OAK RIDGE National Laboratory

Toward the Detection of Polyglot Files

Authors: Luke Koch, Sean Oesch, Mary Adkisson, Samantha Erwin, Brian Weber, Amul Chaulagain

UT-Battelle Business Sensitive

ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

What is a Polyglot?

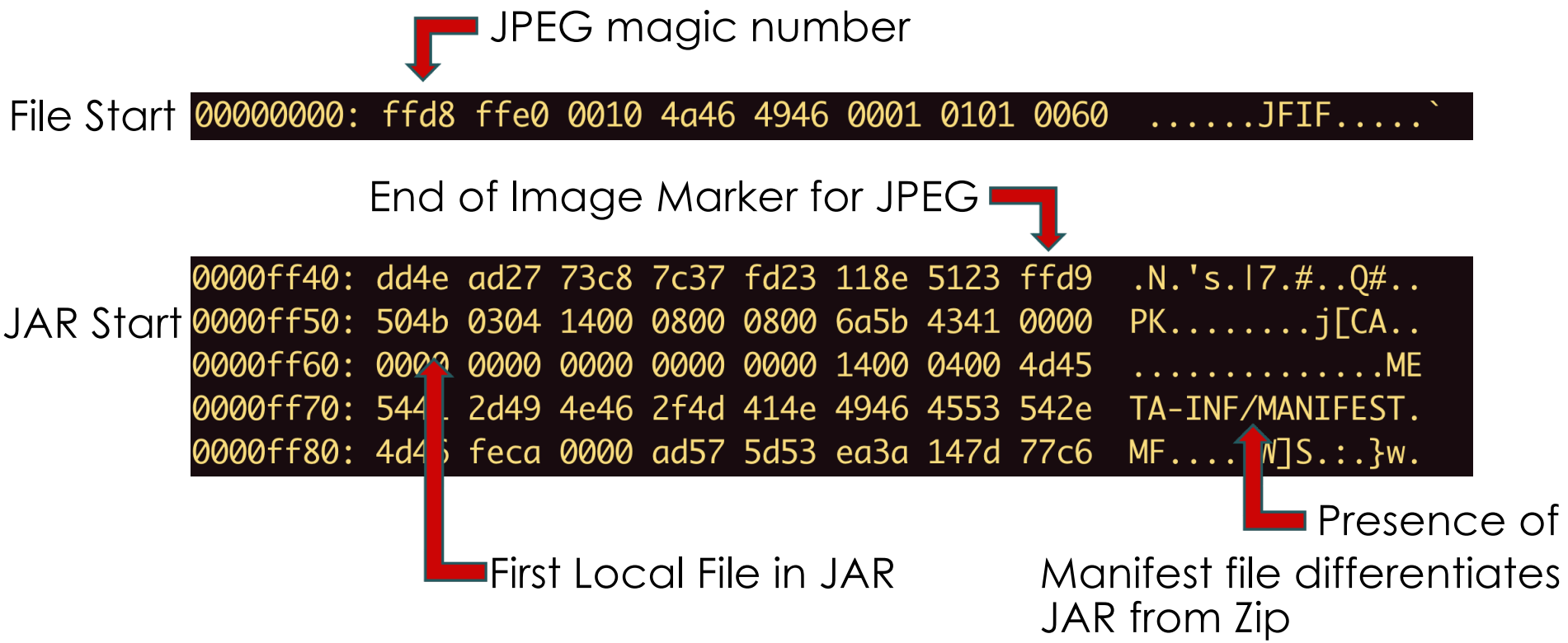
- A single file that is fully valid in two different formats
- Example: JPG+JAR polyglot can execute Java code or yield an image depending on the interpreting program.
- Danger: Interpreting programs simply ignore content from second format
- Large-scale analysis of leading commercial malware detectors revealed failure to detect 100% of polyglot malware in data set[1]

[1]Robert A Bridges, Sean Oesch, Miki E Verma, Michael D Iannacone, Kelly MT Huffer, Brian Jewell, Jeff A Nichols, Brian Weber, Justin M Beaver, Jared M Smith, et al . 2020. Beyond the Hype: A Real-World Evaluation of the Impact and Cost of Machine Learning-Based Malware Detection. arXiv preprint arXiv:2012.09214 (2020).

Generating Polyglots

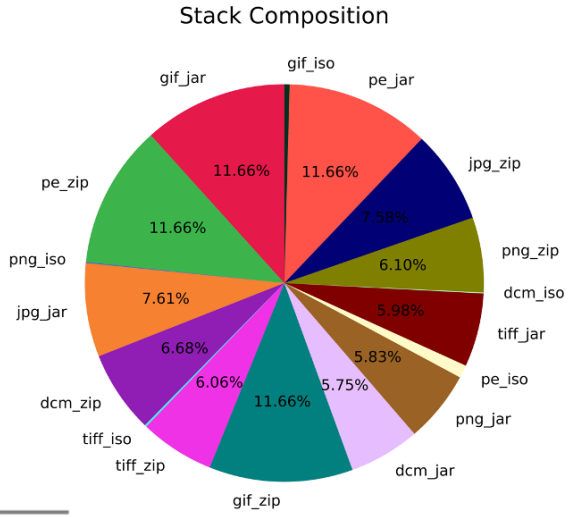
- Open-source *Mitra* tool combines compatible files into polyglots
  - Created by Ange Albertini
  - https://github.com/corkami/mitra
- Creates 4 types of polyglots from pairs of donor files
  - Stacks: file 2 appended to end of file 1
  - Parasites: file 2 placed inside comment markers of file 1
  - Zippers: both files placed within each other's comment markers
  - Cavities: file 2 placed inside padding space of file 1
- *Mitra* attempts all 4 types of combination on each pair of files

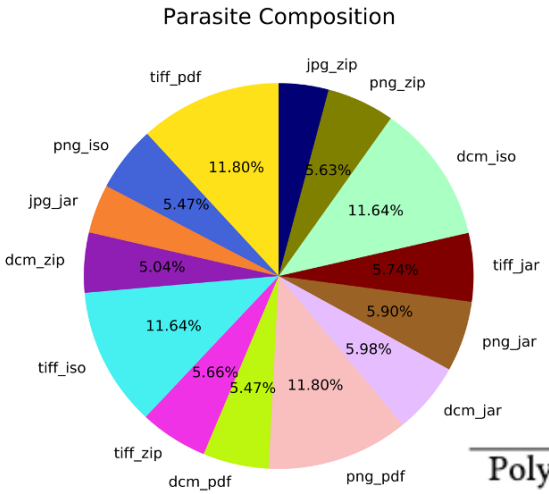
Hexdump of JPEG+JAR Stack-type Polyglot



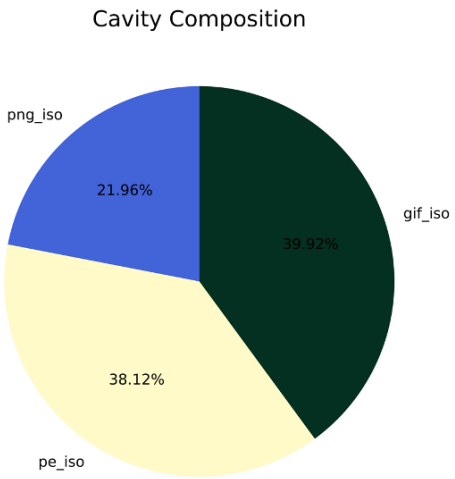
OAK RIDGE National Laboratory

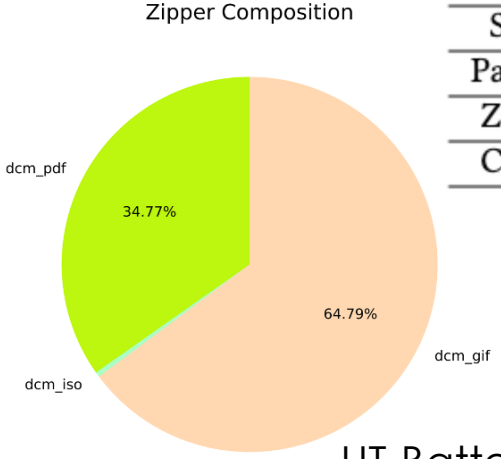
Polyglot Data Set for ML/DL Training





Data Set	Train	Test
Monoglot	31,199	7,799
Polyglot	25,210	6,303





Polyglot Type	Train	Test
Stack	10,267	2,574
Parasite	10,301	2,542
Zipper	1,795	463
Cavity	2,847	724

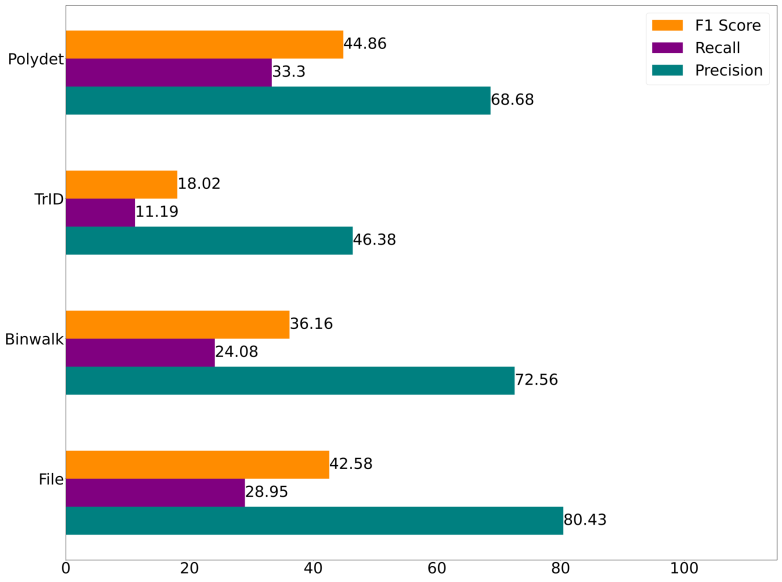
OAK RIDGE National Laboratory

Existing Tools for Polyglot Detection

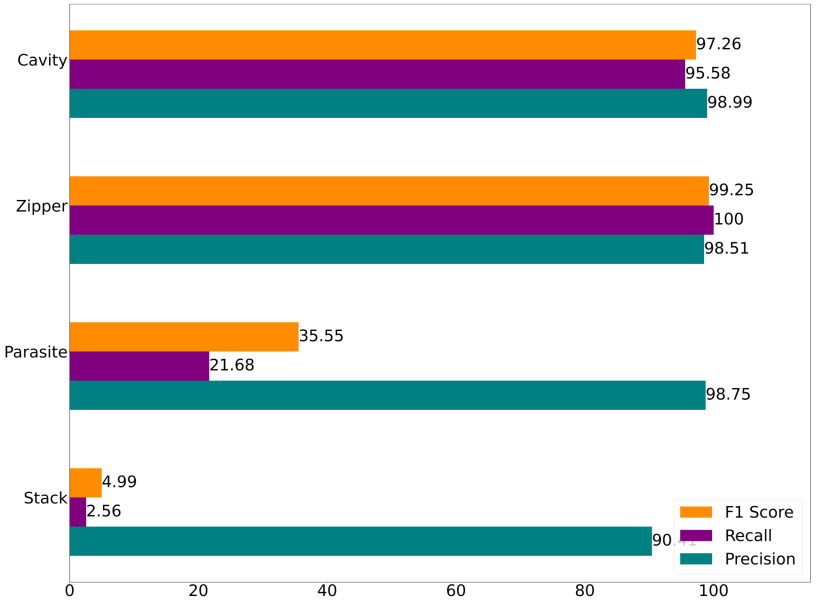
- *File* is the ubiquitous tool for identifying a file's format
- *TrID* is a similar tool used by Virustotal
- *Binwalk*, a file carving utility, has been used to detect polyglots
- *Polyfile*, a DARPA-funded tool, is a utility designed to analyze the structure of abnormal files, including polyglots

Existing Tool Performance

Recall, precision, F1 score for all 4 tools

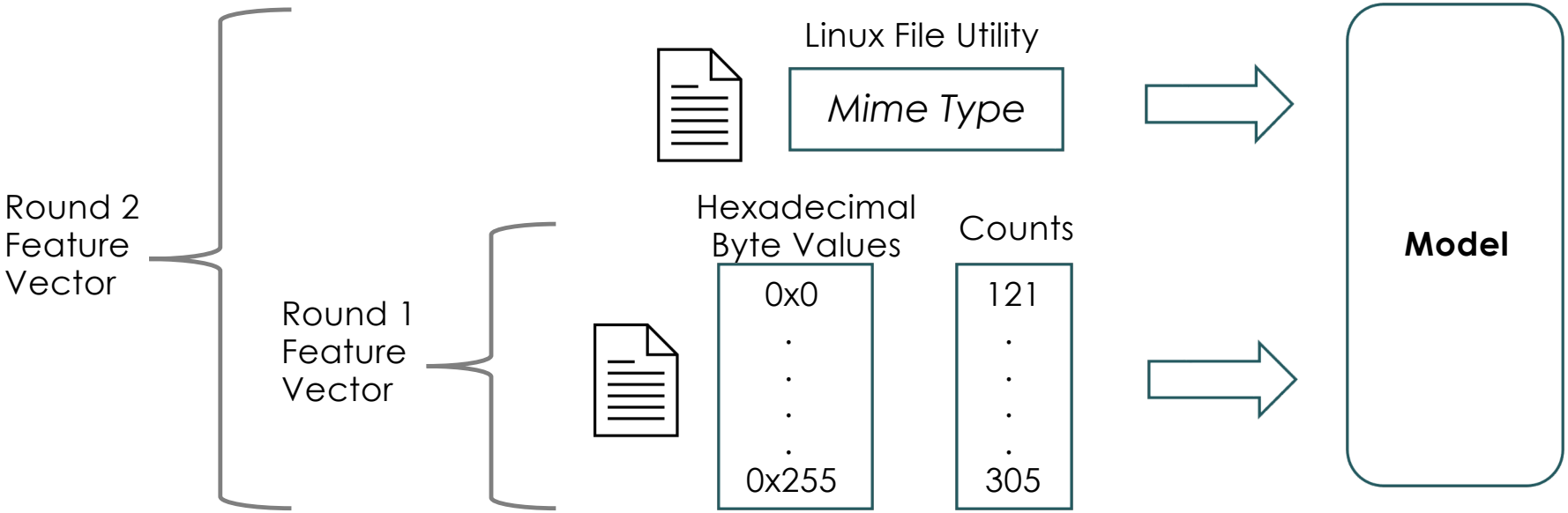


Breakdown of *file* utility's performance



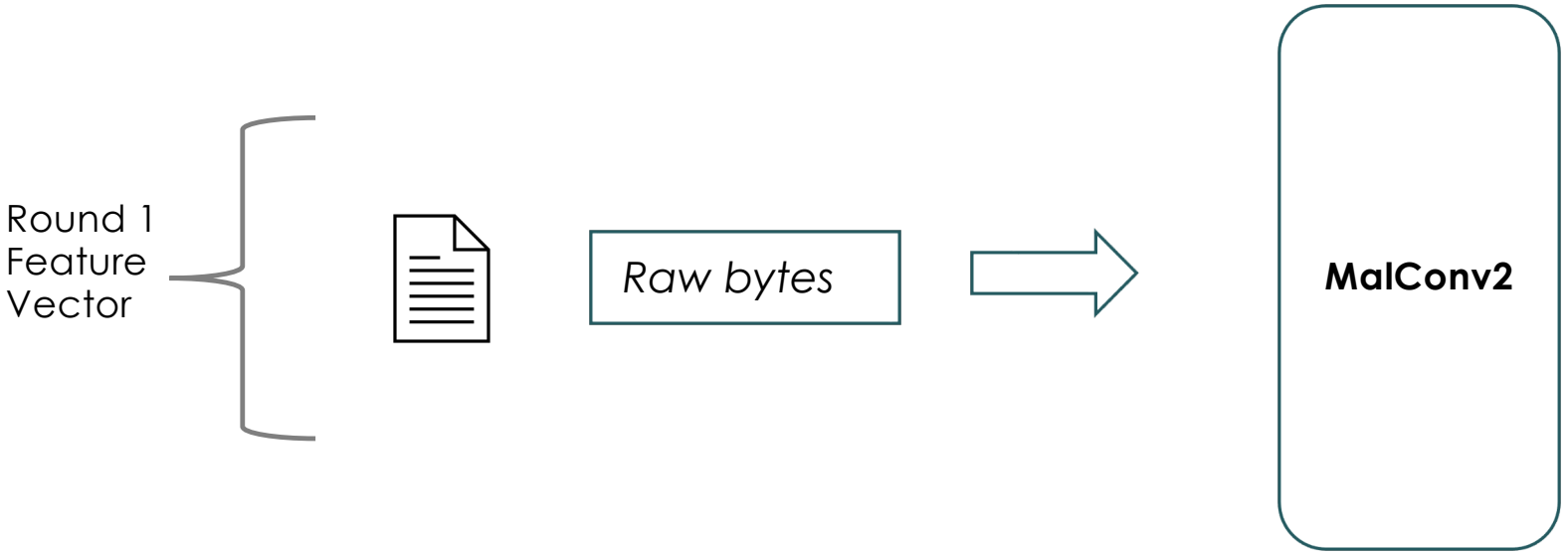
OAK RIDGE National Laboratory

Machine Learning – Input Features



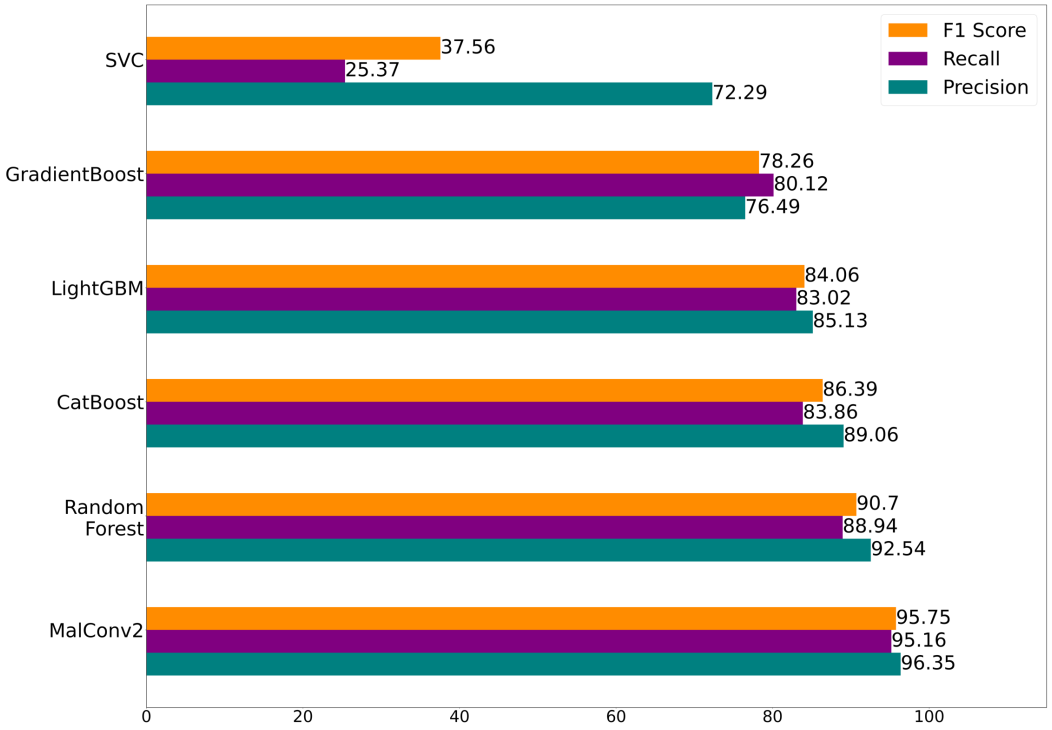
OAK RIDGE National Laboratory

Deep Learning – Input Features



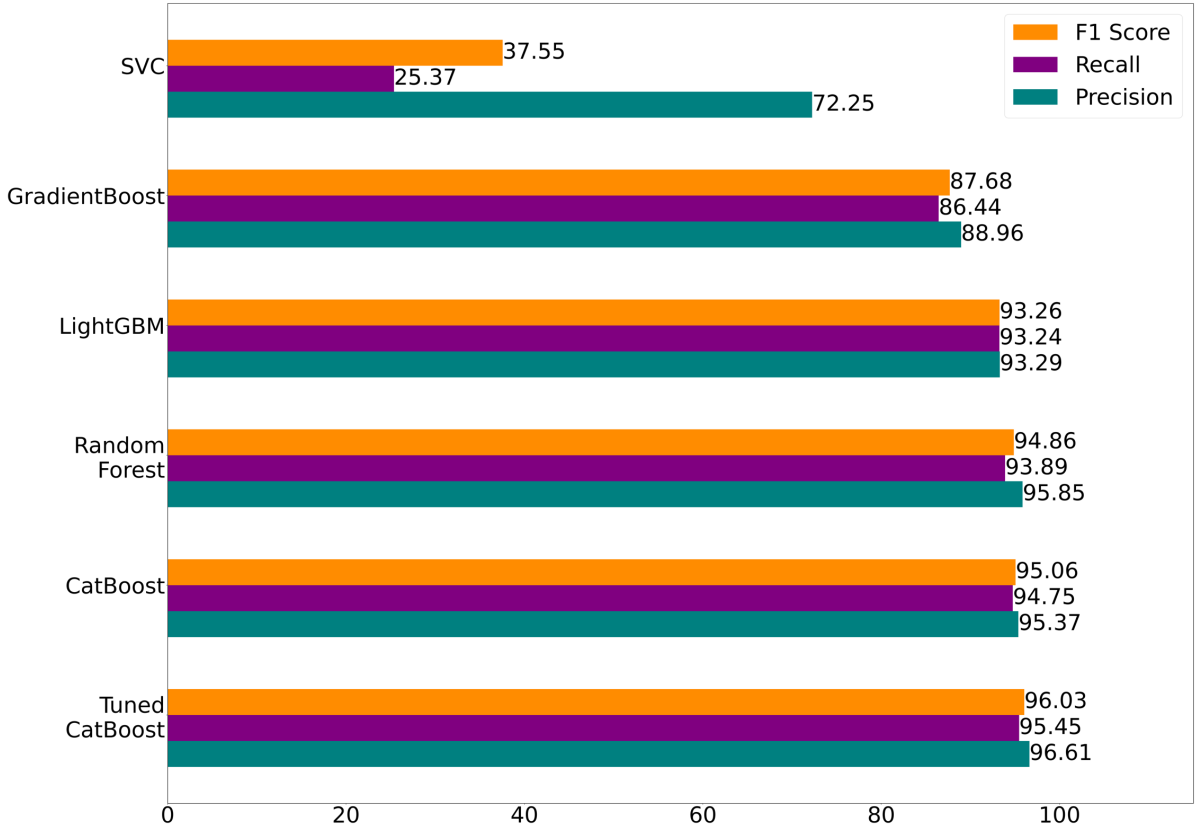
ML/DL Performance: Round 1

Results of several ML models and one deep learning model



- MalConv2 trains on the raw bytes of the file
- The ML models, on the other hand, train on a byte occurrence vector
- This 255-character vector summarizes the occurrence of all possible hexadecimal values per file

ML Performance: Round 2



- New feature vector for ML models consists of byte occurrence vector concatenated with mime-type output from *file* utility
- MalConv2 did not improve from additional feature
- Tuned CatBoost performed best across all metrics

OAK RIDGE National Laboratory

Conclusion and Next Steps

- Polyglot files present a serious challenge for existing utilities
- ML/DL show promise for detection
- Need to perform tuning and feature selection to maximize performance
- Need to demonstrate high throughput for practical use in industry

Questions?

- Contact: kochlr@ornl.gov or oeschts@ornl.gov

Supplemental: Detection of Novel Polyglots

- Tested our models against small dataset of novel polyglot malware
  - Cannot release data due to NDA
- Random Forest: 75% accuracy
- MalConv2: 83.80% accuracy
- CatBoost: 99.4% accuracy

Supplemental: Deep Learning Model Selection

- We chose Malconv2
  - Binary classifier designed for malware detection
  - 1-D convolutional neural network
- Reasons for selection
  - File type agnostic
  - Feature vector = raw bytes
  - Larger input capacity (16MB) than competing models