Automatic Program Protection Using Fuzzing Driven Classifiers

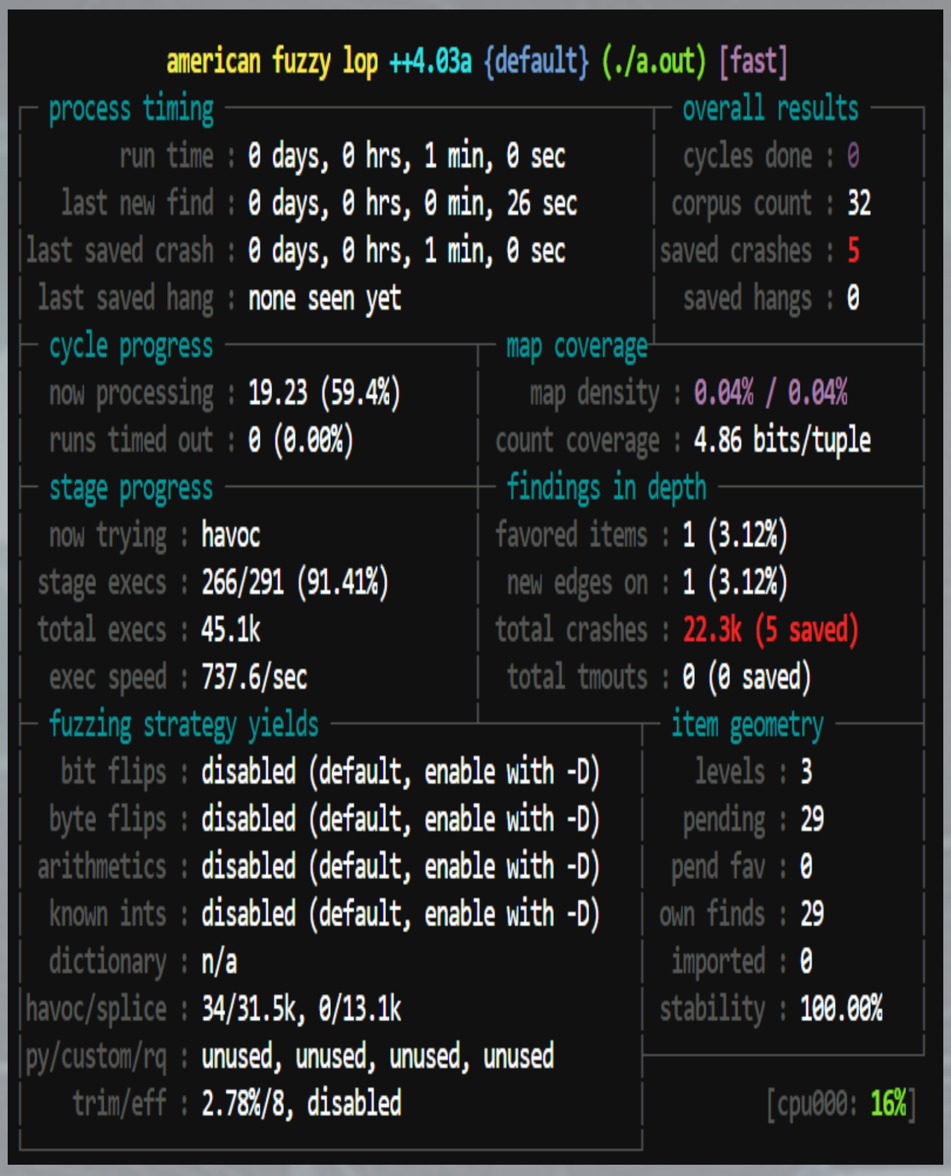
Introduction

Goal

- Classify program inputs to avoid crashing states that may lead to vulnerabilities
- Using fuzzer driven ML models to create automatic program protection filters
- Leveraging novel techniques in generating high-quality data for attention-based architectures

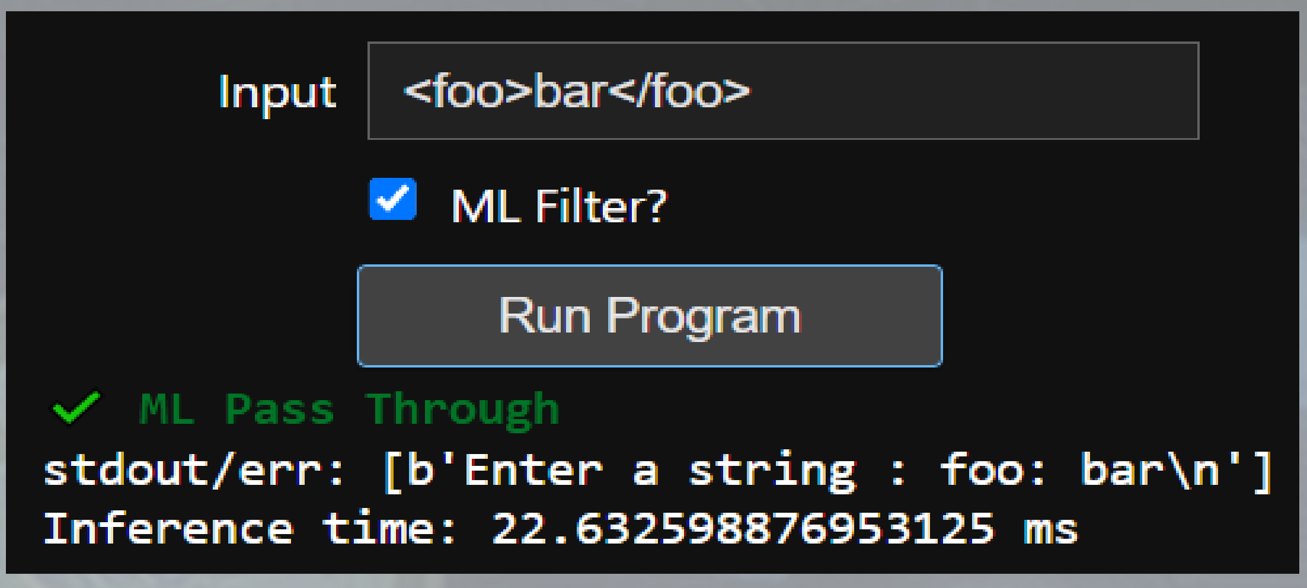
Address Limitations in Past Work

- Generates large (500k+) training datasets
- Usage for protection and not fuzzing guidance
- Models capable of extracting syntactic features
- Analysis of corpus coverage and layer contributions



```
            american fuzzy lop ++4.03a {default} (./a.out) [fast]
┌─ process timing ────────────────────────────────────┬─ overall results ──┐
│        run time : 0 days, 0 hrs, 1 min, 0 sec        │  cycles done : 0   │
│   last new find : 0 days, 0 hrs, 0 min, 26 sec       │ corpus count : 32  │
│last saved crash : 0 days, 0 hrs, 1 min, 0 sec        │saved crashes : 5   │
│ last saved hang : none seen yet                      │  saved hangs : 0   │
├─ cycle progress ─────────────────┬─ map coverage─────┴────────────────────┤
│  now processing : 19.23 (59.4%)  │    map density : 0.04% / 0.04%         │
│  runs timed out : 0 (0.00%)      │ count coverage : 4.86 bits/tuple       │
├─ stage progress ─────────────────┼─ findings in depth ────────────────────┤
│  now trying : havoc              │ favored items : 1 (3.12%)              │
│ stage execs : 266/291 (91.41%)   │  new edges on : 1 (3.12%)              │
│ total execs : 45.1k              │ total crashes : 22.3k (5 saved)        │
│  exec speed : 737.6/sec          │  total tmouts : 0 (0 saved)            │
├─ fuzzing strategy yields ────────┴──────────────────┬─ item geometry ─────┤
│   bit flips : disabled (default, enable with -D)    │    levels : 3       │
│  byte flips : disabled (default, enable with -D)    │   pending : 29      │
│ arithmetics : disabled (default, enable with -D)    │  pend fav : 0       │
│  known ints : disabled (default, enable with -D)    │ own finds : 29      │
│  dictionary : n/a                                   │  imported : 0       │
│havoc/splice : 34/31.5k, 0/13.1k                     │ stability : 100.00% │
│py/custom/rq : unused, unused, unused, unused        ├─────────────────────┘
│    trim/eff : 2.78%/8, disabled                     │          [cpu000: 16%]
└─────────────────────────────────────────────────────┘
```

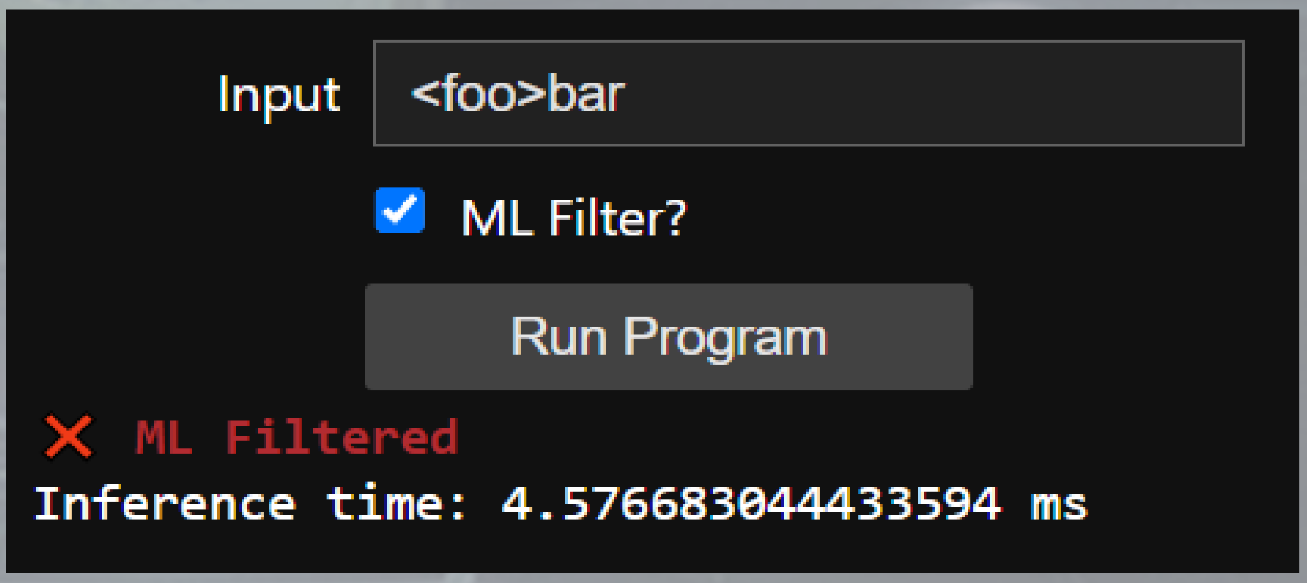
Automatic Supervised Learning Corpus



Input <foo>bar</foo>

☑ ML Filter?

```
✓ ML Pass Through
stdout/err: [b'Enter a string : foo: bar\n']
Inference time: 22.632598876953125 ms
```



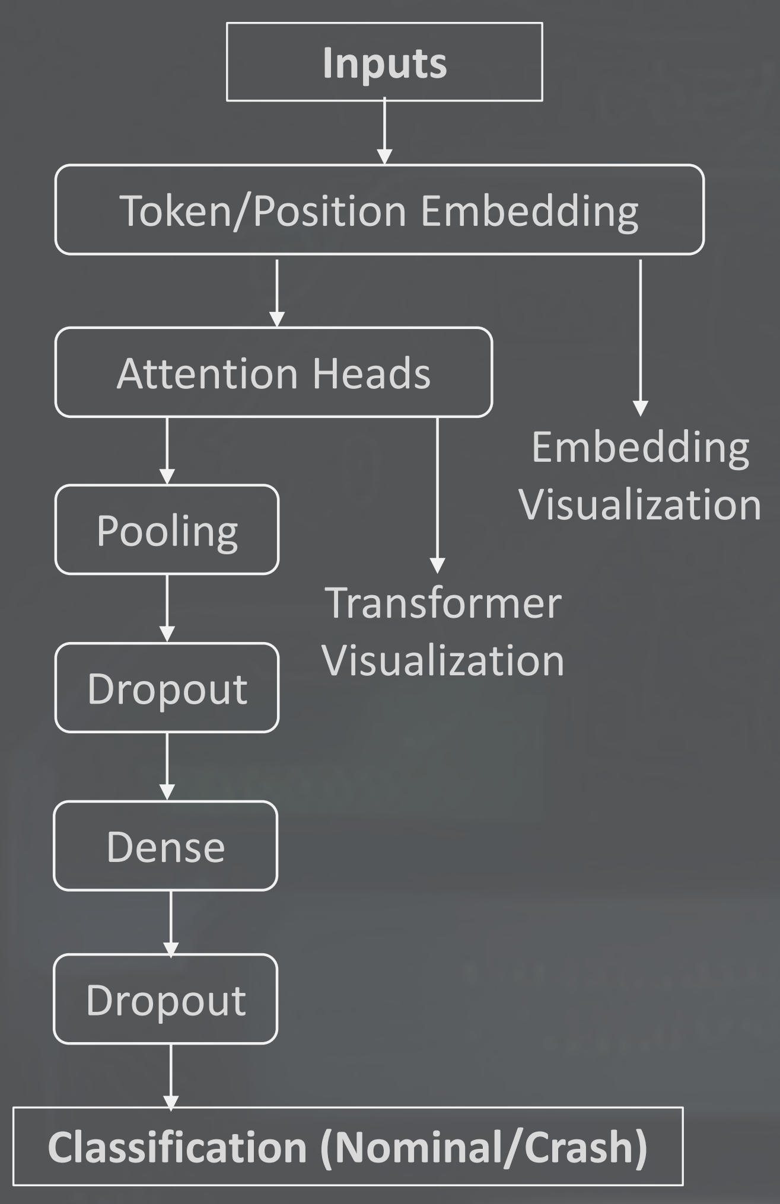
Input <foo>bar

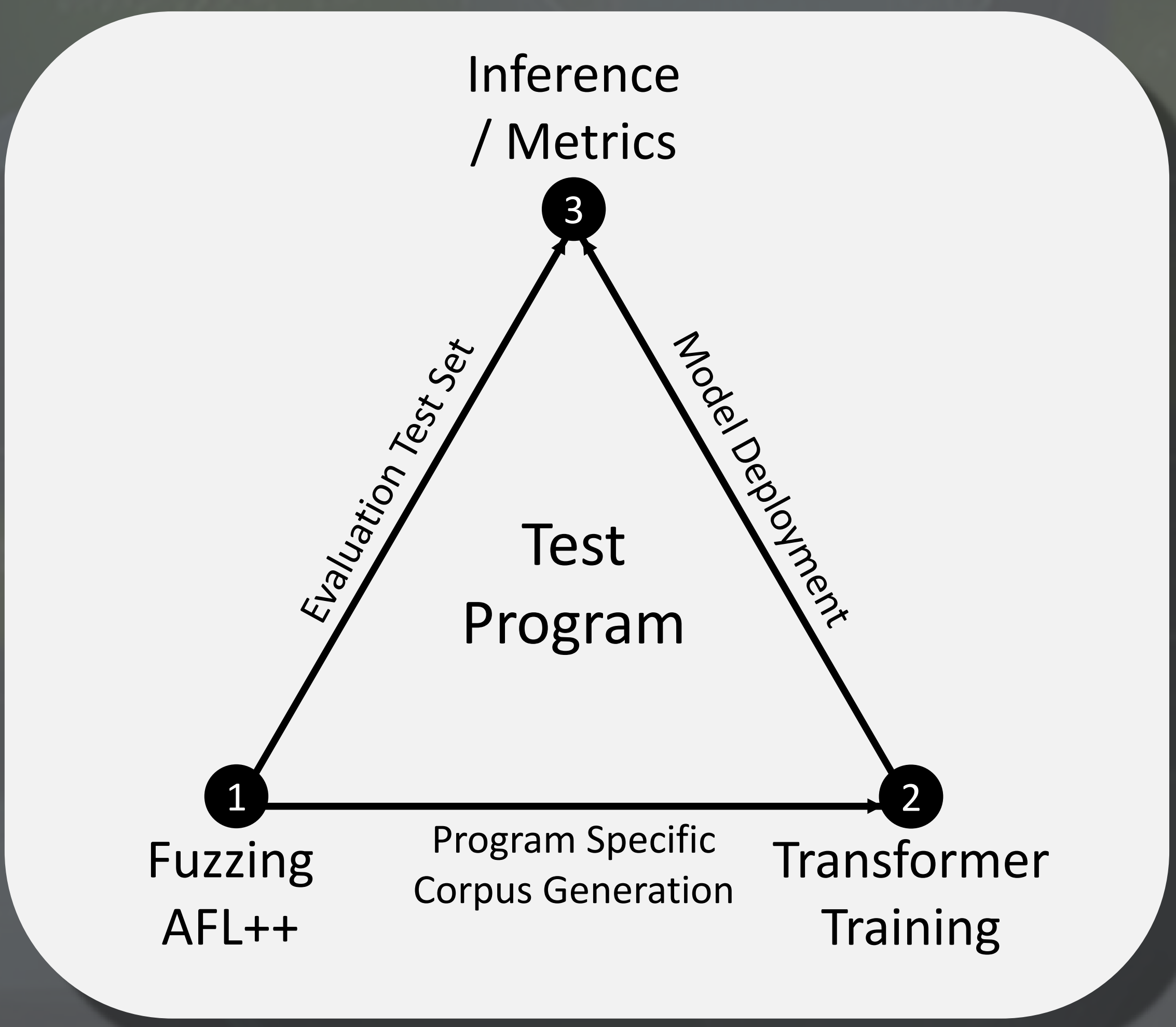
☑ ML Filter?

```
✗ ML Filtered
Inference time: 4.576683044433594 ms
```

Provide Tailored Program Protection via Classifier

Transformer Architecture





Experiment

Approach

1. Modified fuzzer to generate training data for supervised learning
2. Binary classifier model trained with attention heads to learn nominal vs. crashing
3. Inference for collecting metrics and providing tailored application protection

Metrics

- Ablation study: Statistics collected for reduced models: dense only, positional, transformer
- F1/AUC Scores: Measure performance of classifier with false positives/negatives
- PaCMAP: Qualitative analysis of manifold coverage for both datasets and activations

Findings

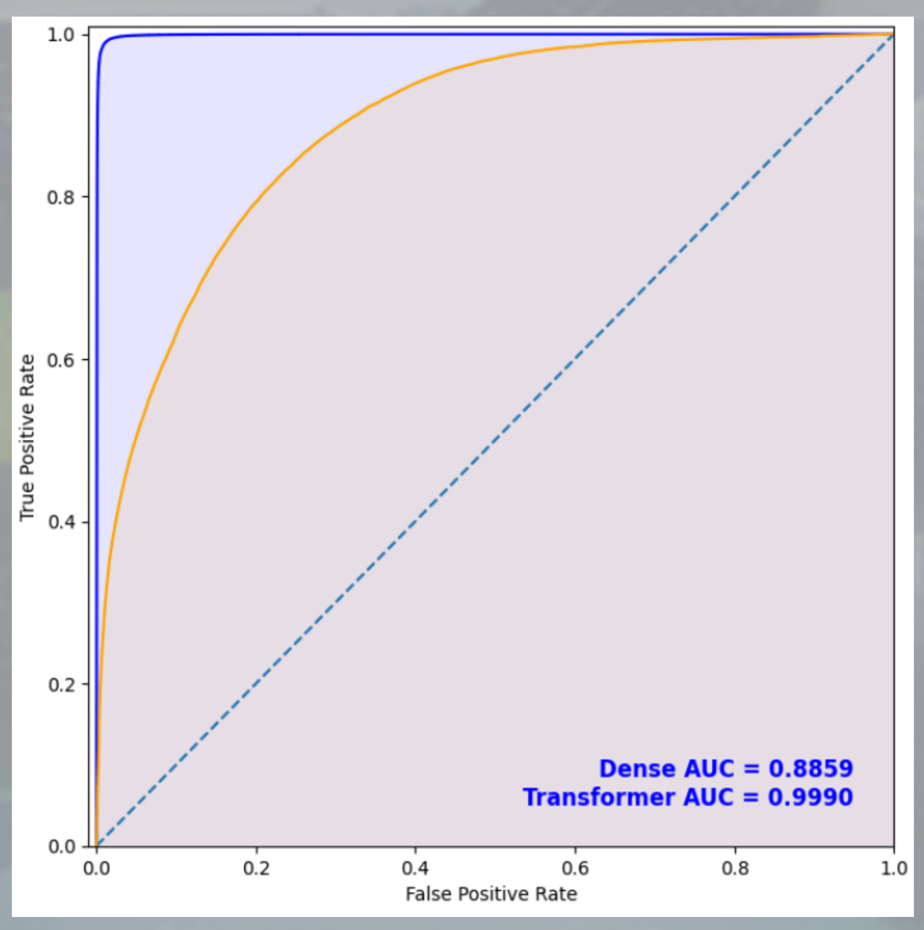
	Dense		Position		Attention	
	F1	AUC	F1	AUC	F1	AUC
Test	0.8044	0.8666	0.9695	0.9880	0.9911	0.9993
Fuzzgoat	0.9527	0.9851	0.9355	0.9746	0.9906	0.9994
XML CVE	0.9744	0.9747	0.9860	0.9964	0.9992	1.0000

Quantitative

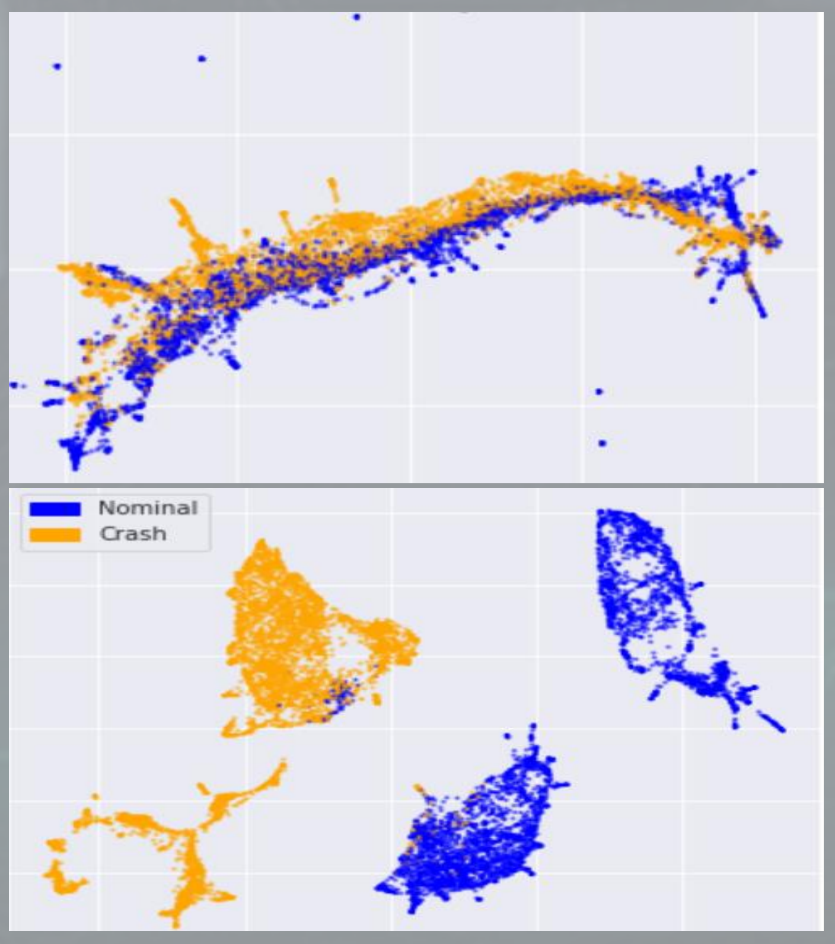
- Transformer model more performant in all cases
- Sequence information needed for inputs with heavy overlap, as seen in Test Program
- Able to correctly identify CVE inducing inputs
- Dense model achieves similar performance when special characters are present

Qualitative

- Good solution-space coverage from fuzzer, acts as a form of dataset augmentation
- Extra structure added by transformer to create distinct groupings on learned manifold



Test Classifier Scores



Learned Dataset Structure