ISM6HG256X: machine learning core

Introduction

This document provides information on the machine learning core feature available in the ISM6HG256X. The machine learning processing capability allows moving some algorithms from the application processor to the MEMS sensor, enabling consistent reduction of power consumption.

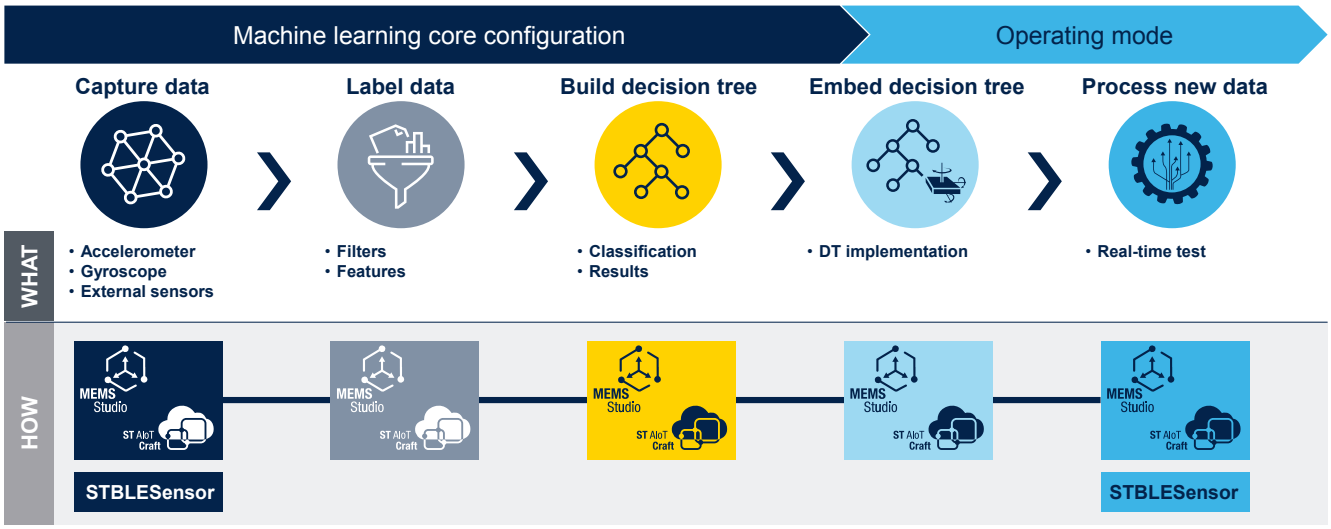
The machine learning processing capability is obtained through decision tree logic. A decision tree is a mathematical tool composed of a series of configurable nodes. Each node is characterized by an “if-then-else” condition, where an input signal (represented by statistical parameters calculated from the sensor data) is evaluated against a threshold.

The ISM6HG256X can be configured to run up to eight decision trees simultaneously and independently. The decision trees are stored in the device and generate results in the dedicated output registers.

The results of the decision tree can be read from the application processor at any time. Furthermore, there is the possibility to generate an interrupt for every change in the result in the decision tree.

The figure below illustrates the workflow for implementing the five key steps in developing an MLC solution, along with the various software tools available. STMicroelectronics offers a comprehensive set of MLC configurations that cover a range of ready-to-use use cases, as well as tutorials demonstrating how to create MLC solutions using different ST hardware kits and software tools. Examples and tutorials for the machine learning core feature are available in the public STMicroelectronics GitHub repository (see st-mems-machine-learning-core).

Figure 1. Machine learning core supervised approach



1 Machine learning core in the ISM6HG256X

The machine learning core (together with the finite state machine) is one of the main embedded features available in the ISM6HG256X. It is composed of a set of configurable parameters and decision trees able to implement algorithms in the sensor itself.

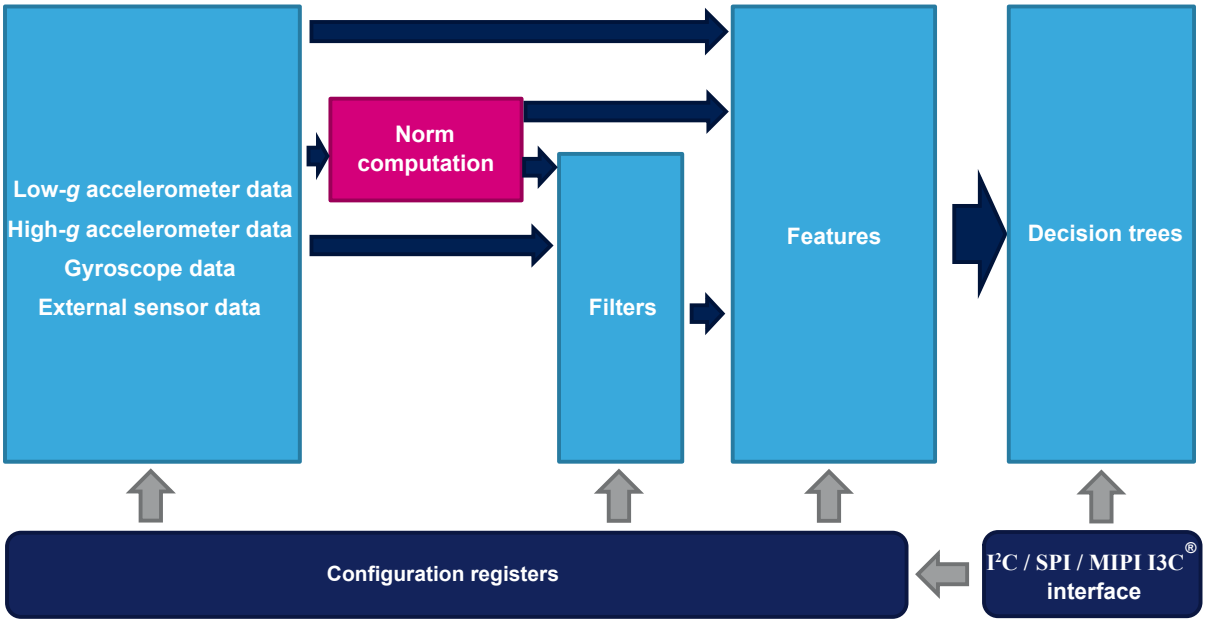
The classification of algorithms suitable for the machine learning core are those which can be implemented by following an inductive approach, which involves searching patterns from observations. Some examples of algorithms that follow this approach are vibration intensity detection, asset tracking, context awareness, equipment status classification, and similar industrial applications.

The idea behind the machine learning core is to use the low-*g* accelerometer, high-*g* accelerometer, gyroscope, and external sensor data (readable through the I²C controller interface) to compute a set of statistical parameters selectable by the user (such as mean, variance, energy, peak, zero-crossing, and so on) in a defined time window. In addition to the sensor input data, some filtered inputs can be defined by applying some configurable filters available in the device.

The machine learning core parameters are called “features” and can be used as input for a configurable decision tree that can be stored in the device.

The decision tree that can be stored in the ISM6HG256X is a binary tree composed of a series of nodes. In each node, a statistical parameter (feature) is evaluated against a threshold to establish the evolution in the next node. When a leaf (one of the last nodes of the tree) is reached, the decision tree generates a result that is readable through a dedicated device register.

Figure 2. Machine learning core in the ISM6HG256X



The machine learning core output data rate can be configured among one of the seven available rates from 15 Hz to 960 Hz. The MLC_ODR bits in the embedded functions register MLC_ODR (60h) allow selecting one of the rates as shown in the following table.

Table 1. Machine learning core output data rates

MLC_ODR bits in MLC_ODR (60h)	Machine learning core output data rate
000	15 Hz
001	30 Hz (default)
010	60 Hz
011	120 Hz
100	240 Hz
101	480 Hz
110	960 Hz

In order to implement the machine learning processing capability of the ISM6HG256X, it is necessary to use a “supervised learning” approach that consists of:

- Identifying some classes to be recognized
- Collecting multiple data logs for each class
- Performing some data analysis from the collected logs to learn a generic rule, which allows mapping inputs (data logs) to outputs (classes to be recognized)

In an activity recognition algorithm, for instance, the classes to be recognized might be: stationary, walking, jogging, biking, driving, and so forth. Multiple data logs have to be acquired for every class, for example multiple people performing the same activity.

The analysis on the collected data logs has the purpose of:

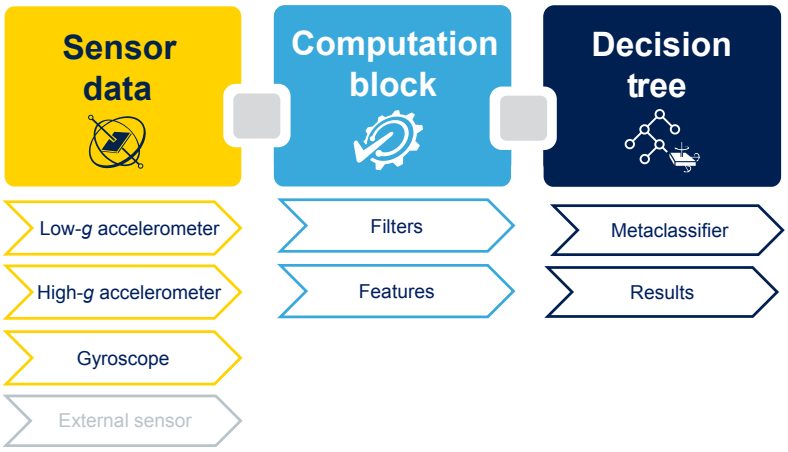
- Defining the features to be used to correctly classify the different classes
- Defining the filters to be applied to the input data to improve the performance using the selected features
- Generating a dedicated decision tree able to recognize one of the different classes (mapping inputs to outputs)

Once a decision tree has been defined, a configuration for the device can be generated by the software tools provided by STMicroelectronics. The decision tree runs on the device, minimizing power consumption.

Going deeper in detail in the machine learning core feature inside the ISM6HG256X, it can be thought of as three main blocks (Figure 3):

- Sensor data
- Computation block
- Decision tree

Figure 3. Machine learning core blocks



The first block called “Sensor data”, is composed of data coming from the low-*g* accelerometer, high-*g* accelerometer, and gyroscope, which are built in the device, or from an additional external sensor, which might be connected to the ISM6HG256X through the I²C controller interface (sensor hub).

The machine learning core inputs defined in the first block are used in the second block, the “Computation block”, where filters and features can be applied. The features are statistical parameters computed from the input data (or from the filtered data) in a defined time window, selectable by the user.

The features computed in the computation block are used as input for the third block of the machine learning core. This block called “Decision tree”, includes the binary tree that evaluates the statistical parameters computed from the input data. In the binary tree, these parameters are compared against certain thresholds to generate results (in the example of the activity recognition described above, the results were: stationary, walking, jogging, biking, and so on). The decision tree results might also be filtered by an optional filter called "metaclassifier". The machine learning core results are the decision tree results after the optional metaclassifier processing.

The machine learning core memory is organized in a “dynamic” or “modular” way, in order to maximize the number of computation blocks, which can be configured in the device (filters, features, and so on). ST's software tools automatically handle memory management and generate MLC configurations for the ISM6HG256X sensor.

The following sections explain in detail the three main blocks of the machine learning core in the ISM6HG256X described in Figure 3.

2 Inputs

The ISM6HG256X works as a combo (low-*g* accelerometer + high-*g* accelerometer + gyroscope) sensor, generating acceleration and angular rate output data. The 3-axis data of the acceleration and angular rate can be used as input for the machine learning core. Figure 4, Figure 5, and Figure 6 show the inputs of the machine learning core block in the low-*g* accelerometer, high-*g* accelerometer, and gyroscope digital chains. The position of the machine learning core (MLC) block in the digital chains is the same for all the connection modes described in the device datasheet.

Note: Data as input to the MLC block has a higher resolution of data available in the output registers.

Note: High-g accelerometer data as input to the MLC block is mutually exclusive with external sensor data as input to the MLC block.

Figure 4. MLC inputs (low-*g* accelerometer)

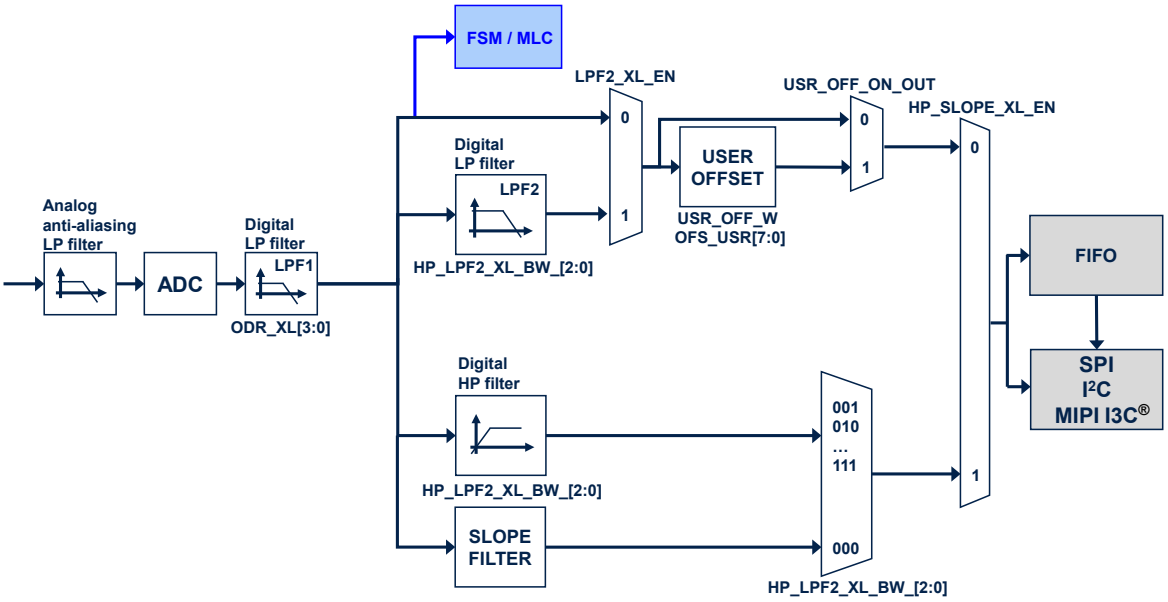


Figure 5. MLC inputs (high-*g* accelerometer)

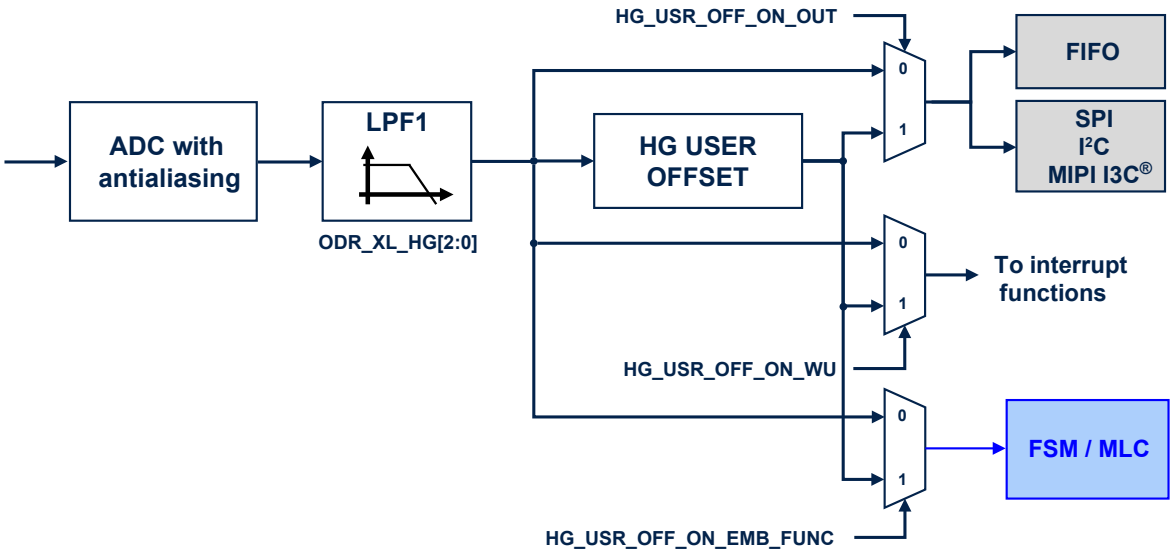
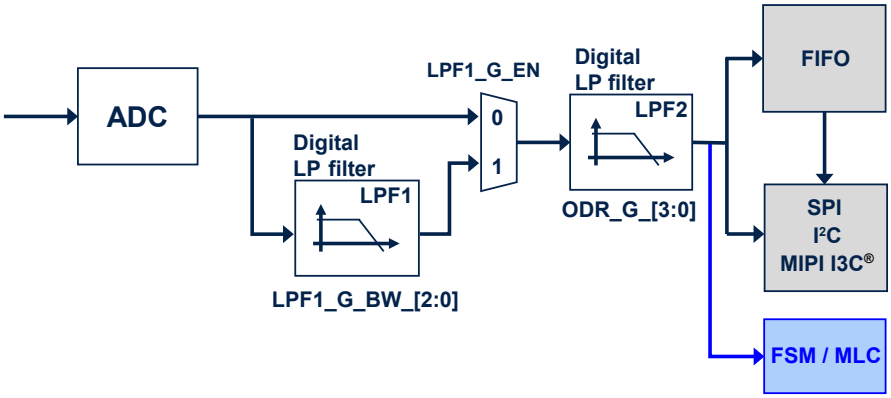


Figure 6. **MLC inputs (gyroscope)**



The rate of the input data must be equal to or higher than the machine learning core data rate configurable through the embedded functions register MLC_ODR (60h), as described in Table 1.

Example: In an activity recognition algorithm running at 30 Hz, the machine learning core ODR must be selected at 30 Hz, while the sensor ODRs must be equal to or higher than 30 Hz.

The machine learning core uses the following unit conventions:

- Low-*g* accelerometer data in [*g*]
- High-*g* accelerometer data in [h*g*]
- Gyroscope data in [rad/sec]
- External sensor data in [gauss] for a magnetometer

Since it is possible to connect an external sensor (for example, a magnetometer or a pressure sensor) to the ISM6HG256X through the sensor hub interface, the data coming from an external sensor can also be used as input for machine learning processing. In this case, the first six sensor hub bytes (two bytes per axis) are considered as input for the MLC. In other words, the MLC directly takes as input the content of the ISM6HG256X registers from SENSOR_HUB_1 (02h) to SENSOR_HUB_6 (07h).

Depending on the EXT_FORMAT_SEL bit in the EXT_FORMAT (00h) register of the embedded advanced features page 2, the MLC is able to read one of three values of 16-bit each (for example, axes of a magnetometer) or one value of 24-bit (for example, pressure sensor value). In the former case, the user may select the external sensor sensitivity by writing a half-precision floating-point (HFP) encoded value to registers MLC_EXT_SENSITIVITY_L (E8h) and MLC_EXT_SENSITIVITY_H (E9h) in the embedded advanced features page 1, that is used to convert the LSB value to the correct unit that the MLC uses for its computations.

Example: For a magnetometer like the LIS2MDL, the sensitivity is 1.5 mG/LSB, which becomes 0.0015 G/LSB after converting it to gauss, and becomes 1624h converted as HFP (half-precision floating-point value for the ISM6HG256X sensitivity registers).

Sensitivity [mG/LSB]	Sensitivity [G/LSB]	Sensitivity HFP
1.5 mG/LSB	0.0015 G/LSB	1624h

Note: The half-precision floating-point format is expressed as:

SEEEEEFFFFFFFFFF (S: 1 sign bit; E: 5 exponent bits; F: 10 fraction bits).

The following procedure allows changing the conversion factor for the external magnetometer data:

1. Write 80h to register 01h // Enable access to the embedded functions registers
2. Write 40h to register 17h // PAGE_RW (17h) = 40h: enable write transaction
3. Write 11h to register 02h // PAGE_SEL (02h) = 11h
4. Write E8h to register 08h // PAGE_ADDRESS (08h) = E8h
5. Write [LSB] conversion factor (LIS2MDL example, 24h) to register 09h
6. Write [MSB] conversion factor (LIS2MDL example, 16h) to register 09h
7. Write 00h to register 17h // PAGE_RW (17h) = 00h: disable read / write transaction
8. Write 00h to register 01h // Disable access to the embedded functions registers

The MLC in the ISM6HG256X also has the possibility to use external sensors with a single axis value of 24-bit (for example, a pressure sensor). In this case, some dedicated registers are available in the embedded advanced features page 2, in order to properly set sensitivity and offset:

- EXT_3BYTE_SENSITIVITY_L (02h)
- EXT_3BYTE_SENSITIVITY_H (03h)
- EXT_3BYTE_OFFSET_XL (06h)
- EXT_3BYTE_OFFSET_L (07h)
- EXT_3BYTE_OFFSET_H (08h)

The EXT_3BYTE_OFFSET_XL/L/H registers must be written with the three bytes composing an LSB value that is subtracted from the raw LSB value in the registers SENSOR_HUB_1 to SENSOR_HUB_3. The EXT_3BYTE_SENSITIVITY_L/H must instead be written with an HFP value that is multiplied by the result of the previous operation to obtain the converted value in the desired unit of measurement.

Note: The procedures above to change the sensitivity for the external sensor are included in the configuration generated by ST's machine learning core tools, so the user just needs to set a sensitivity value in the tool, which translates it to the corresponding register setting.

To summarize the machine learning core inputs:

- Low-*g* and high-*g* accelerometer data conversion factor is automatically handled by the device.
- Gyroscope data conversion factor is automatically handled by the device.
- External sensor data conversion factor is not automatically handled by the device. A conversion factor must be set by the user in order to make the machine learning core work with the correct unit of measurement.

An additional input available for all sensor data (low-*g* accelerometer, high-*g* accelerometer, gyroscope, and external sensor) is the norm. From the 3-axis data, the machine learning core (in the ISM6HG256X) internally computes the norm and the squared norm. These two additional signals can be used as inputs for machine learning processing.

The norm and the squared norm of the input data are computed with the following formulas:

$$V = \sqrt{x^2 + y^2 + z^2}$$

$$V^2 = x^2 + y^2 + z^2$$

Norm and squared norm data can be used in the decision trees in order to guarantee a high level of program customization for the user.

Note: The data rate for MLC inputs is set through the MLC_ODR bits. If the sensor ODR is higher than MLC_ODR, MLC automatically decimates the input data (without any additional filtering).

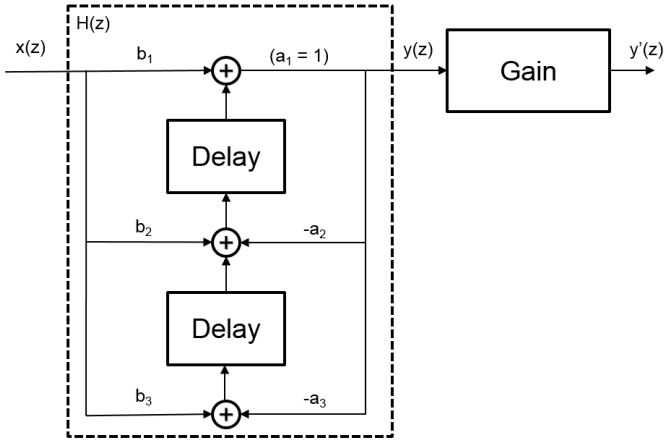
It is recommended to select MLC_ODR equal to the sensor ODR to avoid decimation of the MLC inputs. Selecting MLC_ODR lower than the sensor ODR is also possible, but the different frequency response could lower the accuracy of the MLC solution.

3 Filters

The input data seen in the previous section can be filtered by different types of filters available in the machine learning core logic. The basic element of the machine learning core filtering is a second order IIR filter, as shown in Figure 7.

Note: The filters available in the MLC block are independent of any other filter available on the device (the filters described in this section are illustrated in the MLC block of Figure 4, Figure 5, and Figure 6).

Figure 7. Filter basic element



The transfer function of the generic IIR 2nd order filter is the following:

$$H\left(z\right) = \frac{b_1 + b_2 z^{-1} + b_3 z^{-2}}{1 + a_2 z^{-1} + a_3 z^{-2}}$$

From Figure 7, the outputs can be defined as:

$$y(z) = H(z) \cdot x(z)$$

$$y'(z) = y(z) \cdot Gain$$

To optimize memory usage, the machine learning core has default coefficients for the different types of filters (high-pass, bandpass, IIR1, IIR2). The machine learning core software tools help in configuring the filter by asking for the filter coefficients needed after selecting the type of filter. The following table shows the default values and the configurable values for the coefficients, depending on the filter type chosen. By setting different coefficients, it is possible to tune the filter for the specific application.

Table 2. Filter coefficients

Filter type / Coefficients	b_1	b_2	b_3	a_2	a_3	Gain
High-pass filter	0.5	-0.5	0	0	0	1
Bandpass filter	1	0	-1	Configurable	Configurable	Configurable
IIR1 filter	Configurable	Configurable	0	Configurable	0	1
IIR2 filter	Configurable	Configurable	Configurable	Configurable	Configurable	1

The filter coefficient values are expressed as half-precision floating-point format: SEEEEEFFFFFFFFFF (S: 1 sign bit; E: 5 exponent bits; F: 10 fraction bits).

3.1 Filter coefficients

The IIR filter coefficients can be computed with different tools, including MATLAB®, Octave, and Python. In MATLAB®, for instance, the following function can be used to generate coefficients for a low-pass filter:

```
[b, a] = butter(N, f_cut / (ODR / 2), 'low')
```

Where:

- N is the order of the IIR filter (1 for IIR1, 2 for IIR2)
- f_cut is the cutoff frequency [Hz] of the filter
- ODR is the machine learning core data rate [Hz]
- 'low' (or 'high') is the type of filter to be implemented (low-pass or high-pass)

Note: It is possible to configure a high-pass filter with the cutoff at half of the bandwidth (ODR/4) without inserting the coefficients. The machine learning core has some predefined coefficients for this configuration.

The following function instead allows generating bandpass filter coefficients through MATLAB®:

```
[b, a] = butter(1, [f1 f2] / (ODR / 2), 'bandpass')
```

Note: Since only a_2, a_3 and gain are configurable for a bandpass filter, the b vector should be normalized by setting gain = b(1). Bandpass filters are generated as first-order filters in MATLAB® and Python.

Example:

b = [0.2929 0 -0.2929]; a = [1.0 -0.5858 0.4142];

can be written as b = [1 0 -1] and gain = 0.2929.

So, the bandpass filter coefficients are:

a_2 = -0.5858; a_3 = 0.4142; gain = 0.2929.

Table 3 shows some examples of filter coefficients (most of them considering an ODR of 30 Hz).

When designing high-pass and bandpass IIR filters, the stability of the filters in half-precision floating-point must be considered, since the resolution loss can cause some divergence in the signal if the filters are not very stable.

It is recommended to use first-order IIR filters when the cutoff frequency (normalized) is below 0.02 [2*f_cutoff/ODR] or increasing the cutoff frequency.

Table 3. Examples of filter coefficients

Filter type / Coefficients	b_1	b_2	b_3	a_2	a_3	Gain
High-pass IIR1, f_cut = 1 Hz, ODR = 30 Hz	0.905	-0.905	-	-0.8096	-	1
High-pass IIR1, f_cut = 2 Hz, ODR = 30 Hz	0.8247	-0.80247	-	-0.6494	-	1
High-pass IIR1, f_cut = 5 Hz, ODR = 30 Hz	0.634	-0.634	-	-0.268	-	1
High-pass IIR1, f_cut = 10 Hz, ODR = 30 Hz	0.366	-0.366	-	0.268	-	1
High-pass IIR2, f_cut = 1 Hz, ODR = 30 Hz	0.8623	-1.725	0.8623	-1.705	0.7437	1
High-pass IIR2, f_cut = 2 Hz, ODR = 30 Hz	0.743	-1.486	0.743	-1.419	0.553	1
High-pass IIR2, f_cut = 5 Hz, ODR = 30 Hz	0.465	-0.93	0.465	-0.62	0.2404	1
High-pass IIR2, f_cut = 10 Hz, ODR = 30 Hz	0.155	-0.31	0.155	0.62	0.2404	1
Low-pass IIR1, f_cut = 1 Hz,	0.0951	0.0951	-	-0.8096	-	1

Filter type / Coefficients	b_1	b_2	b_3	a_2	a_3	Gain
ODR = 30 Hz						
Low-pass IIR1, f_cut = 2 Hz, ODR = 30 Hz	0.1753	0.1753	-	-0.6494	-	1
Low-pass IIR1, f_cut = 5 Hz, ODR = 30 Hz	0.366	0.366	-	-0.268	-	1
Low-pass IIR1, f_cut = 10 Hz, ODR = 30 Hz	0.634	0.634	-	0.268	-	1
Low-pass IIR2, f_cut = 1 Hz, ODR = 30 Hz	0.00953	0.01906	0.00953	-1.705	0.7437	1
Low-pass IIR2, f_cut = 2 Hz, ODR = 30 Hz	0.03357	0.06714	0.03357	-1.419	0.553	1
Low-pass IIR2, f_cut = 5 Hz, ODR = 30 Hz	0.155	0.31	0.155	-0.62	0.2404	1
Low-pass IIR2, f_cut = 10 Hz, ODR = 30 Hz	0.465	0.93	0.465	0.62	0.2404	1
Bandpass IIR2, f1 = 1.5 Hz, f2 = 5 Hz, ODR = 30 Hz	1	0	-1	-1.203	0.4453	0.2773
Bandpass IIR2, f1 = 0.2 Hz, f2 = 1 Hz, ODR = 120 Hz	1	0	-1	-1.958	0.959	0.02052

4 Features

The features are the statistical parameters computed from the machine learning core inputs. The machine learning core inputs that can be used for the computation of the features are:

- Sensor input data, which includes
  - Sensor data from the X, Y, Z axes (for example, AccLG_X, AccLG_Y, AccLG_Z, AccHG_X, AccHG_Y, AccHG_Z, Gyro_X, Gyro_Y, Gyro_Z)
  - External sensor data (for example, ExtSens_X, ExtSens_Y, ExtSens_Z)
  - Norm and squared norm signals of sensor / external sensor data (AccLG_V, AccLG_V2, AccHG_V, AccHG_V2, Gyro_V, Gyro_V2, Ext_V, Ext_V2)
- Filtered data (for example, high-pass on AccLG, bandpass on AccLG_V2, and so forth)

All the standard features are computed within a defined time window, which is also called “window length” since it is expressed as the number of samples. The size of the window has to be determined by the user and is very important for the machine learning processing, since all the statistical parameters in the decision tree are evaluated in this time window. It is not a moving window, features are computed just once for every WL sample (where WL is the size of the window).

The window length can have values from 1 to 255 samples. The choice of the window length value depends on the MLC data rate (MLC_ODR bits in the embedded functions register MLC_ODR (60h)), which introduces a latency for the generation of the machine learning core result, and in the specific application or algorithm. In an activity recognition algorithm for instance, it can be decided to compute the features every 2 or 3 seconds, which means that considering sensors running at 30 Hz, the window length should be around 60 or 90 samples respectively.

Besides the standard features, the machine learning core provides the recursive features, which are computed at every sample and are evaluated by the decision trees once for every WL sample analogously to the standard features. For more details, refer to Section 4.13: Recursive features.

Some of the features in the machine learning core require some additional parameters for the evaluation (for example, an additional threshold). The following table shows all the features available in the machine learning core including additional parameters.

Note: The maximum number of features that can be configured in the MLC is 31. Feature values are limited to the range ±65504.

Table 4. **Features**

Feature	Additional parameter
MEAN	-
VARIANCE	-
ENERGY	-
PEAK-TO-PEAK	-
ZERO-CROSSING	Threshold
POSITIVE ZERO-CROSSING	Threshold
NEGATIVE ZERO-CROSSING	Threshold
PEAK DETECTOR	Threshold
POSITIVE PEAK DETECTOR	Threshold
NEGATIVE PEAK DETECTOR	Threshold
MINIMUM	-
MAXIMUM	-
RECURSIVE MEAN	b_1, b_2, a_2
RECURSIVE RMS	b_1, b_2, a_2
RECURSIVE VARIANCE	b_1, b_2, a_2
RECURSIVE MAXIMUM	ths, c_{start}
RECURSIVE MINIMUM	ths, c_{start}
RECURSIVE PEAK-TO-PEAK	ths, c_{start}

4.1 Mean

The feature “*Mean*” computes the average of the selected input (*I*) in the defined time window (*WL*) with the following formula:

$$Mean = \frac{1}{WL}\sum_{k=0}^{WL-1} I_k$$

4.2 Variance

The feature “*Variance*” computes the variance of the selected input (*I*) in the defined time window (*WL*) with the following formula:

$$Variance = \left(\frac{\sum_{k=0}^{WL-1} {I_k}^2}{WL}\right) - \left(\frac{\sum_{k=0}^{WL-1} I_k}{WL}\right)^2$$

4.3 Energy

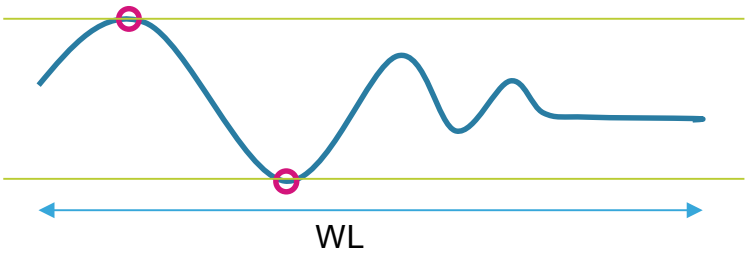
The feature “*Energy*” computes the energy of the selected input (*I*) in the defined time window (*WL*) with the following formula:

$$Energy = \sum_{k=0}^{WL-1} {I_k}^2$$

4.4 Peak-to-peak

The feature “*Peak-to-peak*” computes the maximum peak-to-peak value of the selected input in the defined time window.

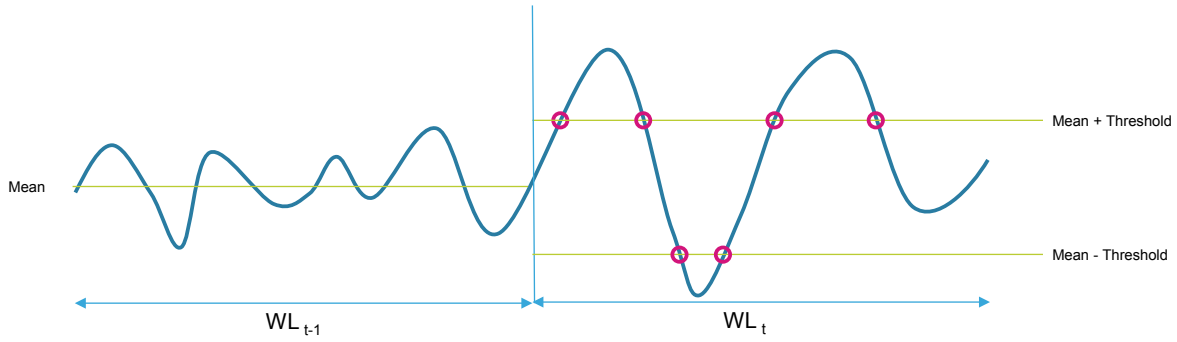
Figure 8. Peak-to-peak



4.5 Zero-crossing

The feature “*Zero-crossing*” computes the number of times the selected input crosses one of two levels; the values of these levels are computed as the average of the input signal computed in the previous window (feature “Mean”) plus or minus a configurable threshold. If the threshold is set to zero, then the two values coincide and each crossing is counted twice, as both levels are always crossed at the same time.

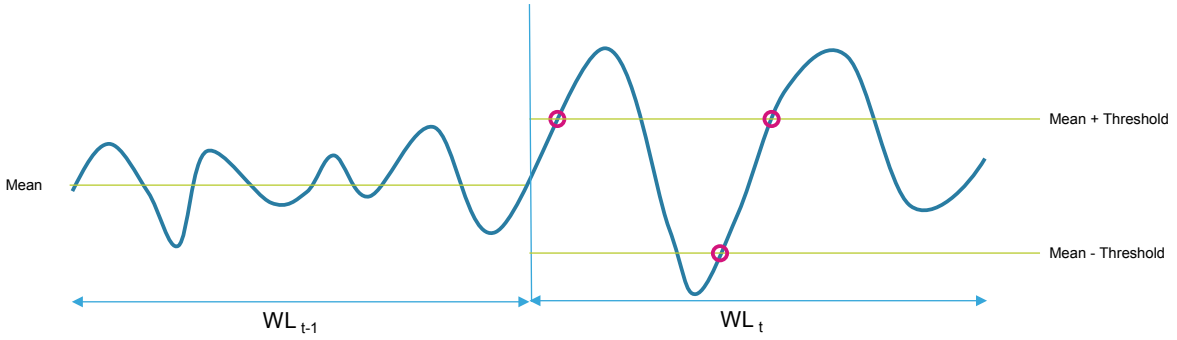
Figure 9. Zero-crossing



4.6 Positive zero-crossing

The feature "*Positive zero-crossing*" computes the number of times the selected input crosses one of two levels; the values of these levels are computed as the average of the input signal computed in the previous window (feature "Mean") plus or minus a configurable threshold. If the threshold is set to zero, then the two values coincide and each crossing is counted twice, as both levels are always crossed at the same time. Only crossings with a positive slope are considered for this feature.

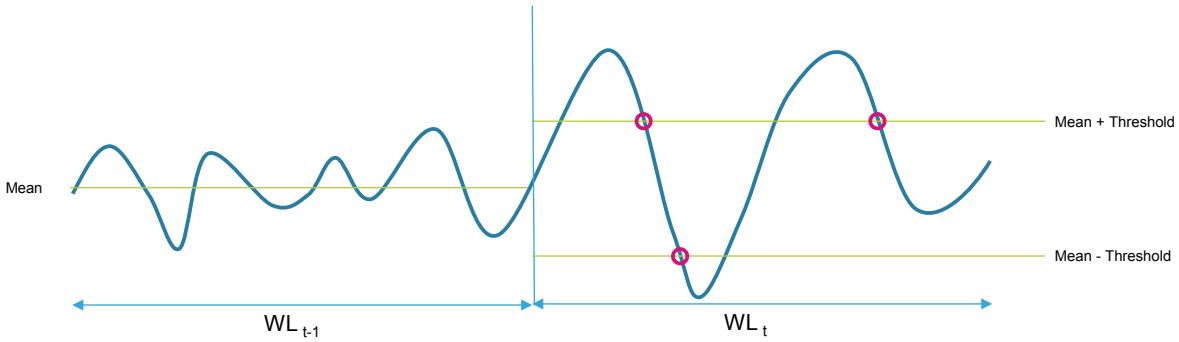
Figure 10. Positive zero-crossing



4.7 Negative zero-crossing

The feature "*Negative zero-crossing*" computes the number of times the selected input crosses one of two levels; the values of these levels are computed as the average of the input signal computed in the previous window (feature "Mean") plus or minus a configurable threshold. If the threshold is set to zero, then the two values coincide and each crossing is counted twice, as both levels are always crossed at the same time. Only crossings with a negative slope are considered for this feature.

Figure 11. Negative zero-crossing



4.8 Peak detector

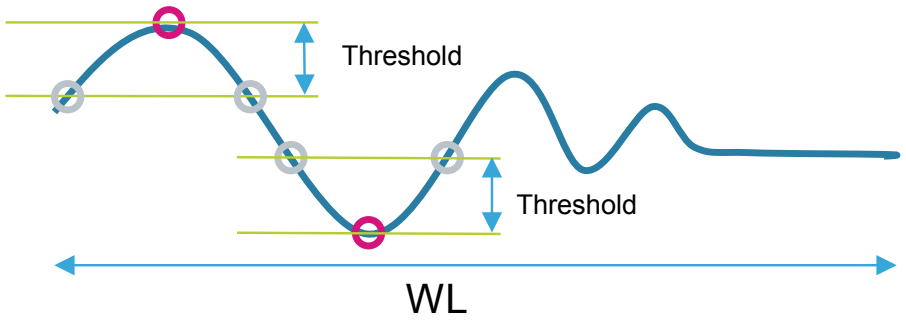
The feature "*Peak detector*" counts the number of peaks (positive and negative) of the selected input in the defined time window.

A threshold has to be defined by the user for this feature, and a buffer of three values is considered for the evaluation. If the second value of the three values buffer is higher (or lower) than the other two values of a selected threshold, the number of peaks is increased.

The buffer of three values considered for the computation of this feature is a moving buffer inside the time window.

The following figure shows an example of the computation of this feature, where two peaks (one positive and negative) have been detected in the time window.

Figure 12. Peak detector



4.9 Positive peak detector

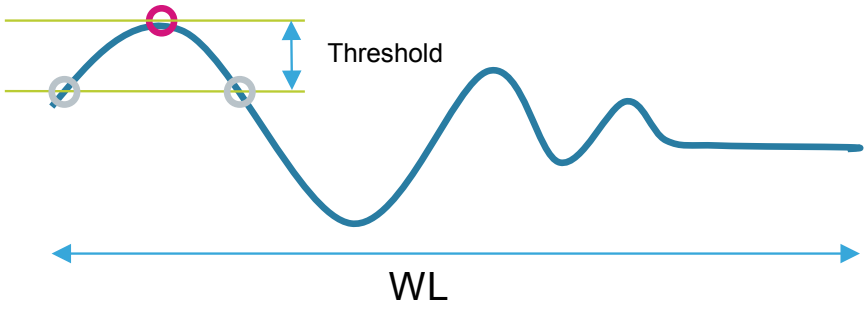
The feature "*Positive peak detector*" counts the number of positive peaks of the selected input in the defined time window.

A threshold has to be defined by the user for this feature, and a buffer of three values is considered for the evaluation. If the second value of the three values buffer is higher than the other two values of a selected threshold, the number of peaks is increased.

The buffer of three values considered for the computation of this feature is a moving buffer inside the time window.

The following figure shows an example of the computation of this feature, where just one peak (positive) has been detected in the time window.

Figure 13. Positive peak detector



4.10 Negative peak detector

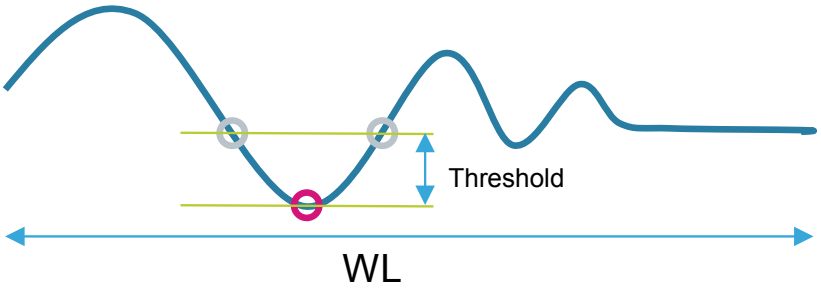
The feature "*Negative peak detector*" counts the number of negative peaks of the selected input in the defined time window.

A threshold has to be defined by the user for this feature, and a buffer of three values is considered for the evaluation. If the second value of the three values buffer is lower than the other two values of a selected threshold, the number of peaks is increased.

The buffer of three values considered for the computation of this feature is a moving buffer inside the time window.

The following figure shows an example of the computation of this feature, where just one peak (negative) has been detected in the time window.

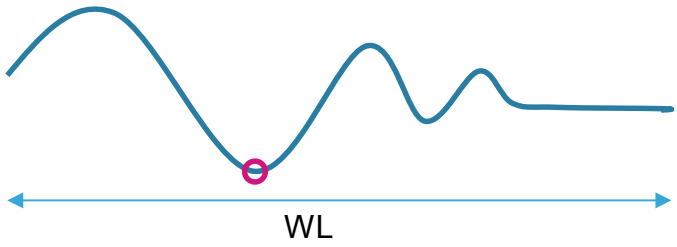
Figure 14. Negative peak detector



4.11 Minimum

The feature “*Minimum*” computes the minimum value of the selected input in the defined time window.
The following figure shows an example of minimum in the time window.

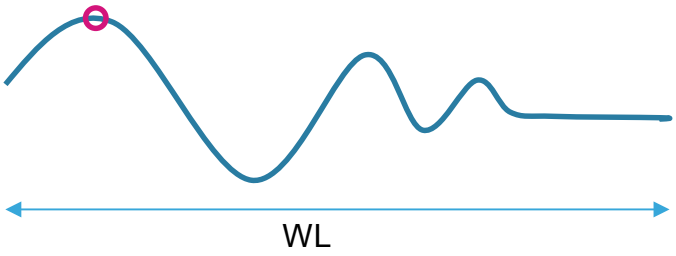
Figure 15. Minimum



4.12 Maximum

The feature “*Maximum*” computes the maximum value of the selected input in the defined time window.
The following figure shows an example of maximum in the time window.

Figure 16. Maximum



4.13 Recursive features

Besides the standard features previously described, the ISM6HG256X has a set of advanced features called "recursive".

Recursive algorithms are widely used in real-time embedded systems for the low computational complexity and minimum storage needs.

In the MLC, recursive features overcome some of the limitations associated with the features computed every window length (nonoverlapped window), allowing better detection of short events and simplifying the coexistence of different algorithms.

Compared to windowed features, the recursive features:

- Provide a valid value every MLC_ODR
- Allow tuning the "memory time" of the feature using weighting

Recursive features in the MLC are divided in two groups:

- First-order mean, RMS, variance
- Maximum, minimum, peak-to-peak

4.13.1 Recursive mean, RMS, variance

Recursive features on the mean, RMS, and variance are based on the IIR low-pass filter. The parameters of these features are b_1, b_2, and a_2. Example: $b_1 = 1$; $b_2 = 0.25$; $a_2 = 0.75$

The following table shows the three formulas of the first group of recursive features.

Table 5. First group of recursive features

Recursive feature	Formula
Recursive mean	$IIR_{LP}[x_i]$
Recursive variance	$IIR_{LP}[x_i^2]$-IIR_{LP}^2 $[x_i]$
Recursive RMS	$\sqrt{IIR_{LP}[x_i^2]}$

4.13.2 Recursive maximum, minimum, peak-to-peak

Recursive features on maximum, minimum, peak-to-peak are based on a concept similar to the envelope detector:

- Maximum/minimum value equal to input value when higher/lower than current values
- Otherwise, the maximum/minimum value starts to decay with a configurable coefficient

The following table shows the three formulas of the second group of recursive features.

The configurable parameters for these features are:

- Threshold, which sets a "rest value" (it should be set as the average value of the input signal)
- c_{start}, which is the initial value of the decay coefficient and must be set between 0 and 1 (lower values mean faster updates)

Table 6. Second group of recursive features

Recursive feature	Formula	
Recursive maximum (Max_i)(1)	$Max_i = in_i$	if $in_i \geq Max_{i-1}$ $c_{max} = c_{start}$
	$Max_i = THS + (Max_{i-1}-THS) \cdot c_{max}$	if $in_i < Max_{i-1}$ $c_{max} = c_{start}\ c_{max}$
Recursive minimum (Min_i)(2)	$Min_i = in_i$	if $in_i \leq Min_{i-1}$ $c_{min} = c_{start}$
	$Min_i = THS - (THS - Min_{i-1}) \cdot c_{min}$	if $in_i > Min_{i-1}$ $c_{min} = c_{start}\ c_{min}$
Recursive peak-to-peak	$pk2pk_i = Max_i - Min_i$	

1. *Recursive maximum (Max_i) cannot be lower than the threshold (THS).*
2. *Recursive minimum (Min_i) cannot be higher than the threshold (THS).*

4.14 Selection of features

The selection of the features to be used for the machine learning core configuration depends on the specific application.

Considering that the use of too many features may lead to overfitting and too large decision trees, it is recommended to start first by selecting the four most common features:

- Mean
- Variance
- Energy
- Peak-to-peak

If the performance is not good with these features, and in order to improve the accuracy, other features can be considered to better separate the classes.

Input data for the features calculation (from the low-*g* accelerometer, high-*g* accelerometer, gyroscope) and axes (for example, X, Y, Z, V) have to be chosen according to the specific application as well. Some classes are strongly correlated with sensor orientation (that is, applications which use the device carry position), so it is better to use individual axis (X, Y, Z). Other classes (like walking) are independent of orientation, so it is better to use the norm (V or V2).

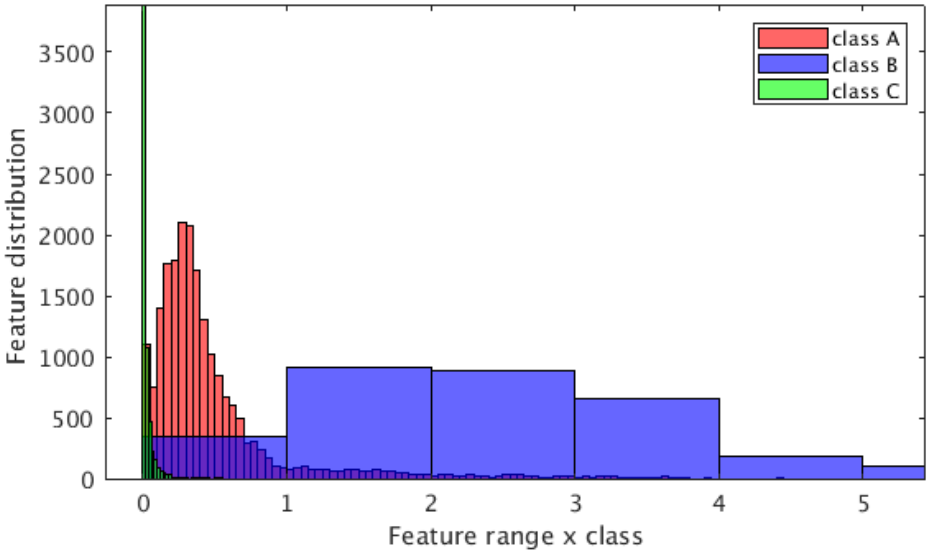
Sometimes the basic features (mean, variance, energy, and so forth) might not help in distinguishing the dominating frequency, so embedded digital filters can be enabled to select a specific region of frequency. Using the filtered signal, certain classes may be distinguished more precisely. For instance, if the user is walking, the typical signal is around 1-2 Hz, while if the user is jogging, the typical signal is around 2.5-4 Hz.

The information contribution from a single feature can be evaluated by a measure of how much different classes are separated (from each other). This analysis can be done in a graphical way, by plotting 1D/2D graphs as described in the following examples.

4.14.1 Histogram of a single feature (1D plot)

The following figure shows a histogram of the computed values of a single feature for three different classes. These three classes are reasonably separated, so an important level of information is expected with this feature. For reference, the computed classification accuracy with this single feature is around 75%.

Figure 17. Distribution of single feature for three different classes



4.14.2 Visualization of two features (2D plot)

The following figure shows a 2D plot related to a 2-class classification problem with the selection of two features:

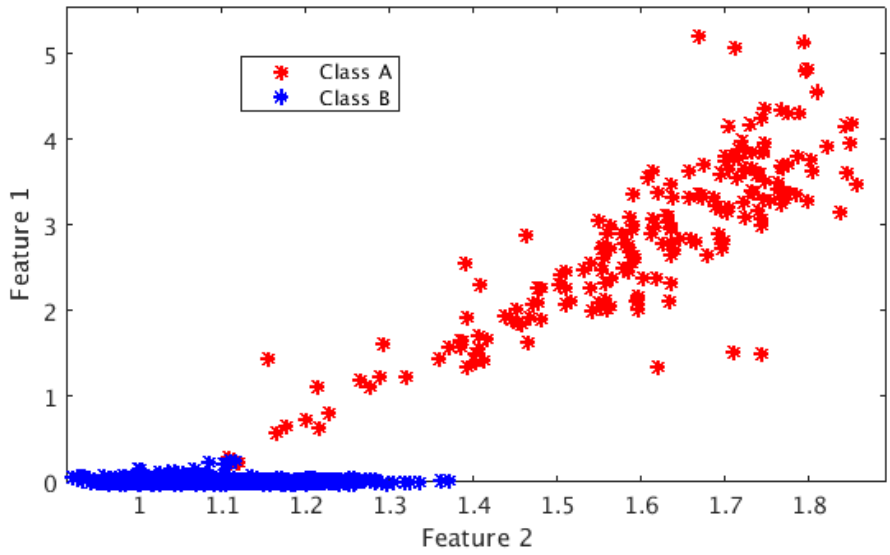
- Feature 1 on the graph vertical axis
- Feature 2 on the graph horizontal axis

In this case, the strict separation between the two classes is evident:

- Class A in red
- Class B in blue

A good information contribution can be obtained by combining the two features. For reference, the classification accuracy obtained with this example is more than 95%.

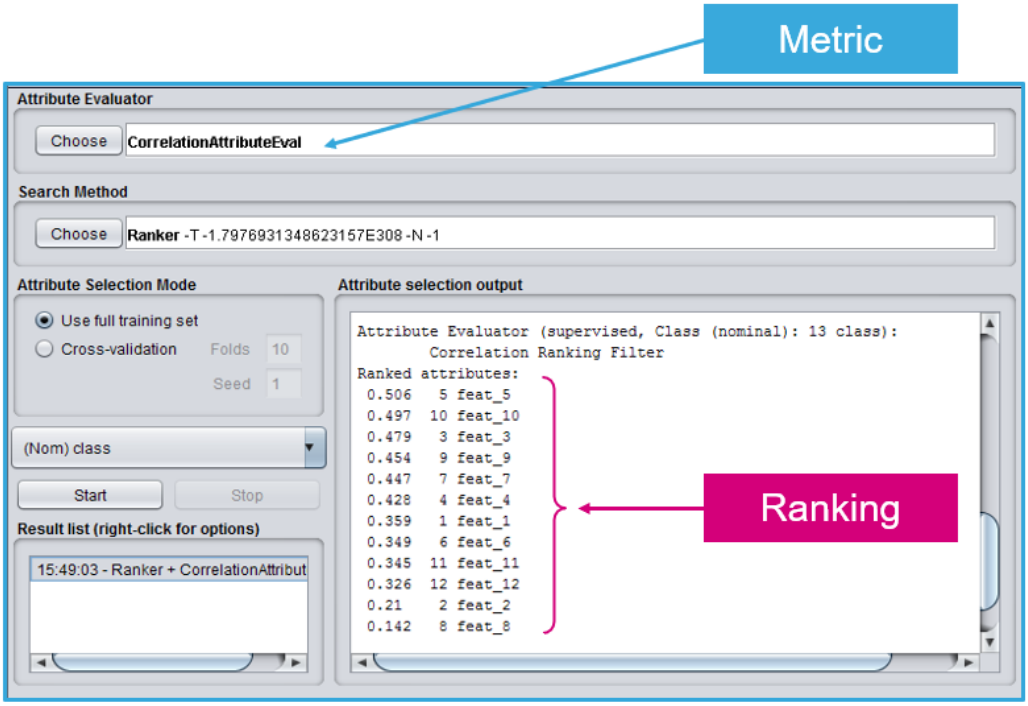
Figure 18. Visualization of two features and two classes



4.14.3 Ranking of features

Different machine learning tools offer automated methods to order features in terms of the information contribution. This form of output ranking is based on criteria/metrics such as correlation, information gain, probabilistic distance, entropy, and more. An example is given by Weka, which automatically handles the calculations needed to generate optimal decision trees as indicated in the figure below.

Figure 19. Ranking from automated output tool



Note that different features could share the same information contribution. This can be evaluated again by visualizing the single feature or by checking the accuracy obtained with the subset of features taken one-by-one, and together, as explained in previous sections.

A final consideration can be done on the number of features that have been selected. In general, the higher the number of features selected:

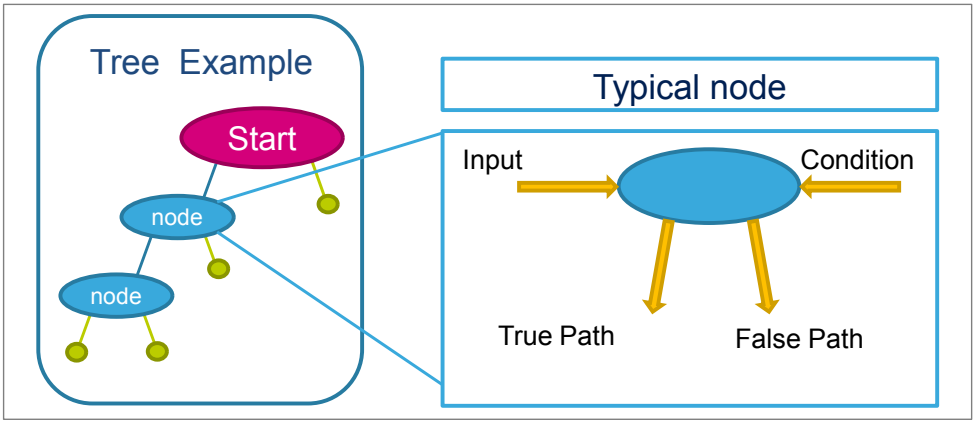
- The higher the risk of overfitting
- The larger the size of the resulting decision tree

5 Decision tree

The decision tree is the predictive model built from the training data, which can be stored in the ISM6HG256X. The training data are the data logs acquired for each class to be recognized (in the activity recognition example the classes might be stationary, walking, jogging, biking, driving, and so forth).

The outputs of the computation blocks described in the previous sections are the inputs of the decision tree. Each node of the decision tree contains a condition, where a feature is evaluated against a certain threshold. If the condition is true, the next node in the true path is evaluated. If the condition is false, the next node in the false path is evaluated. The status of the decision tree evolves node by node until a result is found. The result of the decision tree is one of the classes defined at the beginning of the data collection.

Figure 20. Decision tree node



The decision tree generates a new result at the end of every time window (the parameter "window length" set by the user for the computation of the features). Window length is expressed as a number of samples. The time window can be obtained by dividing the number of samples by the data rate chosen for MLC (MLC_ODR):

Time window = window length / MLC_ODR

For instance, selecting 120 samples for the window length and 120 Hz for the MLC data rate, the obtained time window is:

Time window = 120 samples / 120 Hz = 1 second

The decision tree results can also be filtered by an additional (optional) filter called "metaclassifier", which is described in Section 6: Metaclassifier.

The machine learning core results (decision tree results filtered or not filtered by the metaclassifier) are accessible through dedicated registers in the embedded functions registers (as shown in Table 7). These registers can be continuously read (polled) to check the decision tree outputs. The register MLC_STATUS_MAINPAGE (4Bh) or the embedded functions register MLC_STATUS (15h) contains the interrupt status bits of the eight possible decision trees. These bits are automatically set to 1 when the corresponding decision tree value changes. Furthermore, the interrupt status signal generated using these bits can also be driven to the INT1 pin by setting the MLC_INT1 (0Dh) embedded functions register, or to the INT2 pin by setting the MLC_INT2 (11h) embedded functions register (Table 8). Using the interrupt signals, an MCU performing other tasks or sleeping (to save power), can be awakened when the machine learning core result has changed.

The machine learning core interrupt signal is pulsed by default. The duration of the pulse is equal to 1 / MAX_RATE seconds, where MAX_RATE denotes the maximum rate of the enabled embedded functions.

The machine learning core interrupt signal can also be set latched through the bit EMB_FUNC_LIR in the embedded functions register PAGE_RW (17h).

Table 7. Decision tree results

Register	Content
MLC1_SRC (70h)	Result of decision tree 1
MLC2_SRC (71h)	Result of decision tree 2
MLC3_SRC (72h)	Result of decision tree 3
MLC4_SRC (73h)	Result of decision tree 4
MLC5_SRC (74h)	Result of decision tree 5
MLC6_SRC (75h)	Result of decision tree 6
MLC7_SRC (76h)	Result of decision tree 7
MLC8_SRC (77h)	Result of decision tree 8

Table 8. Decision tree interrupts

Register	Content
MLC_STATUS_MAINPAGE (4Bh)	Contains interrupt status bits for changes in the decision tree result
MLC_STATUS (15h)	Contains interrupt status bits for changes in the decision tree result
MLC_INT1 (0Dh)	Allows routing the interrupt status bits for decision trees to the INT1 pin(1)
MLC_INT2 (11h)	Allows routing the interrupt status bits for decision trees to the INT2 pin(2)

1. *Routing is established if the INT1_EMB_FUNC bit of MD1_CFG (5Eh) is set to 1.*
2. *Routing is established if the INT2_EMB_FUNC bit of MD2_CFG (5Fh) is set to 1.*

5.1 Decision tree limitations in the ISM6HG256X

The ISM6HG256X has limited resources for the machine learning core in terms of number of decision trees, size of the trees, and number of decision tree results.

Up to eight different decision trees can be stored in the ISM6HG256X, but the sum of the number of nodes for all the decision trees must not exceed 256 (*). Every decision tree can have up to 16 results in the ISM6HG256X.

(*) This number might also be limited by the number of features and filters configured. In general, if using few filters and features, there is no further limitation on the size of the decision tree. However, when using many filters and features, the maximum number of nodes for the decision trees is slightly limited. The software tool informs the user of the available nodes for the decision tree.

The table below summarizes the limitations of the ISM6HG256X.

Table 9. Decision tree limitations in the ISM6HG256X

	ISM6HG256X
Maximum number of decision trees	8
Maximum number of nodes (total number for all the decision trees)	256 (*)
Maximum number of results per decision tree	16

Note: When using multiple decision trees, all the parameters described in the previous sections (inputs, filters, features computed in the time window, the time window itself, and also the data rates) are common for all the decision trees.

6 Metaclassifier

A metaclassifier is a filter on the outputs of the decision tree. The metaclassifier uses some internal counters in order to filter the decision tree outputs.

Decision tree outputs can be divided in subgroups (for example, similar classes can be managed in the same subgroup). An internal counter is available for all the subgroups of the decision tree outputs. The counter for the specific subgroup is increased when the result of the decision tree corresponds to one of the classes in the subgroup and it is decreased otherwise. When the counter exceeds a defined value, which is called “end counter” (set by the user), the output of the machine learning core is updated. Values allowed for end counters are from 0 to 14. Each counter can be increased until it exceeds the corresponding end counter, that is, up to the end counter value plus 1.

Table 10. Metaclassifier example

Decision tree result	A	A	A	B	A	B	B	B	A	B	B	B	A	A	A
Counter A (End counter = 2)	1	2	**3**	2	3	2	1	0	1	0	0	0	1	2	**3**
Counter B (End counter = 3)	0	0	0	1	0	1	2	3	2	3	**4**	4	3	2	1
Machine learning core result (including metaclassifier)	x	x	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**A**	**B**	**B**	**B**	**B**	**A**

The previous table shows the effect of filtering the decision tree outputs through a metaclassifier. The first line of the table contains the outputs of the decision tree before the metaclassifier. Counter A and counter B are the internal counters for the two decision tree results (“A” and “B”). In the activity recognition example, the result “A” might be walking and the result “B” jogging. When the internal counter “A” reaches value 3 (which exceeds the end counter for counter “A”), there is a transition to result “A”. When the internal counter “B” reaches value 4 (which exceeds the end counter for counter "B"), there is a transition to result “B”.

The purpose of the metaclassifier is to reduce the false positives, in order to avoid generating an output that is still not stable, and to reduce the transitions on the decision tree result.

6.1 Metaclassifier subgroups in the ISM6HG256X

The metaclassifier has a limited number of subgroups, four subgroups can be used in the ISM6HG256X. The subgroups are fixed and equally subdivide the available decision tree results (subgroup 1 includes results from 0 to 3, subgroup 2 includes results from 4 to 7, and so on). Similar classes may need to be grouped in the same subgroup to use the metaclassifier.

Table 11. Metaclassifier subgroups in the ISM6HG256X

	ISM6HG256X
Maximum number of results per decision tree	16
Result subgroups for metaclassifier per decision tree	4

Note: Multiple metaclassifiers can be configured. One metaclassifier is available for each decision tree configured in the machine learning core.

7 Finite state machine interface

The ISM6HG256X also provides a configurable finite state machine that is suitable for deductive algorithms and in particular gesture recognition.

Finite state machines and decision trees can be combined to work together in order to enhance the accuracy of motion detection.

The decision tree results generated by the machine learning core can be checked by the finite state machine available in the ISM6HG256X: this is possible through the condition CHKDT (described in the application note AN6354 ISM6HG256X: finite state machine). MLC filters and features can also be used by the finite state machines: the identifiers for filters and features are available in the configuration file (.json / .h) generated by STMicroelectronics tools for the MLC (see Figure 22).

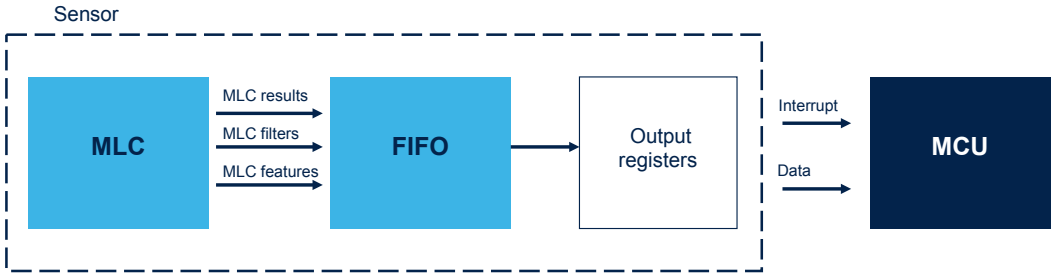
By default, the FSM programs are executed before the MLC algorithms. However, it is possible to change this order through the bit MLC_BEFORE_FSM in the embedded functions register EMB_FUNC_EN_A (04h).

*Note: The MLC and FSM share some part of memory. When using the MLC and FSM together, it is important to properly combine the two configurations by using the merge tool integrated in MEMS Studio. You can find it in the [**Advanced Features**] [**Merge MLC + FSM**] tab of the MEMS Studio main window.*

8 MLC data in FIFO

The ISM6HG256X allows storing MLC data in FIFO.

Figure 21. Storing MLC data in FIFO



- **MLC results** can be stored in FIFO every time that the output of any decision tree changes. One data word is stored containing the MLCx_SRC register corresponding to the decision tree whose output has changed and the decision tree index from 0 to 7. The tag is 1Ah.
- **MLC filters** can be stored in FIFO at every sample processed by the MLC (MLC_ODR). One data word is stored in FIFO for each axis (three words if the MLC filter is configured for XYZ). The tag is 1Bh.
- **MLC standard features** can be stored in FIFO at the end of every time window (window length set by the user). One data word is stored in FIFO and the tag is 1Ch.
- **MLC recursive features** can be stored in FIFO at every sample processed by the MLC (MLC_ODR). Three data words are stored in FIFO and the tag is 1Ch.

Table 12 shows each FIFO entry of the different cases. Identifiers for filters and features are available in the configuration file (.json / .h) generated by STMicroelectronics tools for the MLC (see Figure 22).

Table 12. Data in FIFO

	TAG	X_L	X_H	Y_L	Y_H	Z_L	Z_H
MLC result	1Ah	MLCx_SRC	Index of MLCx_SRC(1)	Timestamp			
MLC filter	1Bh	Value(2)		Filter identifier(3)		Reserved	
MLC feature	1Ch	Value(2)		Feature identifier(3)		Reserved	

1. *MLCx_SRC registers are indexed from 0 to 7 (for example, MLC1_SRC is indexed as 0).*
2. *The value is represented as half-precision floating-point.*
3. *Filter and feature identifiers are indicated in the .json / .h file.*

Figure 22. Identifiers for filters and features

```
"mlc_identifiers": [
    { "fifo_tag": "0x1B", "id": "0x0384", "label": "FILTER_IIR1_ACC_X" },
    { "fifo_tag": "0x1B", "id": "0x0386", "label": "FILTER_IIR1_ACC_Y" },
    { "fifo_tag": "0x1B", "id": "0x0388", "label": "FILTER_IIR1_ACC_Z" },
    { "fifo_tag": "0x1C", "id": "0x038A", "label": "F1_MEAN_ACC_X" },
    { "fifo_tag": "0x1C", "id": "0x038C", "label": "F2_MEAN_ACC_Y" },
    { "fifo_tag": "0x1C", "id": "0x038E", "label": "F3_MEAN_ACC_Z" }
]
```

The following table shows the bits to globally enable batching results and filters/features in FIFO.

Bit	Register	Result
MLC_FIFO_EN	EMB_FUNC_FIFO_EN_A (44h)	It enables batching machine learning core results in the FIFO buffer
MLC_FILTER_FEATURE_FIFO_EN	EMB_FUNC_FIFO_EN_B (45h)	It globally enables batching machine learning core filters and features in the FIFO buffer (Filters and features to be stored in FIFO must be specifically selected when generating the MLC configuration)

Appendix A Glossary

This section contains a glossary of terms used in machine learning. Most of the terms have been taken from https://developers.google.com/machine-learning/glossary/.

ARFF	An ARFF (attribute-relation file format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files were developed by the Machine Learning Project at the Department of Computer Science of The University of Waikato for use with the Weka machine learning software.
Attribute/feature	An attribute is an aspect of an instance (for example, temperature, humidity). Attributes are often called features in machine learning. A special attribute is the class label that defines the class this instance belongs to (required for supervised learning).
Binary classification	A type of classification task that outputs one of two mutually exclusive classes. For example, a machine learning model that evaluates email messages and outputs either "spam" or "not spam" is a binary classifier.
Class	One of a set of enumerated target values for a label. For example, in a binary classification model that detects spam, the two classes are spam and not spam. In a multi-class classification model that identifies dog breeds, the classes would be poodle, beagle, pug, and so on.
Classification model	A type of machine learning model for distinguishing among two or more discrete classes. For example, a natural language processing classification model could determine whether an input sentence was in French, Spanish, or Italian.
Classification threshold	A scalar-value criterion that is applied to a model's predicted score in order to separate the positive class from the negative class. Used when mapping logistic regression results to binary classification. For example, consider a logistic regression model that determines the probability of a given email message being spam. If the classification threshold is 0.9, then logistic regression values above 0.9 are classified as spam and those below 0.9 are classified as not spam.
Class-imbalanced dataset	A binary classification problem in which the labels for the two classes have significantly different frequencies. For example, a disease dataset in which 0.0001 of examples have positive labels and 0.9999 have negative labels is a class-imbalanced problem, but a football game predictor in which 0.51 of examples label one team winning and 0.49 label the other team winning is not a class-imbalanced problem.
Clipping	A technique for handling outliers. Specifically, reducing feature values that are greater than a set maximum value down to that maximum value. Also, increasing feature values that are less than a specific minimum value up to that minimum value.
Confusion matrix	An NxN table that summarizes how successful the classification model predictions were; that is, the correlation between the label and the model classification. One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label.
Cross-validation	A mechanism for estimating how well a model generalizes to new data by testing the model against one or more non-overlapping data subsets withheld from the training set.
Data analysis	Obtaining an understanding of data by considering samples, measurement, and visualization. Data analysis can be particularly useful when a dataset is first received, before one builds the first model. It is also crucial in understanding experiments and debugging problems with the system.
Data augmentation	Artificially boosting the range and number of training examples by transforming existing examples to create additional examples. For example, suppose images are one of your features, but your dataset doesn't contain enough image examples for the model to learn useful associations. Ideally, you'd add enough labeled images to your dataset to enable your model to train properly. If that's not possible, data augmentation can rotate, stretch, and reflect each image to produce many variants of the original picture, possibly yielding enough labeled data to enable excellent training.
Data set or dataset	A collection of examples.
Decision boundary	The separator between classes learned by a model in a binary class or multi-class classification problems.
Decision threshold	Synonym for classification threshold.
Decision tree	A model represented as a sequence of branching statements.
Discrete feature	A feature with a finite set of possible values. For example, a feature whose values may only be animal, vegetable, or mineral is a discrete (or categorical) feature. Contrast with continuous feature.
Discriminative model	A model that predicts labels from a set of one or more features. More formally, discriminative models define the conditional probability of an output given the features and weights.

Downsampling	Overloaded term that can mean either of the following: • Reducing the amount of information in a feature in order to train a model more efficiently. • Training on a disproportionately low percentage of over-represented class examples in order to improve model training on under-represented classes.
Dynamic model	A model that is trained online in a continuously updating fashion. That is, data is continuously entering the model.
Example	One row of a dataset. An example contains one or more features and possibly a label. See also labeled example and unlabeled example.
False negative (FN)	An example in which the model mistakenly predicted the negative class. For example, the model inferred that a particular email message was not spam (the negative class), but that email message actually was spam.
False positive (FP)	An example in which the model mistakenly predicted the positive class. For example, the model inferred that a particular email message was spam (the positive class), but that email message was actually not spam.
False positive rate (FPR)	The x-axis in an ROC curve. The false positive rate is defined as follows: False positive rate = false positives / (false positives + true negatives)
Feature	An input variable used in making predictions.
Feature engineering	The process of determining which features might be useful in training a model, and then converting raw data from log files and other sources into said features. Feature engineering is sometimes called feature extraction.
Feature extraction	Overloaded term having either of the following definitions: • Retrieving intermediate feature representations calculated by an unsupervised or pre-trained model for use in another model as input. • Synonym for feature engineering.
Feature set	The group of features your machine learning model trains on. For example, postal code, property size, and property condition might comprise a simple feature set for a model that predicts housing prices.
Generalization	Refers to your model's ability to make correct predictions on new, previously unseen data as opposed to the data used to train the model.
Ground truth	The correct answer. Reality. Since reality is often subjective, expert raters typically are the proxy for ground truth.
Heuristic	A quick solution to a problem, which may or may not be the best solution.
Imbalanced dataset	Synonym for class-imbalanced dataset.
Independently and identically distributed (i.i.d)	Data drawn from a distribution that doesn't change, and where each value drawn doesn't depend on values that have been drawn previously. An i.i.d. is the ideal gas of machine learning—a useful mathematical construct but almost never exactly found in the real world. For example, the distribution of visitors to a web page may be i.i.d. over a brief window of time; that is, the distribution doesn't change during that brief window and one person's visit is generally independent of another's visit. However, if you expand that window of time, seasonal differences in the web page's visitors may appear.
Interference	In machine learning, often refers to the process of making predictions by applying the trained model to unlabeled examples. In statistics, inference refers to the process of fitting the parameters of a distribution conditioned on some observed data.
Instance	Synonym for example.
Interpretability	The degree to which a model's predictions can be readily explained. Deep models are often non-interpretable; that is, a deep model's different layers can be hard to decipher. By contrast, linear regression models and wide models are typically far more interpretable.
J48	An open source Java implementation of the C4.5 algorithm
Label	In supervised learning, the "answer" or "result" portion of an example. Each example in a labeled dataset consists of one or more features and a label. For instance, in a housing dataset, the features might include the number of bedrooms, the number of bathrooms, and the age of the house, while the label might be the house's price. In a spam detection dataset, the features might include the subject line, the sender, and the email message itself, while the label would probably be either "spam" or "not spam."
Linear regression	A type of regression model that outputs a continuous value from a linear combination of input features.

Machine learning	A program or system that builds (trains) a predictive model from input data. The system uses the learned model to make useful predictions from new (never-before-seen) data drawn from the same distribution as the one used to train the model. Machine learning also refers to the field of study concerned with these programs or systems.
Majority class	The more common label in a class-imbalanced dataset. For example, given a dataset containing 99% non-spam labels and 1% spam labels, the non-spam labels are the majority class.
Matplotlib	An open-source Python 2D plotting library. matplotlib helps you visualize different aspects of machine learning.
Minority class	The less common label in a class-imbalanced dataset. For example, given a dataset containing 99% non-spam labels and 1% spam labels, the spam labels are the minority class.
ML	Abbreviation for machine learning.
Model training	The process of determining the best model.
Multi-class classification	Classification problems that distinguish among more than two classes. For example, there are approximately 128 species of maple trees, so a model that categorized maple tree species would be multi-class. Conversely, a model that divided emails into only two categories (spam and not spam) would be a binary classification model.
Multinomial classification	Synonym for multi-class classification.
Negative class	In binary classification, one class is termed positive and the other is termed negative. The positive class is the thing we're looking for and the negative class is the other possibility. For example, the negative class in a medical test might be "not tumor." The negative class in an email classifier might be "not spam." See also positive class.
Neural network	A model that, taking inspiration from the brain, is composed of layers (at least one of which is hidden) consisting of simple connected units or neurons followed by nonlinearities.
Node (decision tree)	A "test" on an attribute.
Noise	Broadly speaking, anything that obscures the signal in a dataset. Noise can be introduced into data in a variety of ways. For example: • Human raters make mistakes in labeling. • Humans and instruments mis-record or omit feature values.
Normalization	The process of converting an actual range of values into a standard range of values, typically -1 to +1 or 0 to 1. For example, suppose the natural range of a certain feature is 800 to 6,000. Through subtraction and division, you can normalize those values into the range -1 to +1. See also scaling.
Numerical data	Features represented as integers or real-valued numbers.
Outliers	Values distant from most other values. In machine learning, any of the following are outliers: • Weights with high absolute values. • Predicted values relatively far away from the actual values. • Input data whose values are more than roughly 3 standard deviations from the mean. Outliers often cause problems in model training. Clipping is one way of managing outliers.
Overfitting	Creating a model that matches the training data so closely that the model fails to make correct predictions on new data.
Parameter	A variable of a model that the ML system trains on its own.
Performance	Overloaded term with the following meanings: • The traditional meaning within software engineering. Namely: How fast (or efficiently) does this piece of software run? • The meaning within ML. Here, performance answers the following question: How correct is this model? That is, how good are the model's predictions?
Positive class	In binary classification, the two possible classes are labeled as positive and negative. The positive outcome is the thing we're testing for. (Admittedly, we're simultaneously testing for both outcomes, but play along.) For example, the positive class in a medical test might be "tumor." The positive class in an email classifier might be "spam." Contrast with negative class.

Precision	A metric for classification models. Precision identifies the frequency with which a model was correct when predicting the positive class. That is: Precision = true positives / (true positives + false positives)
Prediction	A model's output when provided with an input example.
Pre-trained model	Models or model components that have been already been trained.
Proxy labels	Data used to approximate labels not directly available in a dataset. For example, suppose you want “is it raining?” to be a Boolean label for your dataset, but the dataset doesn't contain rain data. If photographs are available, you might establish pictures of people carrying umbrellas as a proxy label for “is it raining”? However, proxy labels may distort results. For example, in some places, it may be more common to carry umbrellas to protect against sun than the rain.
Rater	A human who provides labels in examples. Sometimes called an "annotator."
Recall	A metric for classification models that answers the following question: “Out of all the possible positive labels, how many did the model correctly identify?” That is: Recall = true positives / (true positives + false negatives)
Regression model	A type of model that outputs continuous (typically, floating-point) values. Compare with classification models, which output discrete values, such as "day lily" or "tiger lily."
Reinforcement learning	A machine learning approach to maximize an ultimate reward through feedback (rewards and punishments) after a sequence of actions. For example, the ultimate reward of most games is victory. Reinforcement learning systems can become expert at playing complex games by evaluating sequences of previous game moves that ultimately led to wins and sequences that ultimately led to losses.
Representation	The process of mapping data to useful features.
ROC curve	ROC = receiver operating characteristic A curve of true positive rate vs. false positive rate at different classification thresholds.
Scaling	A commonly used practice in feature engineering to tame a feature's range of values to match the range of other features in the dataset. For example, suppose that you want all floating-point features in the dataset to have a range of 0 to 1. Given a particular feature's range of 0 to 500, you could scale that feature by dividing each value by 500. See also normalization.
Scikit-learn	A popular open-source ML platform. See www.scikit-learn.org.
Scoring	The part of a recommendation system that provides a value or ranking for each item produced by the candidate generation phase.
Semi-supervised learning	Training a model on data where some of the training examples have labels but others don’t. One technique for semi-supervised learning is to infer labels for the unlabeled examples, and then to train on the inferred labels to create a new model. Semi-supervised learning can be useful if labels are expensive to obtain but unlabeled examples are plentiful.
Sequence model	A model whose inputs have a sequential dependence. For example, predicting the next video watched from a sequence of previously watched videos.
Serving	A synonym for inferring.
Static model	A model that is trained offline.
Stationarity	A property of data in a dataset, in which the data distribution stays constant across one or more dimensions. Most commonly, that dimension is time, meaning that data exhibiting stationarity doesn't change over time. For example, data that exhibits stationarity doesn't change from September to December.
Supervised machine learning	Training a model from input data and its corresponding labels. Supervised machine learning is analogous to a student learning a subject by studying a set of questions and their corresponding answers. After mastering the mapping between questions and answers, the student can then provide answers to new (never-before-seen) questions on the same topic. Compare with unsupervised machine learning.
Target	Synonym for label.
Training	The process of determining the ideal parameters comprising a model.
Training set	The subset of the dataset used to train a model.

	Contrast with validation set and test set.
True negative (TN)	An example in which the model correctly predicted the negative class. For example, the model inferred that a particular email message was not spam, and that email message really was not spam.
True positive (TP)	An example in which the model correctly predicted the positive class. For example, the model inferred that a particular email message was spam, and that email message really was spam.
True positive rate (TPR)	Synonym for recall. That is: True positive rate = true positives / (true positives + false negatives) True positive rate is the y-axis in an ROC curve.
Underfitting	Producing a model with poor predictive ability because the model hasn't captured the complexity of the training data. Many problems can cause underfitting, including: • Training on the wrong set of features. • Training for too few epochs or at too low a learning rate. • Training with too high a regularization rate. • Providing too few hidden layers in a deep neural network.
Unlabeled example	An example that contains features but no label. Unlabeled examples are the input to inference. In semi-supervised and unsupervised learning, unlabeled examples are used during training.
Unsupervised machine learning	Training a model to find patterns in a dataset, typically an unlabeled dataset. The most common use of unsupervised machine learning is to cluster data into groups of similar examples. For example, an unsupervised machine learning algorithm can cluster songs together based on various properties of the music. The resulting clusters can become an input to other machine learning algorithms (for example, to a music recommendation service). Clustering can be helpful in domains where true labels are hard to obtain. For example, in domains such as anti-abuse and fraud, clusters can help humans better understand the data. Another example of unsupervised machine learning is principal component analysis (PCA). For example, applying PCA on a dataset containing the contents of millions of shopping carts might reveal that shopping carts containing lemons frequently also contain antacids. Compare with supervised machine learning.
Validation	A process used, as part of training, to evaluate the quality of a machine learning model using the validation set. Because the validation set is disjoint from the training set, validation helps ensure that the model's performance generalizes beyond the training set. Contrast with test set.
Validation set	A subset of the dataset—disjoint from the training set—used in validation. Contrast with training set and test set.
Weka	A collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.

Revision history

Table 13. **Document revision history**

Date	Version	Changes
06-Oct-2025	1	Initial release

Contents

List of tables

List of figures

IMPORTANT NOTICE – READ CAREFULLY

STMicroelectronics NV and its subsidiaries (“ST”) reserve the right to make changes, corrections, enhancements, modifications, and improvements to ST products and/or to this document at any time without notice.

In the event of any conflict between the provisions of this document and the provisions of any contractual arrangement in force between the purchasers and ST, the provisions of such contractual arrangement shall prevail.

The purchasers should obtain the latest relevant information on ST products before placing orders. ST products are sold pursuant to ST’s terms and conditions of sale in place at the time of order acknowledgment.

The purchasers are solely responsible for the choice, selection, and use of ST products and ST assumes no liability for application assistance or the design of the purchasers’ products.

No license, express or implied, to any intellectual property right is granted by ST herein.

Resale of ST products with provisions different from the information set forth herein shall void any warranty granted by ST for such product.

If the purchasers identify an ST product that meets their functional and performance requirements but that is not designated for the purchasers' market segment, the purchasers shall contact ST for more information.

ST and the ST logo are trademarks of ST. For additional information about ST trademarks, refer to www.st.com/trademarks. All other product or service names are the property of their respective owners.

Information in this document supersedes and replaces information previously supplied in any prior versions of this document.