AI FactSheet for Third Party Systems

Please provide details regarding your Artificial Intelligence (AI) product by filling out the FactSheet[1] template below. This information will be kept in the GovAI Coalition vendor registry, and may be made publicly available. You can find an example of a completed FactSheet on page 4.

AI FactSheet

Vendor Name	
System Name	
Overview	Brief summary of the AI system.
Purpose	What function does the AI system perform, and for what purpose? If the system performs multiple functions, list each discretely and reference below. For features that are configurable, please describe all configuration options and default settings.
Intended Domain	What domain is the AI system intended to be applied in?
Training Data	How was the AI system trained? What data was used? How often is data added to the training set? Was all training data legally obtained and its use fully licensed?
Test Data	What data was used to test system performance? Under what conditions has the system been tested?
Model Information	General description of the model(s) used (e.g., large language model, transformer, deep learning, supervised learning, built on an existing open source model, computer vision)
Update procedure	In general, how often are the models updated for users? Will the user have a choice in moving to the updated model or staying on the current model? What documentation is available for new versions of the model?
Inputs and Outputs	What are the inputs to the AI system? What are its outputs? What interfaces and integrations are supported?
Performance Metrics	What are the performance metrics? What is your current level of performance on these metrics? How can the user monitor performance in the deployment environment?

[1] The FactSheet template is heavily inspired by the IBM Research AI FactSheets 360 project.

Bias	What biases does the tool exhibit and how does it handle that bias? This can include but is not limited to biases on human factors such as gender, race, socioeconomic status, disability, culture, age, or other protected classes, or biases on general factors such as a sampling bias, survivorship bias, detection bias, or observer bias.
Robustness	How does the AI system handle outliers? Do overwritten decisions feed back into the system to help calibrate it in the future?
Optimal Conditions	What conditions does the model perform best under? Are there minimum requirements for the quantity of records/observations?
Poor Conditions	What conditions does the model perform poorly under? What are the limitations of the AI system? What kinds of errors can it make (e.g., hallucinations) and what conditions make those errors more likely?
Explanation	How does the AI system explain its predictions? Are the outcomes of the AI system understandable by subject matter experts, users, impacted individuals, and others?
Jurisdiction-specific Considerations	Please describe any considerations relevant to local, state, industry, or other specific jurisdictional regulations.

Algorithmic Impact Assessment Questionnaire

How is the AI tool monitored to identify any problems in usage? Can outputs (recommendations, predictions, etc.) be overwritten by a human, and do overwritten outputs help calibrate the system in the future?	Problems in usage can include false negatives, false positives, bias, hallucinations, and human-reported quality issues (such as poor translations or poorly generated images).
How is bias managed effectively?	This can include ways to monitor bias, or abilities to toggle parameters to change observed bias in the model.

Have the vendors or an independent party conducted a study on the bias, accuracy, or disparate impact of the system? If yes, can the City of San José review the study? Include methodology and results.	This can include bias impact reports, algorithmic impact reports, or others.[2]
How can the City of San José and its partners flag issues related to bias, discrimination, or poor performance of the AI system?	This can include ways to report inaccurate or concerning decisions/classifications made by the AI system, or ways to retroactively review past system actions.
How has the Human-Computer Interaction aspect of the AI tool been made accessible, such as to people with disabilities?	Has it been assessed against any usability standards, and if so what was the result?
Please share any relevant information, links, or resources regarding your organization's responsible AI strategy.	URL to any broad AI policy or strategy.

[2] See "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms" for an example bias impact report template: https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.

Example FactSheet[3]

This is an example of the AI FactSheet above completed by a fictitious company. This is only here for reference and does not need to be included in the completed form.

<table>
<tr><td><b>Vendor Name</b></td><td>XYZ Technologies, Inc.</td></tr>
<tr><td><b>System Name</b></td><td>Audio Classifier</td></tr>
<tr><td><b>Overview</b></td><td>This document is a FactSheet accompanying the Audio Classifier model on IBM Developer Model Asset eXchange.</td></tr>
<tr><td><b>Purpose</b></td><td>This model classifies an input audio clip.</td></tr>
<tr><td><b>Intended Domain</b></td><td>This model is intended for use in the audio processing and classification domain.</td></tr>
<tr><td><b>Training Data</b></td><td>The model is trained on the AudioSet dataset by Google. New data is added to the training set daily. The AudioSet database was legally obtained and its use is fully licensed.</td></tr>
<tr><td><b>Test Data</b></td><td>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. The system has been tested in X,Y,Z conditions.</td></tr>
<tr><td><b>Model Information</b></td><td>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.</td></tr>
<tr><td><b>Update procedure</b></td><td>In general, the model is updated annually. If the user does not wish to move to the updated model, the user cannot continue to use the system. Documentation for all new versions of the model can be found on the website at this link.</td></tr>
<tr><td><b>Inputs and Outputs</b></td><td>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.</td></tr>
<tr><td><b>Performance Metrics</b></td><td><table><tr><th>Metric</th><th>Value</th></tr><tr><td>Mean Average Precision</td><td>0.357</td></tr></table></td></tr>
</table>

[3] The example FactSheet is taken from IBM Research AI Factsheet 360's Audio Classifier sample.

Area Under the Curve	0.968
d-prime	2.621

The user can regularly monitor these metrics [here].

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 ---.)

Overwritten decisions are fed back into the system to help calibrate it in the future.

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

The system can misclassify audio:

- When the audio contains more than two distinct classes, and
- 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.

Jurisdiction-specific Considerations

N/A

Algorithmic Impact Assessment Questionnaire

How is the AI tool monitored to identify any problems in usage? Can outputs (recommendations, predictions, etc.) be overwritten by a human, and do overwritten outputs help calibrate the system in the future?	The system can be monitored in usage, and audio classification decisions can be retroactively overwritten by a human. The overwritten decisions can help calibrate the system in the future if desired.
How is bias managed effectively?	Users have access to performance metrics that can be used to understand if the bias in voice-type or music style is harmful.
Have the vendors or an independent party conducted a study on the bias, accuracy, or disparate impact of the system? If yes, can the City of San José review the study? Include methodology and results.	Yes. Results from the third-party study can be provided upon request.
How can the City of San José and its partners flag issues related to bias, discrimination or poor performance of the AI system?	The system provides a web portal to each customer to show the results of the system and its impact on transit performance in the form of reports and graphs.
How has the Human-Computer Interaction aspect of the AI tool been made accessible, such as to people with disabilities?	The system is embedded into a graphics user interface that is compliant with modern screen readers, and provides the option for auto-generated dictation of text on the screen.
Please share any relevant information, links, or resources regarding your organization's responsible AI strategy.	Information about our responsible AI strategy can be found on our website at this link.