results

- Accuracy of 'wake detection' on extreme Low Power MLC

- Naïve Manual configuration:

ML MODEL	CROSS VALIDATION	TEST PERFORMANCE
Decision Tree	0.9 +/- 0.1	0.54

- AutoML configured signal filters + feature selection:

ML MODEL	CROSS VALIDATION	TEST PERFORMANCE
Decision Tree	0.95 +/- 0.04	0.88

- Demonstration

Live Testing

Decision Tree

CLASS LABEL	SENSITIVITY WEIGHTS	DATE
SLEEP	1	3/15/20
WAKE	1	

CLASS LABEL

SLEEP

WAKE

META-CLASSIFIER END COUNTER

0

0

FLASH

Continuous Classification

CLASS LABEL

SLEEP

WAKE


PREDICTION

1.00

0.00

Qeexo

AutoML



SLEEP



Keep in touch!

- Call to Action

Enabling the tinyML Future requires lots of innovative ideas to dissolve the barriers to entry for non-embedded-machine-learning-experts to be able to create innovative products in the tinyML space

- Questions, Ideas, Comments, Suggestions?

- Please stay in touch and let's talk



Elias Fallon
VP OF MACHINE LEARNING

Elias has a wide range of expertise in ML/AI, IoT, Electronics, and software engineering. He has 15+ years' experience leading Machine Learning and Software Engineering research and development teams developing technical software and innovative application of Machine Learning. Previously, Elias worked in the Electronic Design Automation industry at Neoliner and Cadence for more than 20 years. Elias has a M.S. and B.S. in Electrical and Computer Engineering from Carnegie Mellon University.

<https://www.linkedin.com/in/elias-fallon/>

Elias.Fallon@Qeexo.com



T I N Y



Summit 2022

Thank you



ENABLING THE tinyML FUTURE



AONdevices

arm

ASPINITY

brainchip
The Neuromorphic Computing Company

CEVA®

Deeplite

EDGE IMPULSE

emza
visual sense

FotaHub

GREENWAVES
TECHNOLOGIES

Grovetly Inc.

Himax

HOTC

imagimob

infineon

itemis

KLIKA·TECH
GLOBAL IOT SOLUTIONS

LatentAI

LATTICE
SEMICONDUCTOR

Micro.ai

OmniML

NXP

POI

Plumerai

PROPHESSEE

Qeexo

Qualcomm

Rackner

RealityAI®
Engineering Solutions for the Edge

REEXEN
technology

RENESAS

SAP

seeed
The IoT Hardware Enabler

SensiML

Sony Semiconductor
Solutions
Corporation

ST
life.augmented

SA STREAM ANALYZE

synaptics®

SynSense

SYNTIANT

Tensil.ai

TensorFlow

XMOS

Copyright Notice

This presentation in this publication was presented as a tinyML® Summit 2022. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org