1 STATISTICAL ANALYSIS

We performed statistical analysis in results for all research questions (RQs), as follow.

1.1 RQ1: How different window sizes affect the prediction performance of the models?

Figure 1 presents differences with statistical significance for window size values 2, 3 and 4. We used Kruskal Wallis [1] and the p-value was 0.7, considering all results for each size. Results do not have difference with statistical significance between evaluated values.

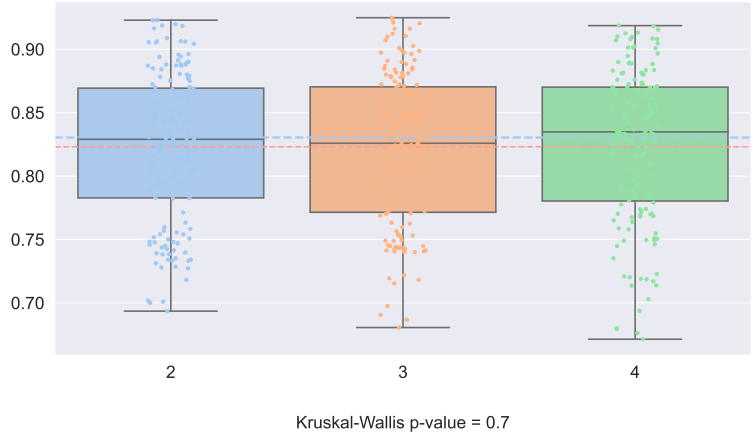

Figure 1: Window size statistical analysis - ROC AUC Score

1.2 RQ2: What is the performance of our approach with respect to the performance of a traditional approach?

We aplied the Wilcoxon [3] test and calculated the p-value for all results of ROC AUC, F1-score, accuracy and sensitivity performance metrics. For all metrics, there's a statistical difference in the distribution of history-based and traditional approaches with >95% of confidence (p-value=0.001), as Figures 2, 4, 3, 5 show.

The effect size for ROC AUC presented a large magnitude and can be seen in Table 1. F1 Score and Accuracy had a medium magnitude and Sensitivity had small effect size magnitude. These results can be seen in Tables 3, 2, and 4.

For effect size analyses, we used the following definitions: Values highlighted with a "★" symbol denotes the algorithm with the best score. A "▼" indicates that the effect size was negligible in relation to the best value, while "▽" denotes a small magnitude, "△" a medium magnitude, and "▲" a large magnitude[2]. The effect size was performed during the post-hoc tests, that is, when there is a statistical difference.

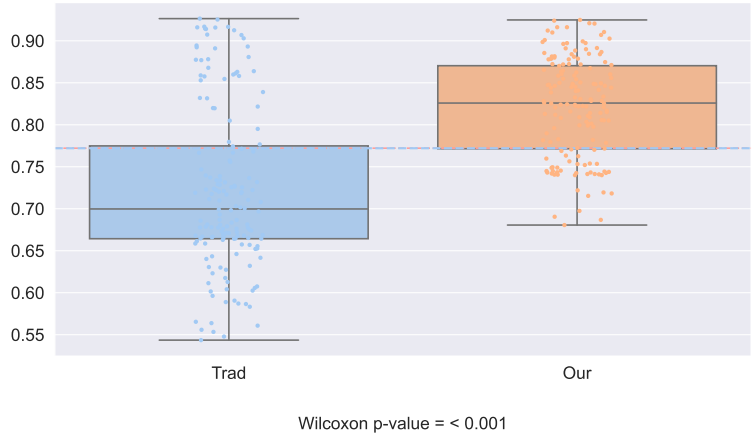

Figure 2: ROC AUC Score

Table 1: Effect size - ROC AUC Score

Approach	ROC	Effect Size
Our	0.8217 ± 0.0590	★ best result
Trad	0.7226 ± 0.0990	▲ large magnitude

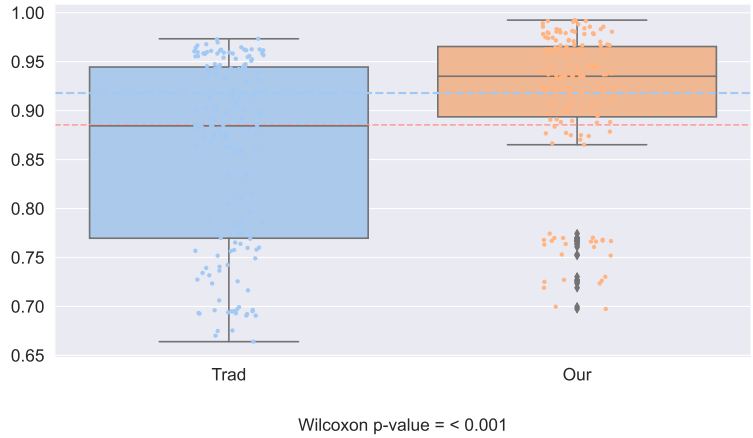

Figure 3: Accuracy

Table 2: Effect size - Accuracy

Approach	Accuracy	Effect Size
Our	0.9118 ± 0.0760	★ best result
Trad	0.8591 ± 0.0950	△ medium magnitude

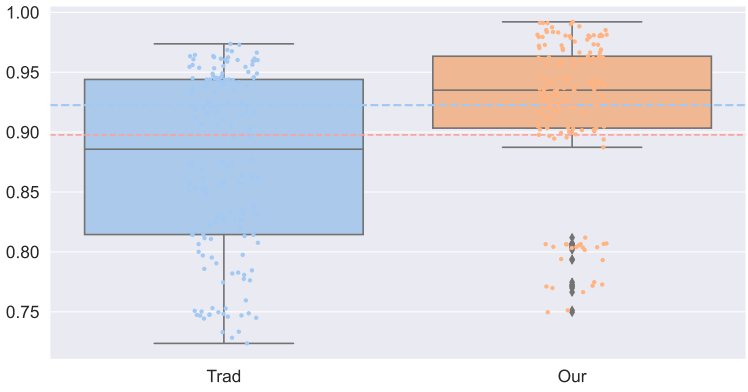

Figure 4: F1 Score

Table 3: Effect size - F1 Score

Approach	F1	Effect Size
Our	0.9208 ± 0.0610	★ best result
Trad	0.8746 ± 0.0740	△ medium magnitude

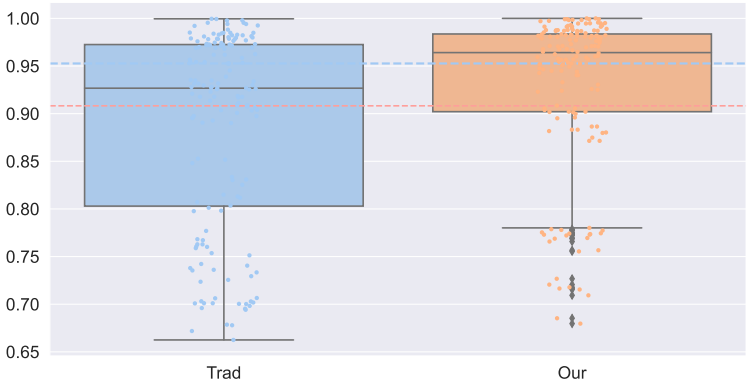

Figure 5: Sensitivity

Table 4: Effect size - Sensitivity

Approach	Sensitivity	Effect Size
Our	0.9293 ± 0.0830	★ best result
Trad	0.8872 ± 0.1030	▽ small magnitude

1.3 RQ3: What is the impact of using smell related information to predict change proneness using our approach?

Models that use smells don't presented statistical difference as shown in Figure 6.

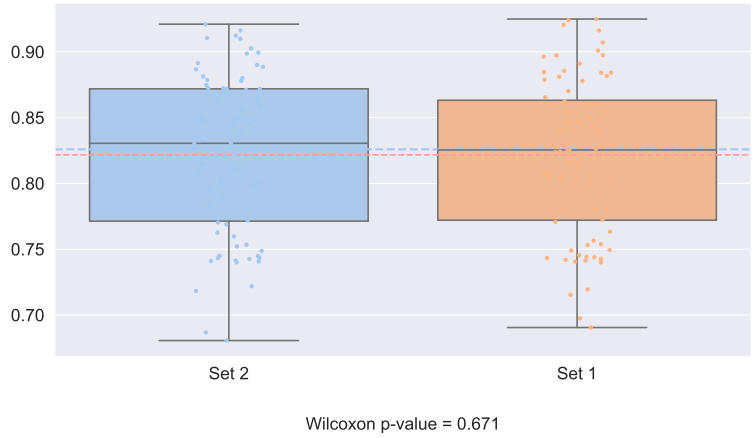

Figure 6: Statistical analysis using smell based information - ROC

1.4 RQ4: What is the algorithm that leads our approach to produce the best results when answering all RQs?

We applied Kruskal-Wallis [1] and calculated the p-value for all results of ROC AUC, F1-score, accuracy and sensitivity performance metrics. The statistical difference for the algorithms had 95% of confidence, as shown in Figures 7, 8, 9. The effect size for ROC AUC presented medium and small magnitudes and can be seen in Table 5. F1 Score and Accuracy had a large magnitude, these results can be seen in Tables 6 and 7.

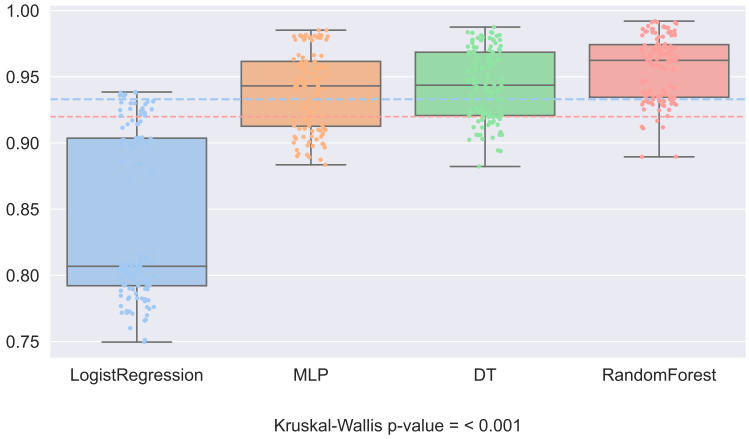

Figure 7: Statistical analysis Algorithms - ROC

Table 5: Effect Size - ROC AUC Score

Algorithm	ROC	Effect Size
DT	0.8367 ± 0.0310	△ medium magnitude
LogistRegression	0.7953 ± 0.0690	△ medium magnitude
MLP	0.8163 ± 0.0560	▽ small magnitude
RandomForest	0.8386 ± 0.0620	★ best result

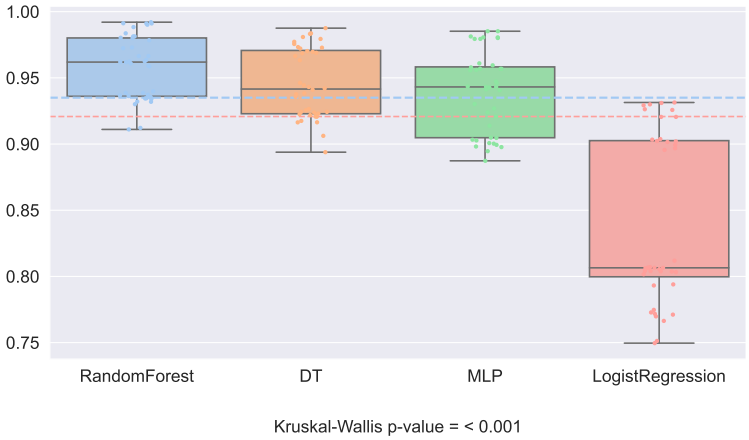

Figure 8: Statistical analysis Algorithms - F1 Score

Table 6: Effect Size - F1 Score

Algorithm	F1	Effect Size
DT	0.9457 ± 0.0260	▲ large magnitude
LogistRegression	0.8403 ± 0.0630	▲ large magnitude
MLP	0.9392 ± 0.0310	▲ large magnitude
RandomForest	0.9579 ± 0.0240	★ best result

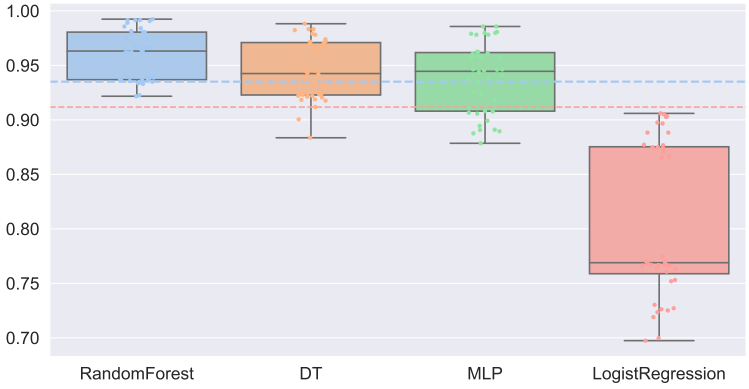

Figure 9: Statistical analysis Algorithms - Accuracy

Table 7: Effect Size - Accuracy

Algorithm	Accuracy	Effect Size
DT	0.9449 ± 0.0270	▲ large magnitude
LogistRegression	0.8040 ± 0.0700	▲ large magnitude
MLP	0.9385 ± 0.0320	▲ large magnitude
RandomForest	0.9596 ± 0.0230	★ best result

REFERENCES

[1] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621.
[2] Jackson A. Prado Lima and Silvia R. Vergilio. 2020. A Multi-Armed Bandit Approach for Test Case Prioritization in Continuous Integration Environments. *IEEE Transactions on Software Engineering* (2020), 12.
[3] Denise Rey and Markus Neuhäuser. 2011. *Wilcoxon-Signed-Rank Test.* Springer Berlin Heidelberg, Berlin, Heidelberg, 1658–1659. https://doi.org/10.1007/978-3-642-04898-2_616