# The Third International Verification of Neural Networks Competition (VNN-COMP 2022): Summary and Results

Stanley Bak[*], Changliu Liu[†], Taylor Johnson[‡]

**Abstract**

This report summarizes the third International Verification of Neural Networks Competition (VNN-COMP 2021), held as a part of the 5th Workshop on Formal Methods for ML-Enabled Autonomous Systems (FOMLAS) that was collocated with the 34th International Conference on Computer-Aided Verification (CAV). The goal of the competition is to provide an objective comparison of the state-of-the-art methods in neural network verification, in terms of scalability and speed. Along this line, we used standard formats (ONNX for neural networks and VNNLIB for specifications), standard hardware (all tools are run by the organizers on AWS), and tool parameters provided by the tool authors. This report summarizes the rules, benchmarks, participating tools, results, and lessons learned from this competition.

## 1 Scored

Table 1: Benchmark `carvana-unet-2022`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 39 | 0 | 39 | 468 | 100.0% |
| 2 | MN BaB | 19 | 0 | 0 | 209 | 44.7% |
| 3 | Verinet | 3 | 0 | 0 | 30 | 6.4% |

Table 2: Benchmark `cifar100-tinyimagenet-resnet`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 69 | 0 | 56 | 813 | 100.0% |
| 2 | Cgdtest | 95 | 0 | 28 | 725 | 89.2% |
| 3 | MN BaB | 60 | 3 | 10 | 674 | 82.9% |
| 4 | Verinet | 48 | 3 | 6 | 541 | 66.5% |

Table 3: Benchmark `cifar-biasfield`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | Cgdtest | 71 | 0 | 55 | 732 | 100.0% |
| 2 | Verinet | 69 | 1 | 0 | 721 | 98.5% |
| 3 | Verapak | 71 | 0 | 0 | 635 | 86.7% |
| 4 | $\alpha$,$\beta$-CROWN | 69 | 0 | 0 | 623 | 85.1% |
| 5 | MN BaB | 36 | 1 | 17 | 404 | 55.2% |
| 6 | Marabou | 27 | 0 | 0 | 270 | 36.9% |
| 7 | Nnenum | 4 | 0 | 0 | 43 | 5.9% |

Table 4: Benchmark `collins-rul-cnn`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | Nnenum | 16 | 45 | 57 | 725 | 100.0% |
| 2 | MN BaB | 16 | 44 | 57 | 715 | 98.6% |
| 3 | Verinet | 16 | 43 | 0 | 590 | 81.4% |
| 4 | Peregrinn | 14 | 42 | 0 | 560 | 77.2% |
| 5 | $\alpha$,$\beta$-CROWN | 15 | 42 | 54 | 278 | 38.3% |
| 6 | Cgdtest | 1 | 42 | 43 | -984 | 0% |

# 2 Unscored

# 3 Stats

*S. Bak is with Stony Brook University, `stanley.bak@stonybrook.edu`.
†C. Liu is with Carnegie Mellon University, `cliu6@andrew.cmu.edu`.
‡T. Johnson is with Vanderbilt University, `taylor.johnson@vanderbilt.edu`.

Table 5: Benchmark `mnist-fc`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | Verinet | 53 | 18 | 50 | 817 | 100.0% |
| 2 | MN BaB | 53 | 18 | 47 | 804 | 98.4% |
| 3 | Debona | 48 | 18 | 38 | 737 | 90.2% |
| 4 | Nnenum | 48 | 11 | 29 | 648 | 79.3% |
| 5 | Marabou | 44 | 16 | 0 | 600 | 73.4% |
| 6 | Peregrinn | 27 | 11 | 7 | 394 | 48.2% |
| 7 | Cgdtest | 66 | 3 | 23 | 241 | 29.5% |
| 8 | Verapak | 40 | 2 | 42 | 104 | 12.7% |
| 9 | $\alpha$,$\beta$-CROWN | 66 | 10 | 47 | 71 | 8.7% |

Table 6: Benchmark `nn4sys`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 152 | 0 | 132 | 1791 | 100.0% |
| 2 | Verinet | 57 | 0 | 43 | 670 | 37.4% |
| 3 | MN BaB | 42 | 0 | 19 | 467 | 26.1% |
| 4 | Peregrinn | 24 | 0 | 22 | 284 | 15.9% |
| 5 | Nnenum | 23 | 0 | 8 | 246 | 13.7% |
| 6 | Debona | 2 | 0 | 2 | 24 | 1.3% |
| 7 | Cgdtest | 2 | 0 | 2 | -176 | 0% |

Table 7: Benchmark `oval21`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | MN BaB | 19 | 1 | 2 | 205 | 100.0% |
| 2 | Verinet | 17 | 1 | 1 | 189 | 92.2% |
| 3 | $\alpha$,$\beta$-CROWN | 25 | 0 | 10 | 180 | 87.8% |
| 4 | Marabou | 19 | 0 | 17 | 125 | 61.0% |
| 5 | Nnenum | 3 | 1 | 0 | 40 | 19.5% |
| 6 | Peregrinn | 1 | 0 | 0 | 10 | 4.9% |
| 7 | Cgdtest | 11 | 0 | 1 | -580 | 0% |

Table 8: Benchmark `reach-prob-density`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | Nnenum | 22 | 14 | 21 | 410 | 100.0% |
| 2 | $\alpha$,$\beta$-CROWN | 22 | 14 | 23 | 406 | 99.0% |
| 3 | Verinet | 22 | 14 | 10 | 383 | 93.4% |
| 4 | MN BaB | 22 | 12 | 14 | 368 | 89.8% |
| 5 | Marabou | 17 | 14 | 12 | 334 | 81.5% |
| 6 | Peregrinn | 18 | 14 | 2 | 324 | 79.0% |
| 7 | Cgdtest | 0 | 5 | 5 | 60 | 14.6% |
| 8 | Debona | 0 | 2 | 2 | 24 | 5.9% |

Table 9: Benchmark `rl-benchmarks`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 193 | 103 | 296 | 3552 | 100.0% |
| 2 | Verinet | 193 | 103 | 292 | 3547 | 99.9% |
| 3 | MN BaB | 193 | 103 | 288 | 3536 | 99.5% |
| 4 | Nnenum | 191 | 103 | 282 | 3504 | 98.6% |
| 5 | Peregrinn | 193 | 103 | 271 | 3502 | 98.6% |
| 6 | Marabou | 191 | 103 | 278 | 3496 | 98.4% |
| 7 | Debona | 153 | 99 | 240 | 3000 | 84.5% |
| 8 | Averinn | 92 | 8 | 16 | 1032 | 29.1% |
| 9 | Cgdtest | 10 | 24 | 29 | 98 | 2.8% |
| 10 | Verapak | 0 | 4 | 0 | 40 | 1.1% |

Table 10: Benchmark `sri-resnet-a`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 20 | 12 | 7 | 356 | 100.0% |
| 2 | Cgdtest | 26 | 6 | 14 | 352 | 98.9% |
| 3 | MN BaB | 18 | 12 | 19 | 343 | 96.3% |
| 4 | Verinet | 12 | 12 | 4 | 248 | 69.7% |
| 5 | Marabou | 3 | 0 | 0 | 30 | 8.4% |

Table 11: Benchmark `sri-resnet-b`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | MN BaB | 27 | 11 | 24 | 431 | 100.0% |
| 2 | Verinet | 20 | 11 | 4 | 319 | 74.0% |
| 3 | Marabou | 61 | 0 | 36 | -411 | 0% |
| 4 | Cgdtest | 22 | 6 | 0 | -111 | 0% |
| 5 | $\alpha$,$\beta$-CROWN | 28 | 0 | 8 | -789 | 0% |

Table 12: Benchmark `tllverifybench`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | Fastbatllnn | 11 | 21 | 32 | 384 | 100.0% |
| 2 | MN BaB | 11 | 21 | 21 | 364 | 94.8% |
| 3 | $\alpha$,$\beta$-CROWN | 11 | 21 | 11 | 351 | 91.4% |
| 4 | Peregrinn | 10 | 21 | 7 | 324 | 84.4% |
| 5 | Verinet | 11 | 21 | 0 | 320 | 83.3% |
| 6 | Nnenum | 1 | 21 | 10 | 240 | 62.5% |
| 7 | Debona | 0 | 19 | 10 | 210 | 54.7% |
| 8 | Marabou | 4 | 15 | 2 | 194 | 50.5% |
| 9 | Cgdtest | 0 | 9 | 6 | 2 | 0.5% |

Table 13: Benchmark `vggnet16-2022`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 14 | 1 | 11 | 176 | 100.0% |
| 2 | Nnenum | 11 | 1 | 0 | 127 | 72.2% |
| 3 | MN BaB | 5 | 1 | 4 | 69 | 39.2% |
| 4 | Verinet | 5 | 1 | 0 | 60 | 34.1% |
| 5 | Cgdtest | 0 | 2 | 1 | -378 | 0% |

Table 14: Overall Score

| # | Tool | Score |
|---|---|---|
| 1 | MN BaB | 1025.5 |
| 2 | $\alpha$,$\beta$-CROWN | 1010.4 |
| 3 | Verinet | 936.8 |
| 4 | Nnenum | 551.7 |
| 5 | Marabou | 410.1 |
| 6 | Peregrinn | 408.2 |
| 7 | Cgdtest | 335.5 |
| 8 | Debona | 236.5 |
| 9 | Verapak | 100.6 |
| 10 | Fastbatllnn | 100.0 |
| 11 | Averinn | 29.1 |

Table 15: Benchmark `acasxu`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | Nnenum | 139 | 47 | 174 | 2218 | 100.0% |
| 2 | $\alpha$,$\beta$-CROWN | 139 | 43 | 56 | 1685 | 76.0% |
| 3 | Cgdtest | 85 | 30 | 115 | 680 | 30.7% |

Table 16: Benchmark `cifar2020`

| # | Tool | Verified | Falsified | Fastest | Score | Percent |
|---|---|---|---|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 148 | 38 | 176 | 1822 | 100.0% |
| 2 | MN BaB | 93 | 28 | 14 | 1313 | 72.1% |
| 3 | Nnenum | 66 | 19 | 0 | 850 | 46.7% |
| 4 | Cgdtest | 86 | 30 | 6 | 499 | 27.4% |
| 5 | Verapak | 0 | 15 | 0 | 150 | 8.2% |
| 6 | Verinet | 91 | 47 | 7 | -3868 | 0% |
| 7 | Marabou | 4 | 0 | 0 | -60 | 0% |

Table 17: Overhead

| # | Tool | Seconds |
|---|---|---|
| 1 | Marabou | 0.2 |
| 2 | Fastbatllnn | 0.5 |
| 3 | Nnenum | 0.8 |
| 4 | Cgdtest | 1.3 |
| 5 | Peregrinn | 1.3 |
| 6 | Debona | 2.0 |
| 7 | Averinn | 3.1 |
| 8 | Verinet | 3.4 |
| 9 | Verapak | 4.6 |
| 10 | $\alpha$,$\beta$-CROWN | 6.7 |
| 11 | MN BaB | 8.2 |

Table 18: Num Benchmarks Participated

| # | Tool | Count |
|---|---|---|
| 1 | Verinet | 13 |
| 2 | MN BaB | 13 |
| 3 | $\alpha$,$\beta$-CROWN | 13 |
| 4 | Cgdtest | 12 |
| 5 | Nnenum | 9 |
| 6 | Marabou | 8 |
| 7 | Peregrinn | 7 |
| 8 | Debona | 5 |
| 9 | Verapak | 3 |
| 10 | Fastbatllnn | 1 |
| 11 | Averinn | 1 |

Table 19: Num Instances Verified

| # | **Tool** | **Count** |
|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 926 |
| 2 | Verinet | 754 |
| 3 | MN BaB | 748 |
| 4 | Nnenum | 515 |
| 5 | Marabou | 514 |
| 6 | Peregrinn | 478 |
| 7 | Cgdtest | 401 |
| 8 | Debona | 341 |
| 9 | Verapak | 117 |
| 10 | Averinn | 100 |
| 11 | Fastbatllnn | 32 |

Table 20: Num SAT

| # | **Tool** | **Count** |
|---|---|---|
| 1 | Verinet | 228 |
| 2 | MN BaB | 227 |
| 3 | $\alpha$,$\beta$-CROWN | 203 |
| 4 | Nnenum | 196 |
| 5 | Peregrinn | 191 |
| 6 | Marabou | 148 |
| 7 | Debona | 138 |
| 8 | Cgdtest | 97 |
| 9 | Fastbatllnn | 21 |
| 10 | Averinn | 8 |
| 11 | Verapak | 6 |

Table 21: Num UNSAT

| # | **Tool** | **Count** |
|---|---|---|
| 1 | $\alpha$,$\beta$-CROWN | 723 |
| 2 | Verinet | 526 |
| 3 | MN BaB | 521 |
| 4 | Marabou | 366 |
| 5 | Nnenum | 319 |
| 6 | Cgdtest | 304 |
| 7 | Peregrinn | 287 |
| 8 | Debona | 203 |
| 9 | Verapak | 111 |
| 10 | Averinn | 92 |
| 11 | Fastbatllnn | 11 |

Table 22: Incorrect Results (or Missing CE)

| # | **Tool** | **Count** |
|---|---|---|
| 1 | Cgdtest | 45 |
| 2 | $\alpha$,$\beta$-CROWN | 25 |
| 3 | Marabou | 12 |
| 4 | Verapak | 5 |