Model Name	Audio Classifier
Overview	This document is a FactSheet accompanying the Audio Classifier model on IBM Developer Model Asset eXchange.
Purpose	This model classifies an input audio clip.
Intended Domain	This model is intended for use in the audio processing and classification domain.
Training Data	The model is trained on the AudioSet dataset by Google.
Model Information	The audio classifier is a two-stage model: • The first model (MAX-Audio-Embedding-Generator) converts each second of input raw audio into vectors or embeddings of size 128 where each element of the vector is a float between 0 and 1. • Once the vectors are generated, there is a second deep neural network that performs classification.
Inputs and Outputs	Input : a 10 second clip of audio in signed 16-bit PCM wavfile format. Output : a JSON with the top 5 predicted classes and probabilities.
Performance Metrics	Metric: Mean Average Precision — Value: 0.357 Metric: Area Under the Curve — Value: 0.968 Metric: d-prime — Value: 2.621
Bias	The majority of audio samples in the training data set represent voice and music content. Potential bias caused by this over-representation has not been evaluated. Careful attention should be paid if this model is to be incorporated in an application where bias in voice type or music genre is potentially sensitive or harmful.
Robustness	This audio classifier is not robust to the L-infinity and L2 norms for the HopSkipJump attack.

	L2	L-Infinity
5th Percentile	887.0 (200.9)	5.5 (4.9)
10th Percentile	1496.6 (720.6)	7.53 (5.73)
15th Percentile	3723.1 (4707.2)	52.8 (41.8)
25th Percentile	7187.9 (---)	187.6 (198.1)
50th Percentile	11538.6 (---)	502.8 (---)

The susceptibility of the model to the two attacks. The parenthetical values in the table above represent the fitted curve evaluated at 11 iterations. (When we are unable to fit a curve, or the result is negative, we denote by ---.)

Domain Shift	No domain shift evaluation occurred.
Test Data	The test set is also part of the AudioSet data. There was a `70:20:10`% split of the data into `train:val:test`. The ratio of samples/class was maintained as much as possible in all the splits.
Optimal Conditions	• When the input audio contains only one or two distinct audio classes. • When the audio quality is high with lesser noise.
Poor Conditions	• When the audio contains more that two distinct classes. • When the audio quality is low with more noise.
Explanation	While the model architecture is well documented, the model is still a deep neural network, which largely remains a black box when it comes to explainability of results and predictions.
Contact Information	Any queries related to the operation of the MAX Audio Classifier model can be addressed on the model GitHub repo.