CEC Continuing Education Center

DesignNews

Machine Learning in Microcontrollers

DAY 5 : Deploying Machine Learning Models

Sponsored by

Digi-Key ELECTRONICS

BENINGO EMBEDDED GROUP

informa markets

Webinar Logistics

- Turn on your system sound to hear the streaming presentation.
- If you have technical problems, click “Help” or submit a question asking for assistance.
- Participate in ‘Group Chat’ by maximizing the chat widget in your dock.

THE SPEAKER

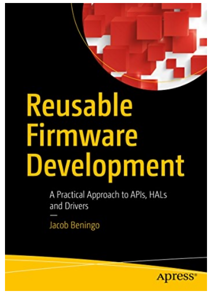
Jacob Beningo

Visit 'Lecturer Profile'

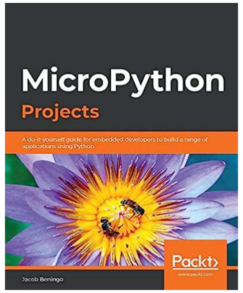
Beningo Embedded Group - President

Focus: Embedded Software Consulting and Training

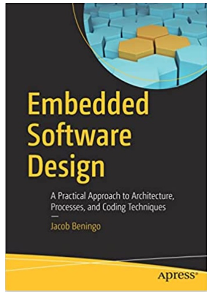
Specializes in creating and promoting embedded software **excellence** in businesses around the world.



Reusable Firmware Development
A Practical Approach to APIs, HALs and Drivers
Jacob Beningo
Apress



MicroPython Projects
A do-it-yourself guide for embedded developers to build a range of applications using Python
Packt
Jacob Beningo



Embedded Software Design
A Practical Approach to Architecture, Processes, and Coding Techniques
Jacob Beningo
Apress

Blogs for:

- DesignNews.com
- Embedded.com
- EmbeddedRelated.com
- MLRelated.com

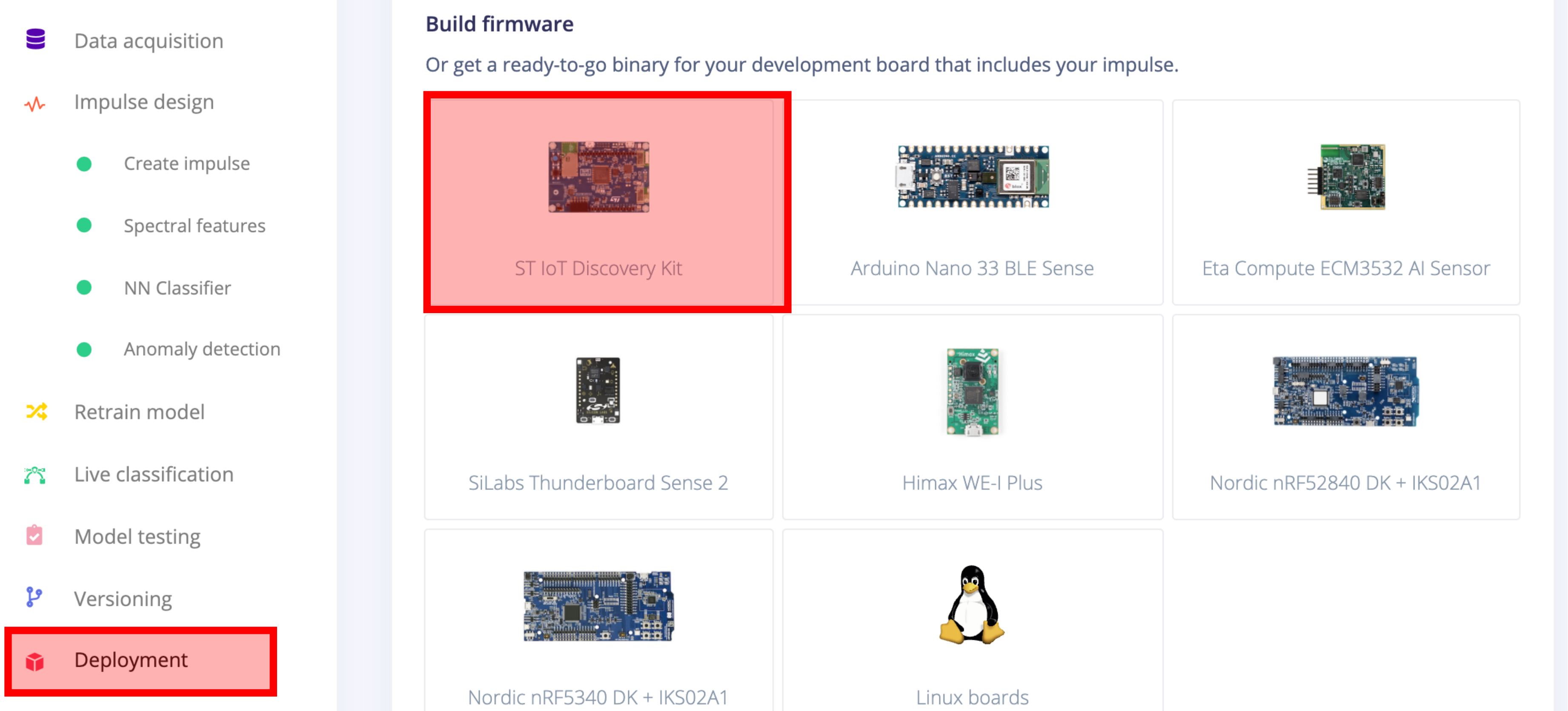
Visit www.beningo.com to learn more …

Course Sessions

- AI and ML for Microcontrollers
- Writing Embedded Software with ChatGPT and Open.AI
- Tools for Machine Learning in Microcontrollers
- Training a Model for the STM32
- **Deploying Machine Learning Models**

1 Preparing the Model for Export

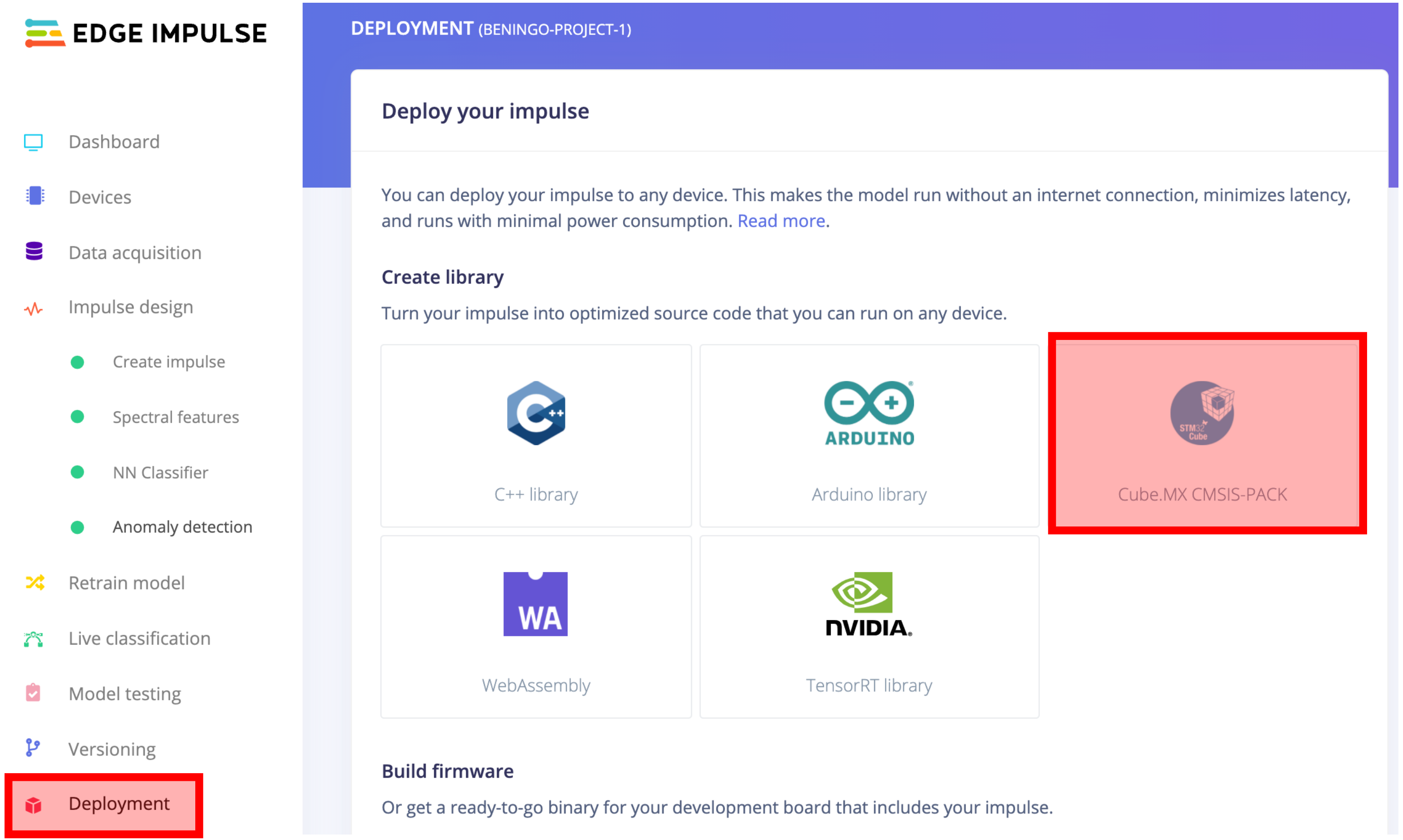
Exporting an STM32 Binary



Build firmware

Or get a ready-to-go binary for your development board that includes your impulse.

Exporting an STM32 AI pack



EDGE IMPULSE

- Dashboard
- Devices
- Data acquisition
- Impulse design
  - Create impulse
  - Spectral features
  - NN Classifier
  - Anomaly detection
- Retrain model
- Live classification
- Model testing
- Versioning
- Deployment

DEPLOYMENT (BENINGO-PROJECT-1)

Deploy your impulse

You can deploy your impulse to any device. This makes the model run without an internet connection, minimizes latency, and runs with minimal power consumption. Read more.

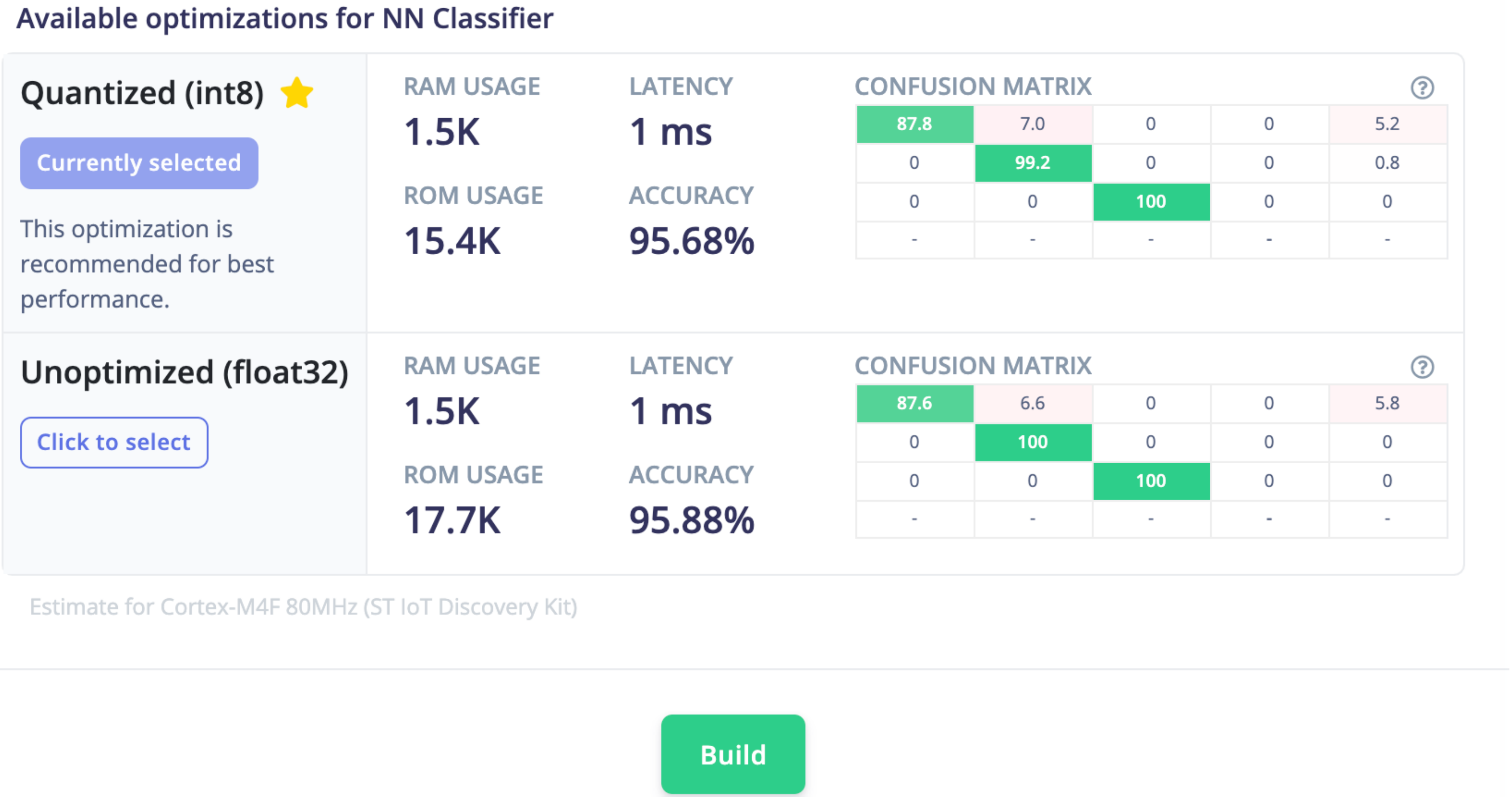
Create library

Turn your impulse into optimized source code that you can run on any device.

Build firmware

Or get a ready-to-go binary for your development board that includes your impulse.

Apply Optimization(s)



Available optimizations for NN Classifier

Quantized (int8) ★

This optimization is recommended for best performance.

RAM USAGE 1.5K

LATENCY 1 ms

ROM USAGE 15.4K

ACCURACY 95.68%

CONFUSION MATRIX

87.8	7.0	0	0	5.2
0	99.2	0	0	0.8
0	0	100	0	0
-	-	-	-	-

Unoptimized (float32)

RAM USAGE 1.5K

LATENCY 1 ms

ROM USAGE 17.7K

ACCURACY 95.88%

CONFUSION MATRIX

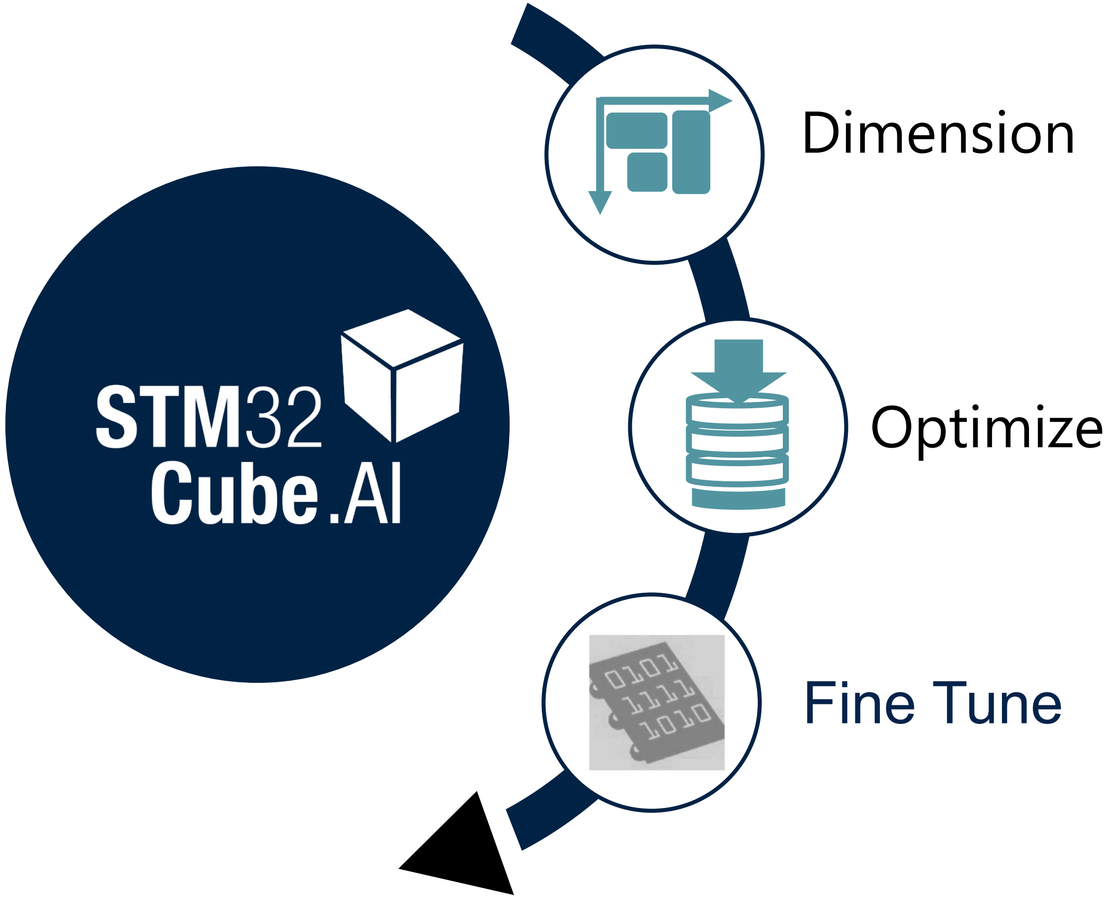
87.6	6.6	0	0	5.8
0	100	0	0	0
0	0	100	0	0
-	-	-	-	-

Estimate for Cortex-M4F 80MHz (ST IoT Discovery Kit)

What method do you prefer for testing?
- Using the prebuilt binary
- Using the pack
- C++ library
- other

2 Importing the Model

STM32Cube.AI Overview



Dimension

- ✓ Quickly assess model footprint requirements
- ✓ Select and configure MCU in STM32CubeMX
- ✓ Review model layers in STM32Cube.AI

Optimize

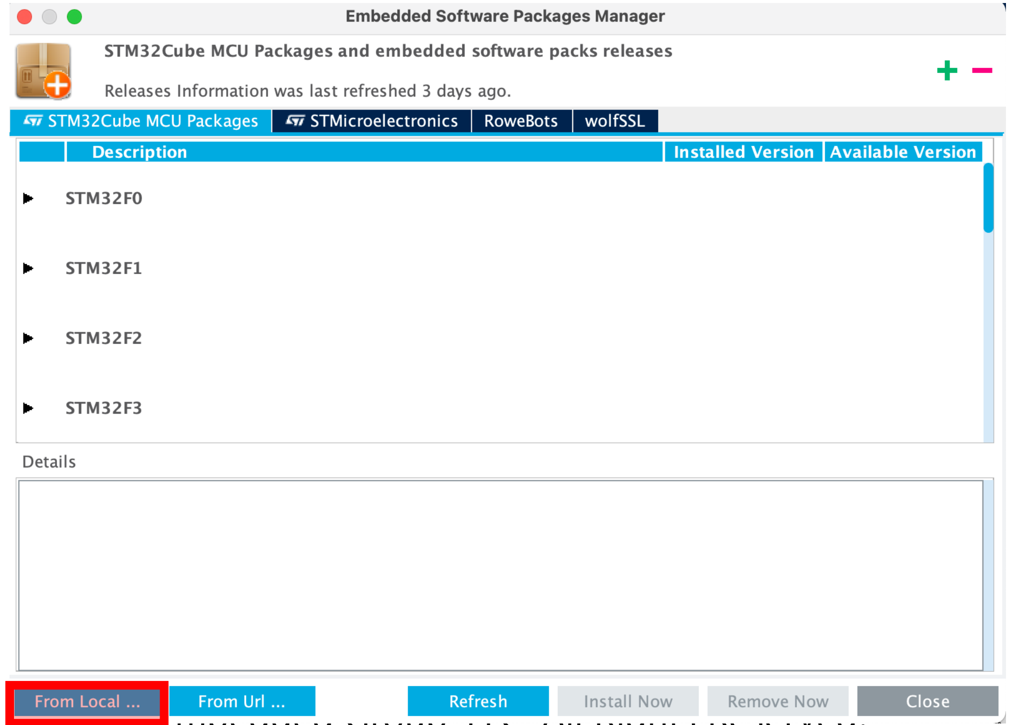
- ✓ Generate C-code for pre-trained model
- ✓ Support quantized models to reduce RAM, flash and latency with minimal loss of accuracy
- ✓ Use light run-time libraries
- ✓ Optimize for performance

Fine Tune

- ✓ Optimize memory allocation
- ✓ Fine control of weight mapping
- ✓ Split between internal and external memory
- ✓ Update model without full FW update

Install Pack into STM32Mx Project

Help -> Manage Embedded Software Packages



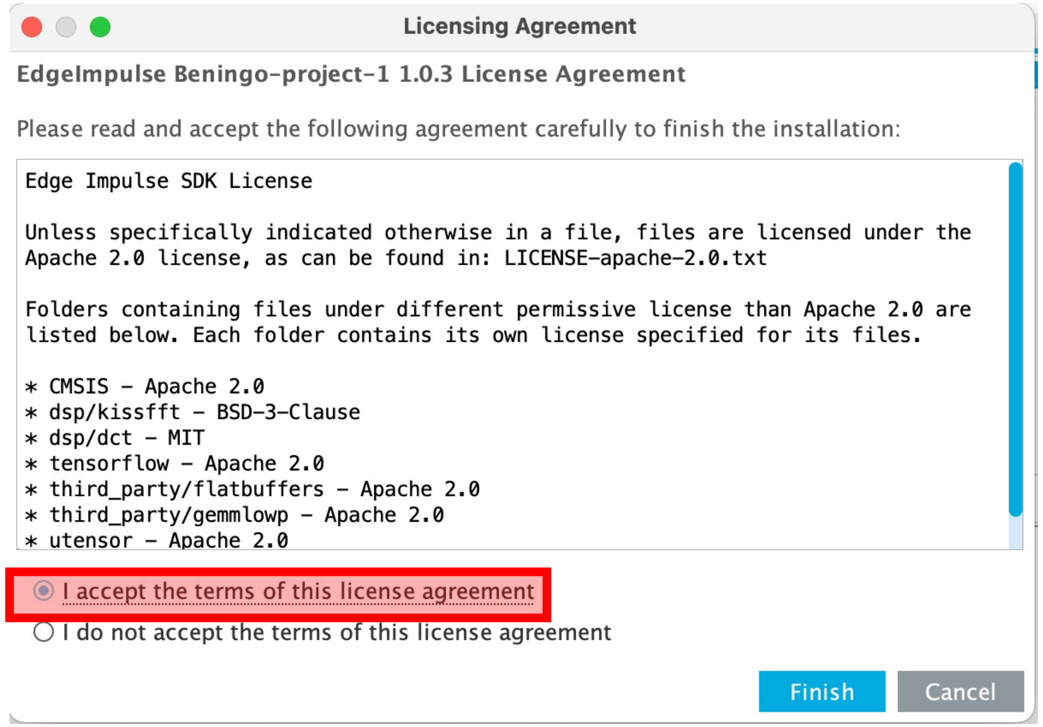
Embedded Software Packages Manager

STM32Cube MCU Packages and embedded software packs releases

Releases Information was last refreshed 3 days ago.

Description	Installed Version	Available Version
▶ STM32F0		
▶ STM32F1		
▶ STM32F2		
▶ STM32F3		

Details



Licensing Agreement

EdgeImpulse Beningo-project-1 1.0.3 License Agreement

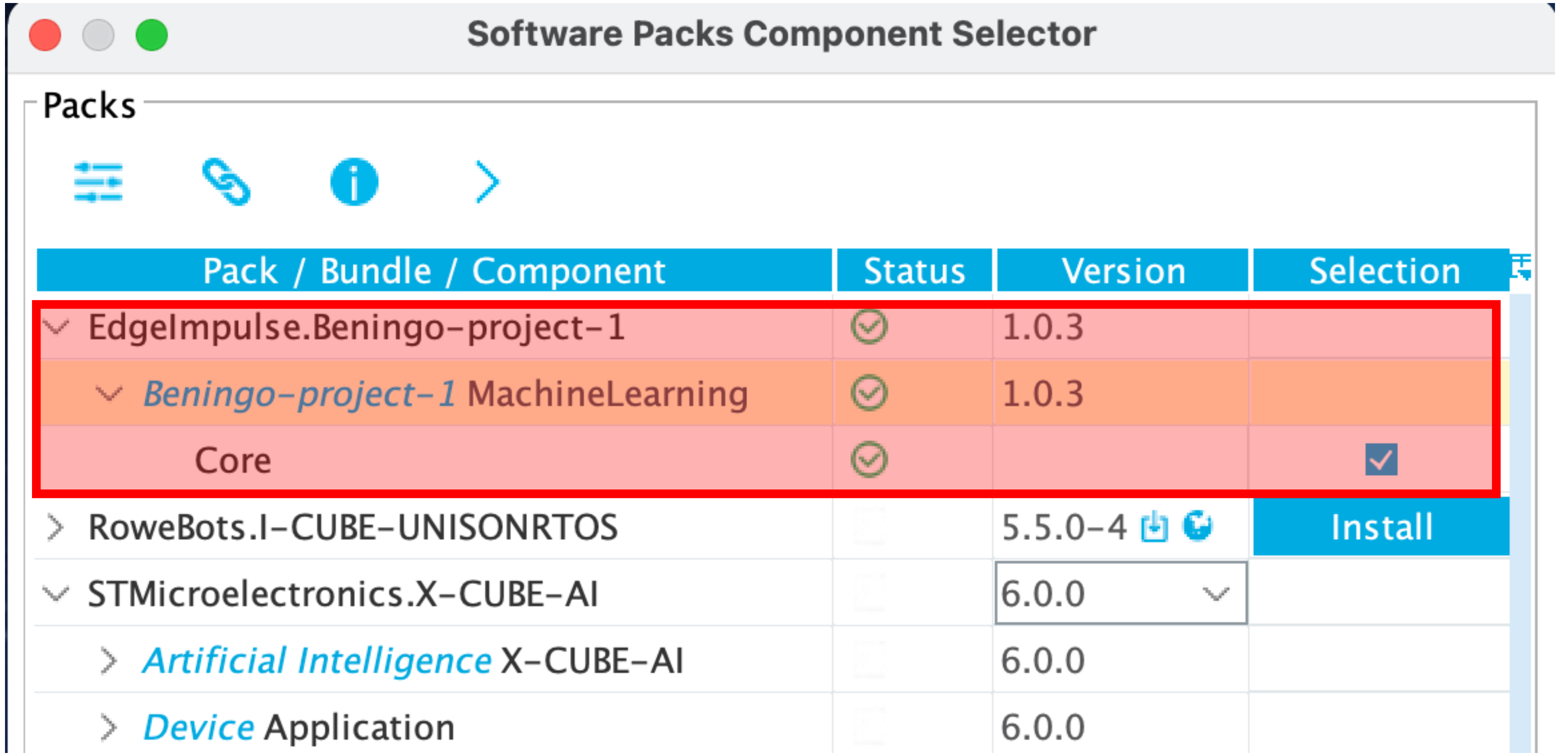
Please read and accept the following agreement carefully to finish the installation:

```
Edge Impulse SDK License

Unless specifically indicated otherwise in a file, files are licensed under the
Apache 2.0 license, as can be found in: LICENSE-apache-2.0.txt

Folders containing files under different permissive license than Apache 2.0 are
listed below. Each folder contains its own license specified for its files.

* CMSIS - Apache 2.0
* dsp/kissfft - BSD-3-Clause
* dsp/dct - MIT
* tensorflow - Apache 2.0
* third_party/flatbuffers - Apache 2.0
* third_party/gemmlowp - Apache 2.0
* utensor - Apache 2.0
```

● I accept the terms of this license agreement

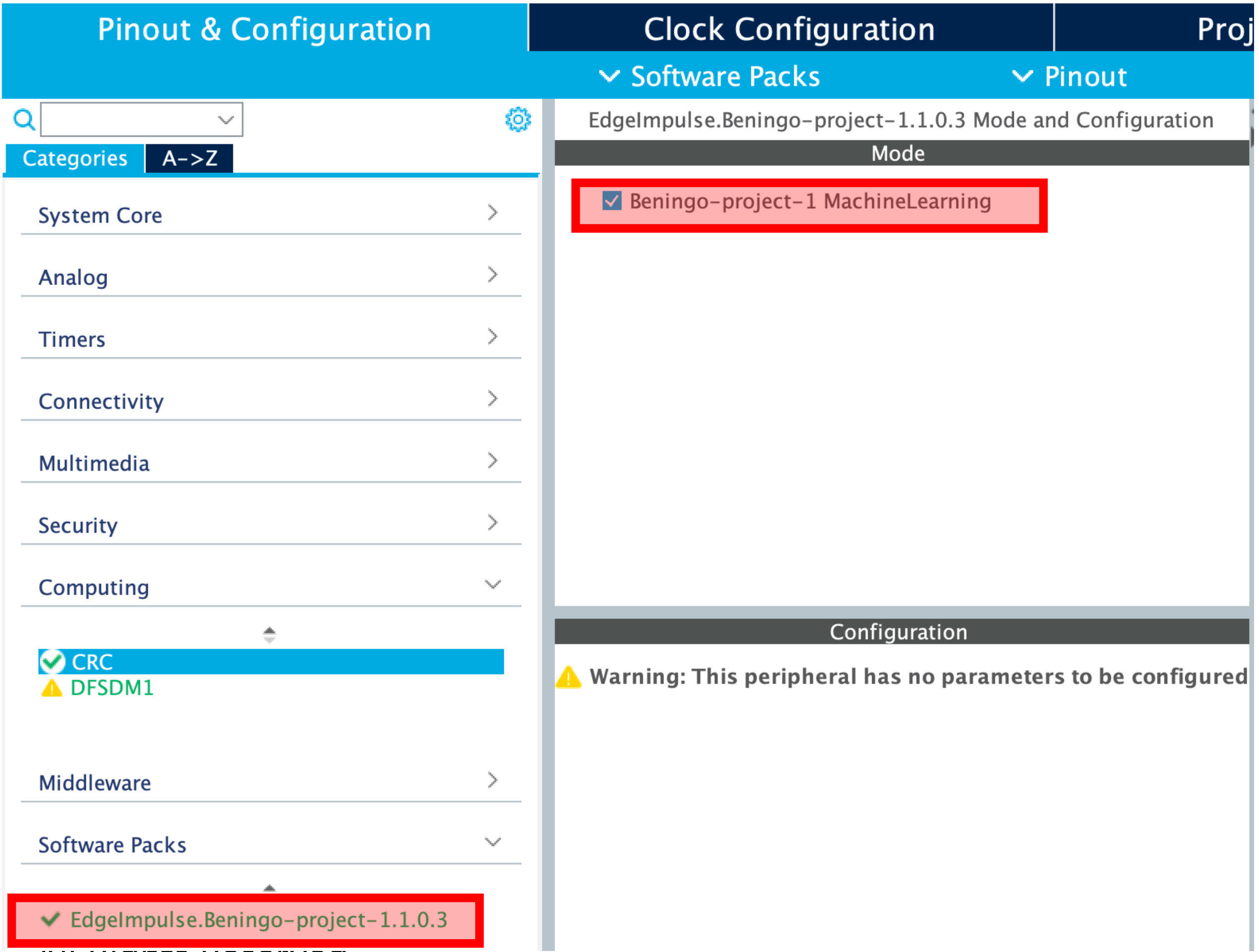
○ I do not accept the terms of this license agreement

Install Pack into STM32Mx Project



Pack / Bundle / Component	Status	Version	Selection
EdgeImpulse.Beningo-project-1	✓	1.0.3	
Beningo-project-1 MachineLearning	✓	1.0.3	
Core	✓		☑
RoweBots.I-CUBE-UNISONRTOS		5.5.0-4	Install
STMicroelectronics.X-CUBE-AI		6.0.0	
Artificial Intelligence X-CUBE-AI		6.0.0	
Device Application		6.0.0	

Install Pack into STM32Mx Project

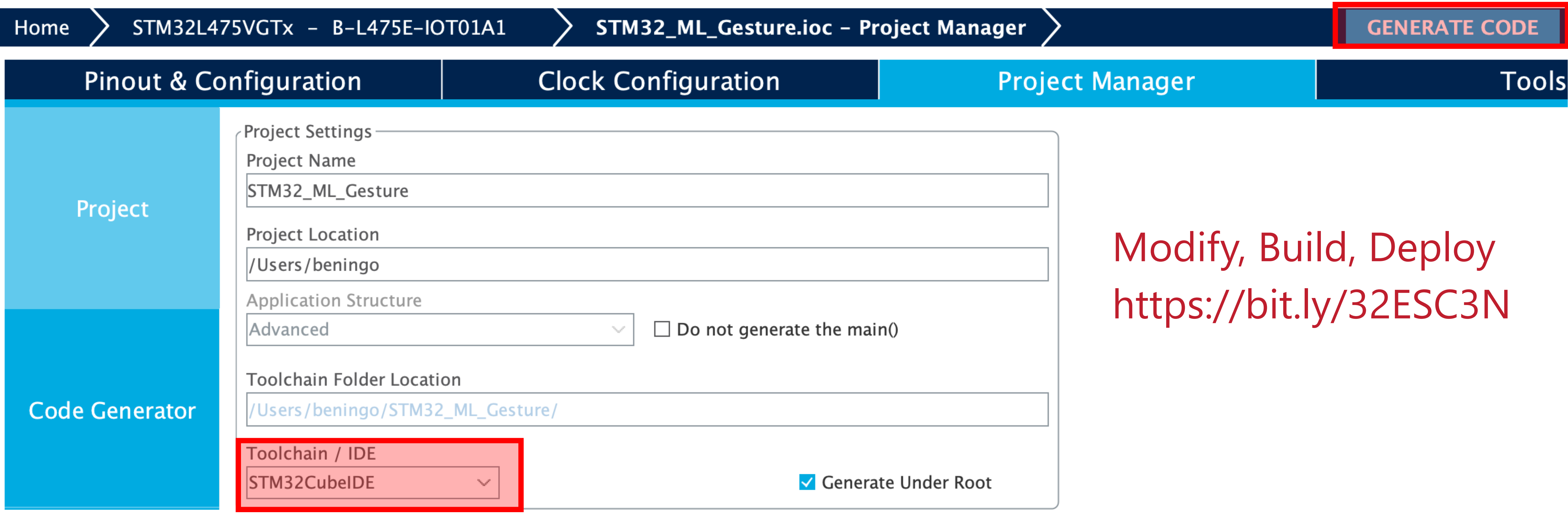


EdgeImpulse.Beningo-project-1.1.0.3 Mode and Configuration

☑ Beningo-project-1 MachineLearning

Warning: This peripheral has no parameters to be configured

Install Pack into STM32Mx Project



Project Settings

Project Name: STM32_ML_Gesture

Project Location: /Users/beningo

Application Structure: Advanced

□ Do not generate the main()

Toolchain Folder Location: /Users/beningo/STM32_ML_Gesture/

Toolchain / IDE: STM32CubeIDE

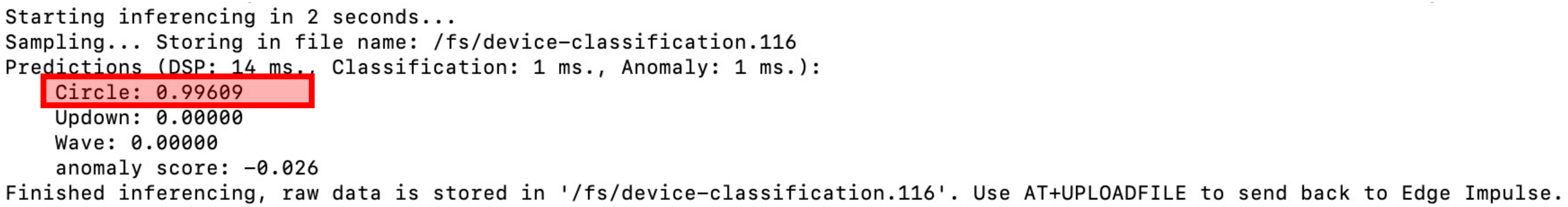
☑ Generate Under Root

Modify, Build, Deploy
https://bit.ly/32ESC3N

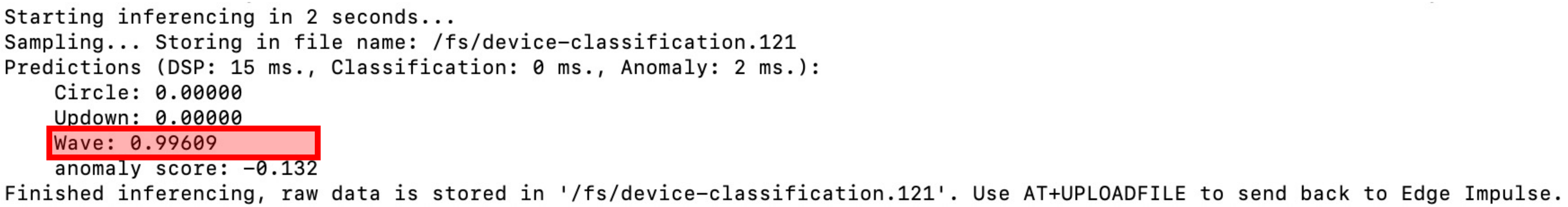
3 Running the Model

Running the Model

In a terminal, run the command: edge-impulse-run-impulse

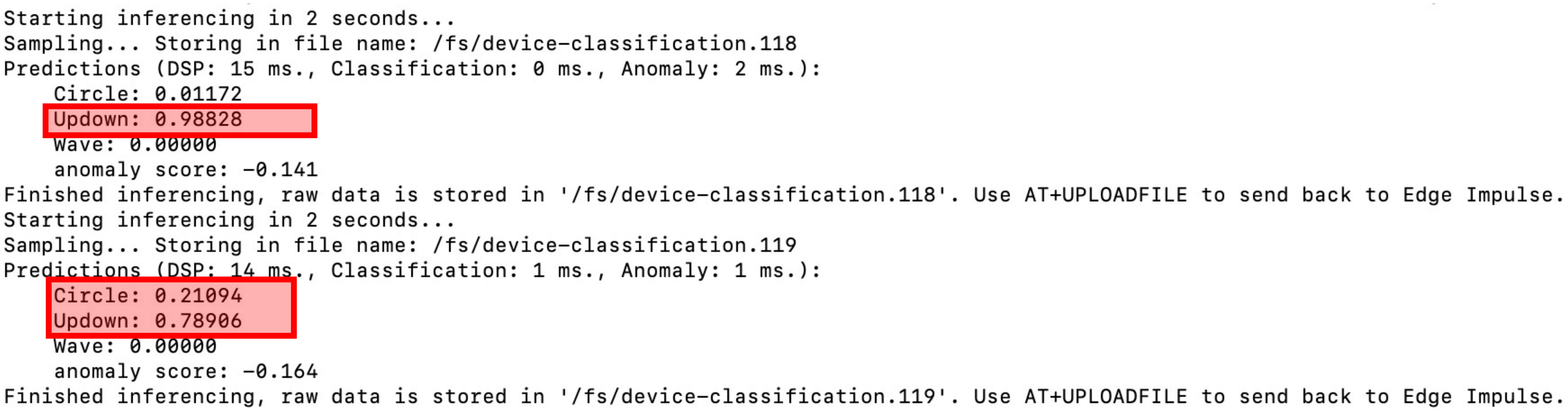


```
Starting inferencing in 2 seconds...
Sampling... Storing in file name: /fs/device-classification.116
Predictions (DSP: 14 ms., Classification: 1 ms., Anomaly: 1 ms.):
    Circle: 0.99609
    Updown: 0.00000
    Wave: 0.00000
    anomaly score: -0.026
Finished inferencing, raw data is stored in '/fs/device-classification.116'. Use AT+UPLOADFILE to send back to Edge Impulse.
```



```
Starting inferencing in 2 seconds...
Sampling... Storing in file name: /fs/device-classification.121
Predictions (DSP: 15 ms., Classification: 0 ms., Anomaly: 2 ms.):
    Circle: 0.00000
    Updown: 0.00000
    Wave: 0.99609
    anomaly score: -0.132
Finished inferencing, raw data is stored in '/fs/device-classification.121'. Use AT+UPLOADFILE to send back to Edge Impulse.
```

Running the Model



```
Starting inferencing in 2 seconds...
Sampling... Storing in file name: /fs/device-classification.118
Predictions (DSP: 15 ms., Classification: 0 ms., Anomaly: 2 ms.):
    Circle: 0.01172
    Updown: 0.98828
    Wave: 0.00000
    anomaly score: -0.141
Finished inferencing, raw data is stored in '/fs/device-classification.118'. Use AT+UPLOADFILE to send back to Edge Impulse.
Starting inferencing in 2 seconds...
Sampling... Storing in file name: /fs/device-classification.119
Predictions (DSP: 14 ms., Classification: 1 ms., Anomaly: 1 ms.):
    Circle: 0.21094
    Updown: 0.78906
    Wave: 0.00000
    anomaly score: -0.164
Finished inferencing, raw data is stored in '/fs/device-classification.119'. Use AT+UPLOADFILE to send back to Edge Impulse.
```

What methods can be used to improve classifaction ?

- Running average on the output
- Monitor the anomaly value
- Set a minimum classification percentage
- All the above
- Other

4

Going Further

Next Steps

- Connect the output to a PWM LED channel
- Setup a DAC and drive an output voltage
- Configure the rate at which the inference runs (frequency control)
- Try and compare the Keras model behavior
- Improve the training model to provide a more accurate sine wave

Going Further

FP-AI-SENSING1

People activity recognition

Audio scene classification

FP-AI-NANOEDG1

Condition-based monitoring

FP-AI-VISION1

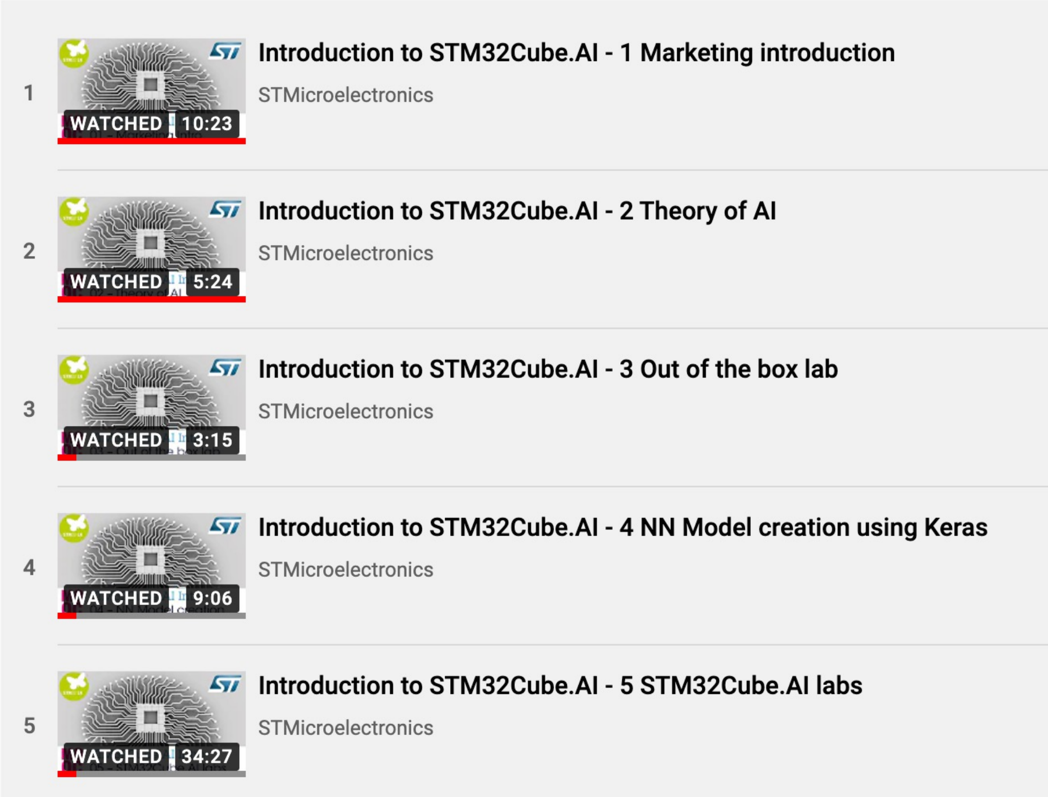
Person presence detection

Food classification

FP-AI-FACEREC1

Face recognition

https://bit.ly/3nf99EZ


1 WATCHED 10:23 Introduction to STM32Cube.AI - 1 Marketing introduction
STMicroelectronics

2 WATCHED 5:24 Introduction to STM32Cube.AI - 2 Theory of AI
STMicroelectronics

3 WATCHED 3:15 Introduction to STM32Cube.AI - 3 Out of the box lab
STMicroelectronics

4 WATCHED 9:06 Introduction to STM32Cube.AI - 4 NN Model creation using Keras
STMicroelectronics

5 WATCHED 34:27 Introduction to STM32Cube.AI - 5 STM32Cube.AI labs
STMicroelectronics

AI and ML Resources

- Jacob's AI Blogs
- Jacob's CEC courses
- Jacob's ML Blogs

- Embedded Bytes Newsletter
  - http://bit.ly/1BAHYXm

BENINGO EMBEDDED GROUP

www.beningo.com

CEC Continuing Education Center

DesignNews

Thank You

Sponsored by

Digi-Key ELECTRONICS

BENINGO EMBEDDED GROUP

informa markets