Metric Evaluation for Multiclass Computer Vision Model

Jeff Lin

Institute for Defense Analyses (IDA) Science, Systems, and Sustainment Division (S3D)

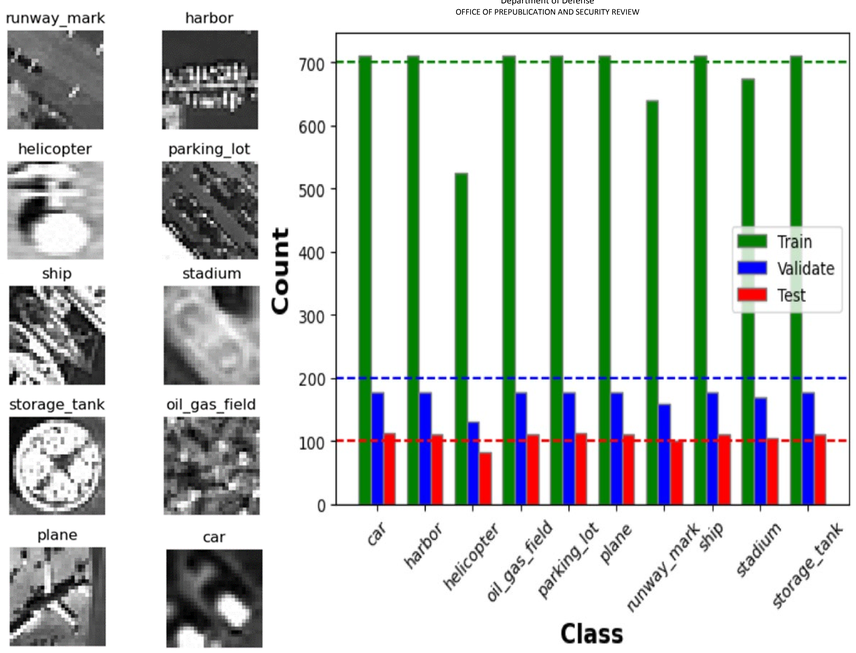
Background and Research Goal

This poster explores strengths and potential risks of popular metrics for evaluating computer vision multiclass classification models to assist the Chief Digital and Artificial Intelligence Office (CDAO) in supporting the test and evaluation of an aided target recognition system. We demonstrate how testers can choose metrics using a tangible overhead MNIST example, highlighting both rollup and per-class results alongside their tradeoffs.

Joint Artificial Intelligence Test Infrastructure Capability (JATIC)

JATIC is a CDAO tool to test and evaluate AI models for performance, effectiveness, robustness, and safety.

Dataset Overview



Key Metrics

Accuracy - Ratio of correctly predicted instances to total number of instances.

- *Intuitively easy to understand but misleading for an imbalanced dataset.*

Precision - Proportion of true positives to all positive predictions.

- *Important when false positives are costly, but good precision in one class does not mean good overall performance.*

Recall - Proportion of true positives to all actual positives.

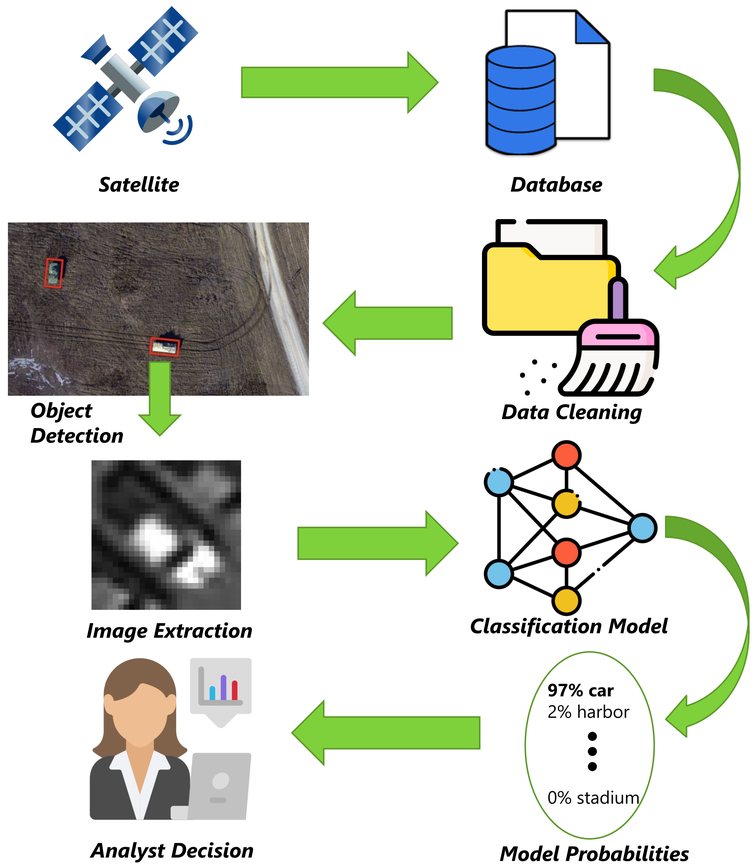
- *Important when false negatives are costly, but good recall in one class does not mean good overall performance.*

F1 - Harmonic means of precision and recall.

- *Balanced evaluation of both precision and recall, but it assumes equal importance and can be difficult to interpret.*

MNIST - Modified National Institute of Standards and Technology

Aided Target Recognition Architecture



Test and Evaluation: Challenges

Testers have no insight into model creation and are expected to evaluate pre-trained models to find the best performer. But how should they identify the best model?

Test and Evaluation of Multiple Models

To demonstrate metric evaluation of model performance, we generated several competing models using the parameters listed below.

Design to Generate Multiple Models	
Parameter	**Value**
Optimization schema	Accuracy, F1 Score
Number of layers	4, 5
Number of blocks per layer	2, 3
Training Dataset	Unmodified, Unmodified Plus Blurred
Testing Dataset	Unmodified, Blurred

Additional models are trained on blurred + original data to assess model robustness.

Model Performance: Average Across All Classes

- Best performance - 2nd best performance - 3rd best performance

Model Performance on Blurred Dataset

	Accuracy	Precision	Recall	F1	Training Time	Validation Time
Original Model	20.20%	25.73%	20.20%	11.73%	216.90 seconds	20.50 seconds
Model 1	23.84%	18.23%	23.84%	15.82%	299.17 seconds	25.02 seconds
Model 2	28.95%	19.89%	28.95%	19.84%	199.43 seconds	16.44 seconds
Model 3	26.74%	21.13%	26.74%	16.05%	171.49 seconds	15.19 seconds
Model 4	21.98%	18.17%	21.98%	10.95%	271.61 seconds	23.17 seconds
Model 5	22.34%	22.64%	22.34%	13.54%	198.26 seconds	16.40 seconds
Blurred Model	81.14%	83.96%	81.14%	81.06%	338.70 seconds	28.35 seconds
Model 7	76.64%	79.53%	76.64%	77.19%	331.41 seconds	28.22 seconds
Model 8	72.71%	76.12%	72.71%	72.20%	346.97 seconds	29.81 seconds
Model 9	80.74%	81.82%	80.74%	80.37%	337.56 seconds	28.51 seconds
Model 10	70.80%	74.40%	70.80%	70.83%	333.75 seconds	27.89 seconds

Blurred model outperforms others across all metrics, making it the best choice.

Model Performance on Unmodified Dataset

	Accuracy	Precision	Recall	F1	Training Time	Validation Time
Original Model	92.84%	93.44%	92.84%	92.95%	216.90 seconds	20.50 seconds
Model 1	91.99%	91.99%	91.99%	91.92%	299.17 seconds	25.02 seconds
Model 2	93.01%	93.18%	93.01%	93.03%	199.43 seconds	16.44 seconds
Model 3	92.88%	93.16%	92.88%	92.97%	171.49 seconds	15.19 seconds
Model 4	92.36%	92.83%	92.36%	92.50%	271.61 seconds	23.17 seconds
Model 5	92.50%	92.82%	92.50%	92.60%	198.26 seconds	16.40 seconds
Blurred Model	57.72%	71.79%	57.72%	59.15%	338.70 seconds	28.35 seconds
Model 7	43.96%	71.31%	43.96%	46.06%	331.41 seconds	28.22 seconds
Model 8	59.15%	77.03%	59.15%	62.29%	346.97 seconds	29.81 seconds
Model 9	52.36%	66.31%	52.36%	52.43%	337.56 seconds	28.51 seconds
Model 10	43.46%	68.60%	43.46%	43.32%	333.75 seconds	27.89 seconds

Some models excel in one area but fall short in another. How can we identify the best model?

Performance and Risk Analysis

	Accuracy	Precision	Recall	F1	Training Time	Validation Time
Original Model	92.84%	93.44%	92.84%	92.95%	216.90 seconds	20.50 seconds
Model 2	93.01%	93.18%	93.01%	93.03%	199.43 seconds	16.44 seconds
Model 3	92.88%	93.16%	92.88%	92.97%	171.49 seconds	15.19 seconds

Focusing only on the averages across all classes can obscure key details and fail to capture the entire picture.

Helicopter Performance

	Accuracy	Precision	Recall	F1
Original Model	81.71%	98.53%	81.71%	89.33%
Model 2	91.46%	93.75%	91.46%	92.59%
Model 3	89.02%	97.33%	89.02%	92.99%

Parking Lot Performance

	Accuracy	Precision	Recall	F1
Original Model	85.71%	96.97%	85.71%	91.00%
Model 2	93.75%	92.92%	93.75%	93.33%
Model 3	87.50%	90.74%	87.50%	89.09%

If the impact of lower accuracy for helicopter and parking lot is minimal, the original model may be preferable given its stronger performance in all other classes aside from helicopter and parking lot.

A holistic evaluation of the best model requires the tester to first understand the mission goals, identify appropriate metrics, and make tradeoffs between these metrics based on mission priorities.

Acknowledgement

Thank you to Dr. Sabrina Dimassimo and June Langley for their mentorship throughout the project.