

[Skip to main content](#)

> [cs](#) > arXiv:2203.07561v4

## quick links

- [Login](#)
- [Help Pages](#)
- [About](#)

Computer Science > Cryptography and Security

**arXiv:2203.07561v4** (cs)

*[Submitted on 14 Mar 2022 ([v1](#)), last revised 13 Apr 2022 (this version, v4)]*

# Toward the Detection of Polyglot Files

[Luke Koch](#), [Sean Oesch](#), [Mary Adkisson](#), [Sam Erwin](#), [Brian Weber](#), [Amul Chaulagain](#)

[Download PDF](#)

Standardized file formats play a key role in the development and use of computer software. However, it is possible to abuse standardized file formats by creating a file that is valid in multiple file formats. The resulting polyglot (many languages) file can confound file format identification, allowing elements of the file to evade analysis. This is especially problematic for malware detection systems that rely on file format identification for feature extraction. File format identification processes that depend on file signatures can be easily evaded thanks to flexibility in the format specifications of certain file formats. Although work has been done to identify file formats using more comprehensive methods than file signatures, accurate identification of polyglot files remains an open problem. Since malware detection systems routinely perform file format-specific feature extraction, polyglot files need to be filtered out prior to ingestion by these systems. Otherwise, malicious content could pass through undetected. To address the problem of polyglot detection we assembled a data set using the mitra tool. We then evaluated the performance of the most commonly used file identification tool, file. Finally, we demonstrated the accuracy, precision, recall and


F1 score of a range of machine and deep learning models. Malconv2 and Catboost demonstrated the highest recall on our data set with 95.16% and 95.45%, respectively. These models can be incorporated into a malware detector's file processing pipeline to filter out potentially malicious polyglots before file format-dependent feature extraction takes place.

Subjects: **Cryptography and Security (cs.CR)**; Machine Learning (cs.LG)

Cite as: [arXiv:2203.07561](https://arxiv.org/abs/2203.07561) [cs.CR]

(or [arXiv:2203.07561v4](https://arxiv.org/abs/2203.07561v4) [cs.CR] for this version)

<https://doi.org/10.48550/arXiv.2203.07561>

 Focus to learn more

arXiv-issued DOI via DataCite

## Submission history

From: Luke Koch [[view email](#)]

[\[v1\]](#) Mon, 14 Mar 2022 23:48:22 UTC (88 KB)

[\[v2\]](#) Wed, 16 Mar 2022 19:29:39 UTC (88 KB)

[\[v3\]](#) Wed, 23 Mar 2022 20:52:30 UTC (88 KB)

[\[v4\]](#) Wed, 13 Apr 2022 02:54:22 UTC (2,377 KB)

☒ Bibliographic Tools

## Bibliographic and Citation Tools

☐ Bibliographic Explorer Toggle

Bibliographic Explorer ([What is the Explorer?](#))

☐ Litmaps Toggle

Litmaps ([What is Litmaps?](#))

☐ scite.ai Toggle

scite Smart Citations ([What are Smart Citations?](#))

☐ Code & Data

## Code and Data Associated with this Article

☐ arXiv Links to Code Toggle

arXiv Links to Code & Data ([What is Links to Code & Data?](#))

☐

## Demos

☐ Replicate Toggle

Replicate ([What is Replicate?](#))

☐ Related Papers

# Recommenders and Search Tools

☐ Connected Papers Toggle

Connected Papers ([What is Connected Papers?](#))

☐ Core recommender toggle

CORE Recommender ([What is CORE?](#))

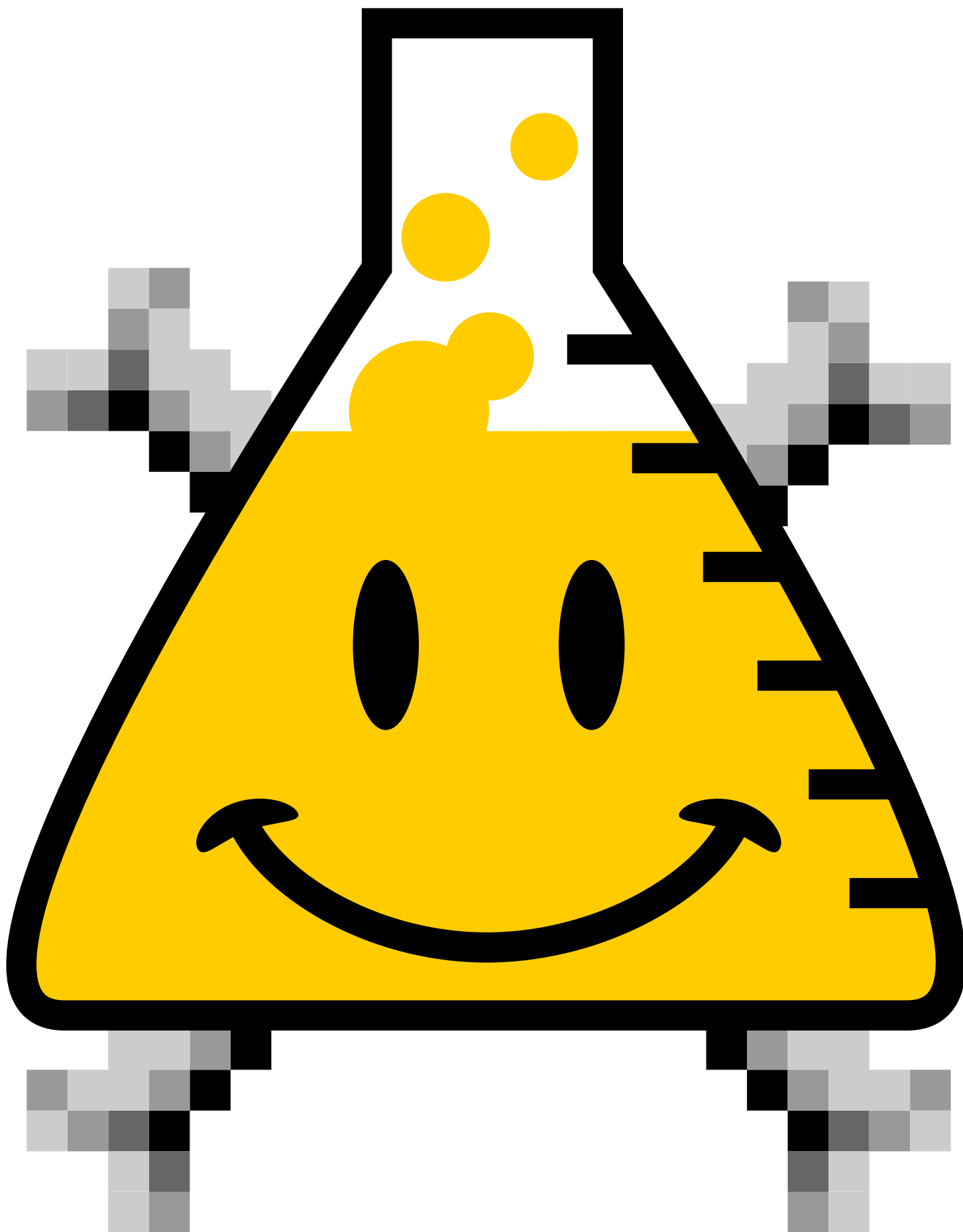
☐ About arXivLabs

## arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [Learn more about arXivLabs](#) and [how to get involved](#).



[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))

