

# Just-In-Time Software Defect Prediction

Ali Mohamed, Yifei Gong, Yunhua Zhao

Project Category: binary **classification** for **buggy** commits

Data: two datasets:

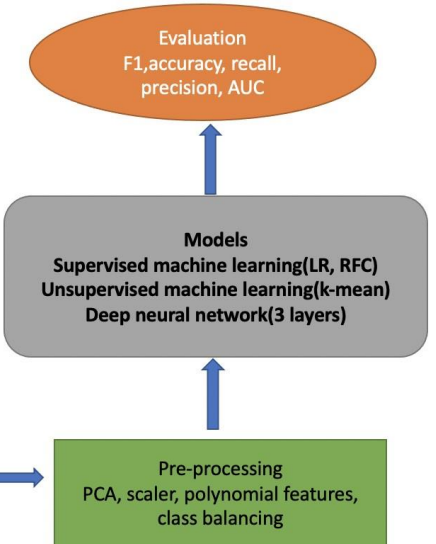
- **Openstack** and **Qt**
- Openstack (10658 negative samples, 1616 positive samples), Qt (23148 negative samples, 2002 positive samples).
- Each sample contains information from a GitHub commit.
- Each sample has 35 features extracted from that commit (lines modified, developer experience, etc.) and 1 target values (buggy or not).

Goals: compete with state-of-the-art work by using handcrafted features

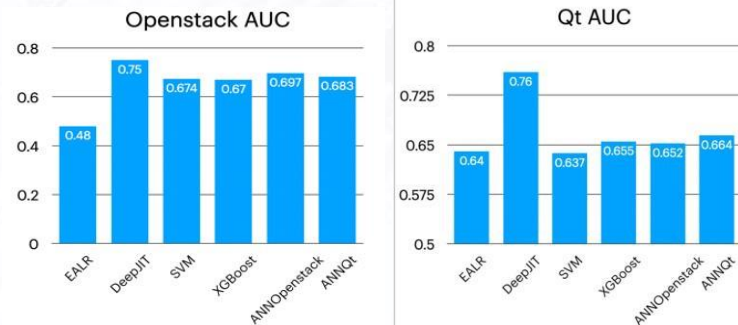
```
diff --git a/qmake/generators/win32/msvc_vcproj.cpp b/qmake/generators/win32/msvc_vcproj.cpp
index e9ba12828..cf80d249 18064
--- a/qmake/generators/win32/msvc_vcproj.cpp
+++ b/qmake/generators/win32/msvc_vcproj.cpp
@@ -585,7 +585,7 @@ ProStringlist VcpkgGenerator::collectDependencies(QMakeProject *proj, QHash<Qt
    wct += where.end(); ++wct;
    const ProStringlist &l = tmp_proj.values(ProKey("wct"));
    for (ProStringlist::ConstIterator it = l.begin(); it != l.end(); ++it) {
        const QString opt = QString::fromLatin1(*it);
        (String opt = (*it).toQstring());
        if (!opt.startsWith("/")) { // Not a switch
            opt = "nologo-target && // Not a switch
            opt = "nologo-target && // We don't care about these libs
        }
    }
}
```

```
19923559 modified file: src/widgets/doc/snippets/macmainwindow.mm
19923560 file status('insertions': 2, 'deletions': 0, 'lines': 2)
19923561 modified file: src/widgets/doc/src/qtwidgets-index.qdoc
19923562 file status('insertions': 2, 'deletions': 2, 'lines': 4)
19923563 modified file: src/widgets/doc/src/widgets-and-layouts/gallery-windows.qdoc
19923564 file status('insertions': 142, 'deletions': 0, 'lines': 142)
19923565 modified file: src/widgets/doc/src/widgets-and-layouts/gallery.qdoc
19923566 file status('insertions': 4, 'deletions': 0, 'lines': 4)
```

```
30c[24]:
commit_id author_date bugcount flcount la ls nt nd ns ent ... rkeep okeep esowr rswor osow
0 00002c8e8c8e363934b0d794d01f852a66 135185555 0 0 2 0 1 1 1 0.000000 ... 82.0 130.0 0.013219 0.018286 0.02886
1 00016c3a2b57683252746293754a8a07a0 1359594574 0 0 0 53 42 0 4 1 0.846452 ... NaN NaN 0.000000 0.000000 0.00000
2 00039a2654e4d18f6a052a555e4e1c3b6 137399362 0 0 28 4 2 1 1 0.974459 ... NaN NaN 0.000000 0.000000 0.00000
3 000625d5d47e042c677103c115e7148e10 136304065 0 0 12 10 4 1 1 0.665790 ... 1406.0 1673.0 0.026888 0.114432 0.12796
4 000842990881068230e5706490d8592a28 1320409350 0 0 53 36 6 4 2 0.731395 ... 246.0 285.0 0.011343 0.087558 0.09429
5 rows x 35 columns
```



Hypothesis 1: two datasets may share intrinsic similarities that can render better results when trained together

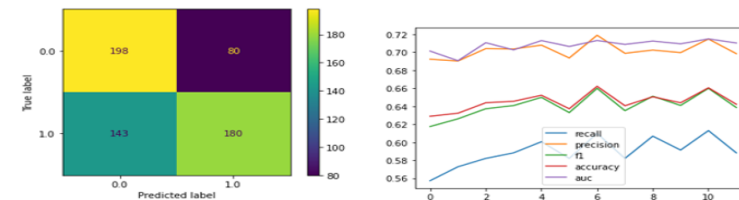


EALR and DeepJIT are state-of-the-art work. The rest four models are trained on combined oversampled datasets. ANNOpenstack and ANNQt uses feed forward neural network.

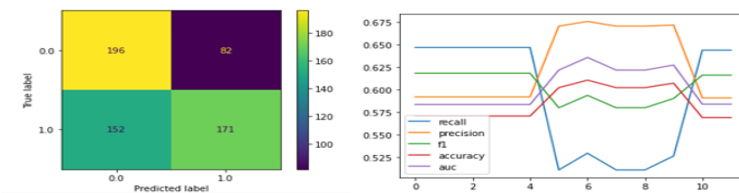
- 4 models can outperform EALR
- Still a gap with DeepJIT

## Comparing the 2 main Supervised Machine Learning Models

Random Forest:



Logistic Regression:



The Random Forest Model seems to be slightly better, partially due to the size of the dataset – performance difference can be attributed to size.