Team 2 Defense

By Nhat Nguyen, Sonjoy Paul, Erik Muller

https://github.com/gnat-n/-ml-based-malware-defender-and-attack

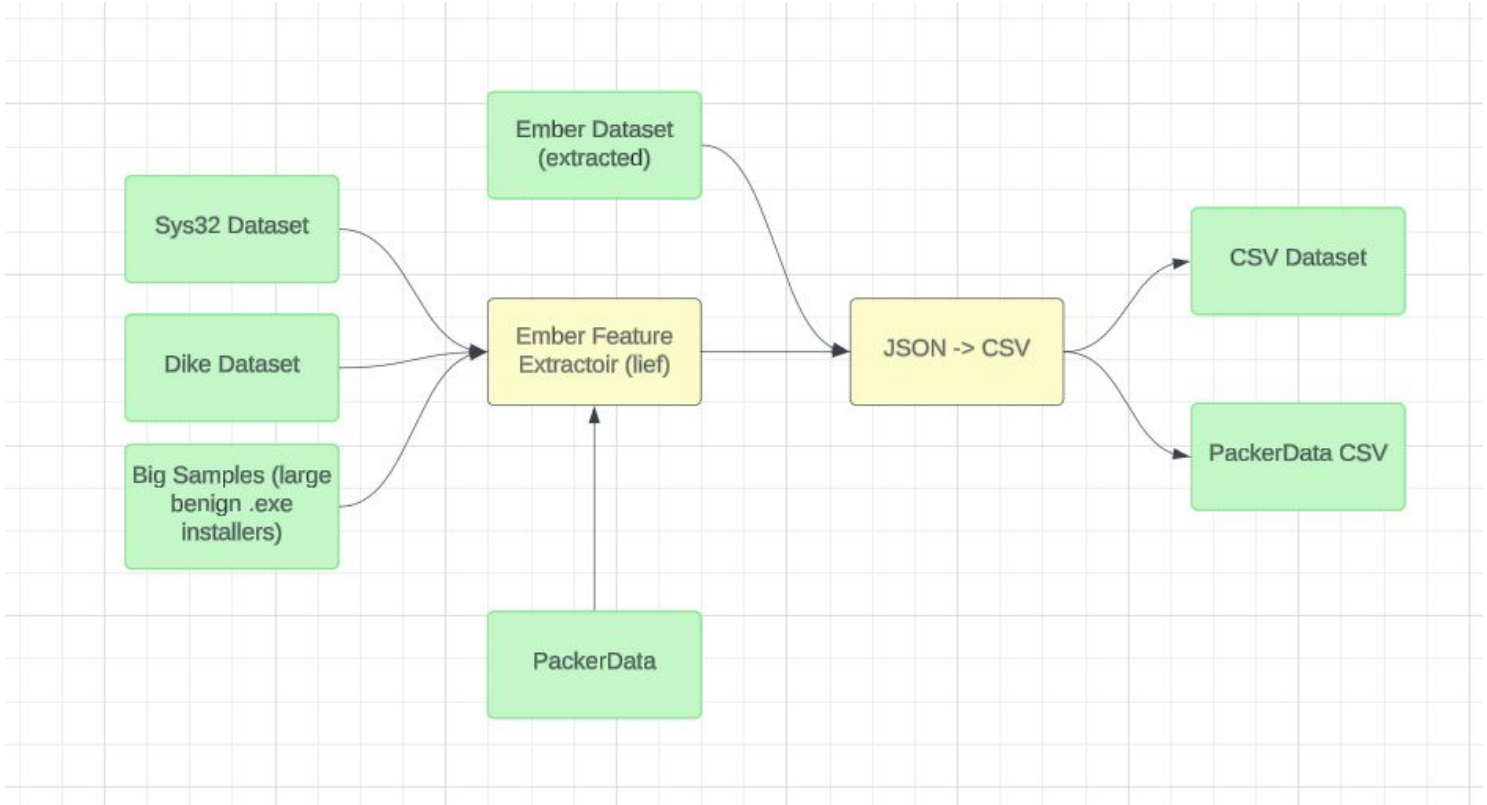
Defense Models

- Model 1 - Random Forest Model (100 estimators) trained on ember dataset, our system 32 folder, the dike dataset, and random installers
  - 816493 samples, each with 2381 features
- Model 2 - Random Forest Model (100 estimators) trained on packing data dataset
  - 3041 samples, each with 2381 features
- Feature Extraction:
  - All numeric features are used
  - Non-numeric (imports/exports, libraries used, etc) features are converted to numeric values using feature hashing
  - Extraction converted to csv to unify all dataset formats

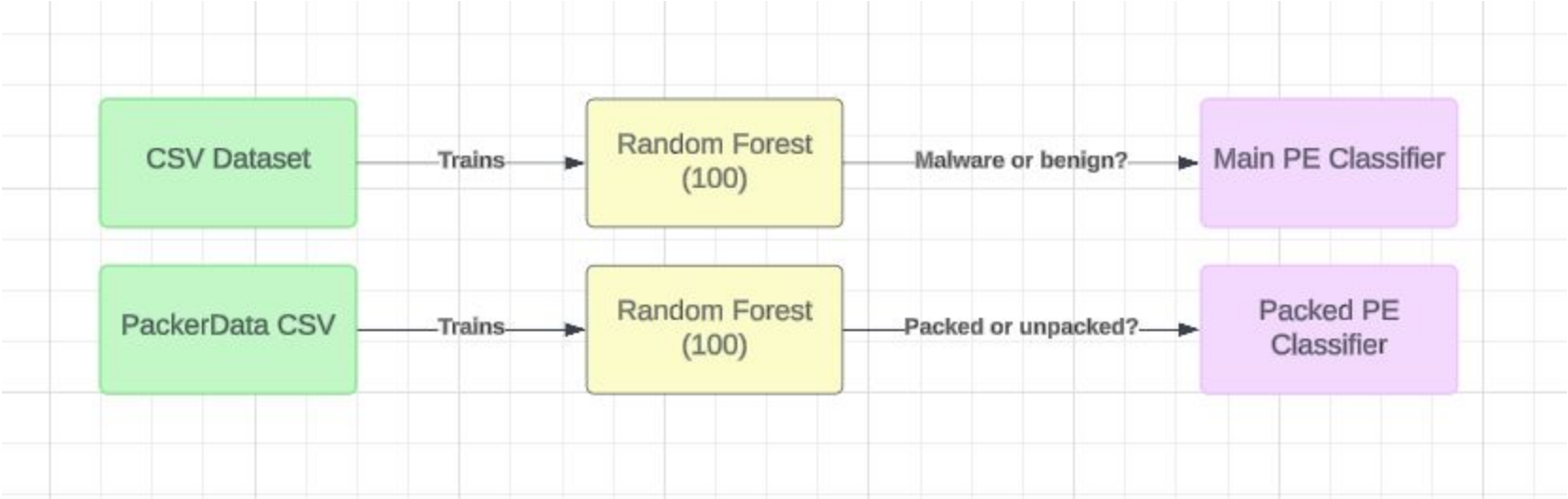
Feature Extractor (Ember Dataset)


```
{
    "sha256": "0abb4fda7d5b13801d63bee53e5e256be43e141faa077a6d149874242c3f02c2",
    "md5": "63956d6417f8f43357d9a8e79e52257e",
    "appeared": "2006-12",
    "label": 0,
    "avclass": "",
    "histogram": [...
    ],
    "byteentropy": [...
    ],
    "strings": {...
    },
    "general": {...
    },
    "header": {...
    },
    "section": {...
    },
    "imports": {...
    },
    "exports": [],
    "datadirectories": [...
    ]
}
```

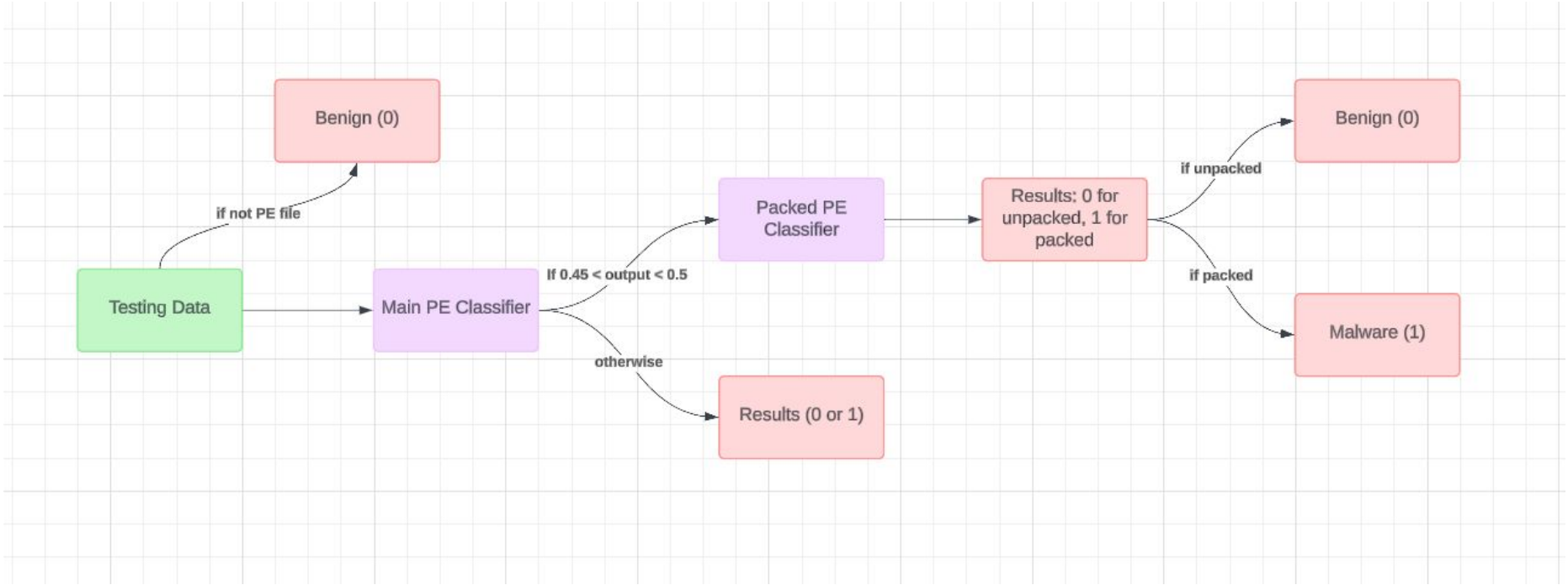
Model Architecture (Data Acquisition)



Model Architecture (Training)



Model Architecture (Deployment)



Other Attempts

- Black box Attempt
  - Random Forest on ember numerical data (640 features)
    - Lower Accuracy
    - Threshold did not translate from dataset to test set
- Other Attempts
  - SVMs
    - Low Accuracy
  - KNN
    - Slow Prediction Speed
    - RAM Intensive
    - Low Accuracy

Issues we faced

- Data
  - Ember dataset format needed to be converted from json into csv
    - Slow, lead to large csv files
  - Handling of non-numerical data
    - One-hot encoding too sparse, used feature hashing
  - Handling of large data
    - 40+ gb of csvs of data used to train model
    - Slow train times on models
    - Only able to be trained on machines with high RAM
- Blackbox Threshold
  - Could not find correct threshold
    - Led to very poor initial black box model
      - Too high, then too low
- Handling Dangerous files
  - PE files had to be handled either in VM or Mac computer
    - Limited options