Comparison with DeepDiagnosis

We have got better results compared to the reported results of DeepDiagnosis on similar benchmarks. The DeepDiagnosis reproducibility package [https://github.com/DeepDiagnosis/ICSE2022] is not publicly available yet.

With the Deeplocalize and Umlaut benchmarks, DeepDiagnosis can successfully detect bugs in 46 out of 53 buggy programs, whereas DL Contract detects bugs in 51 out of 53 buggy programs.

In AUTOTRAINER benchmark, DeepDiagnosis detects bugs in 138 out of 203 buggy programs, whereas DL Contractdetects 195 out of 203 buggy programs.

However, we have also evaluated using the correct (clean) version of 257 programs from all those benchmarks and found only 18 false positives. Our comprehensive evaluation also advocates the utility of DL Contract technique to detect such bugs following the DbC approach. By using DL Contract, DL application developers directly get benefits without writing additional lines of code.

	Deeplocalize+UMLAUT Benchmark	**AUTOTRAINER Benchmark**
Technique	**Detected Bugs (out of 53)**	
DeepDiagnosis	**46**	**138**
DL Contract	**51**	**195**