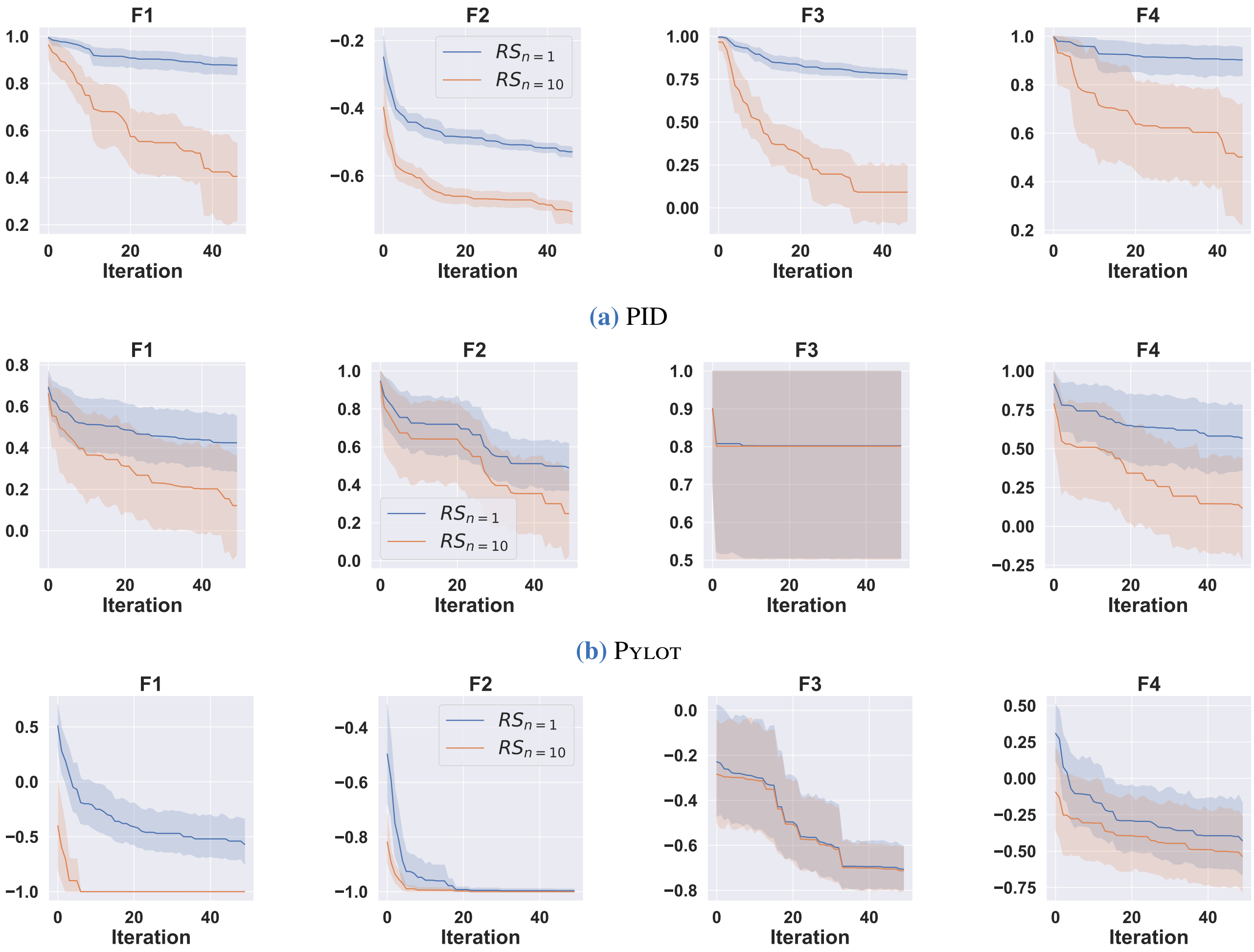
Evaluating the Impact of Flaky Simulators on Testing Autonomous Driving Systems - Supplementary Material

Contents

RQ1

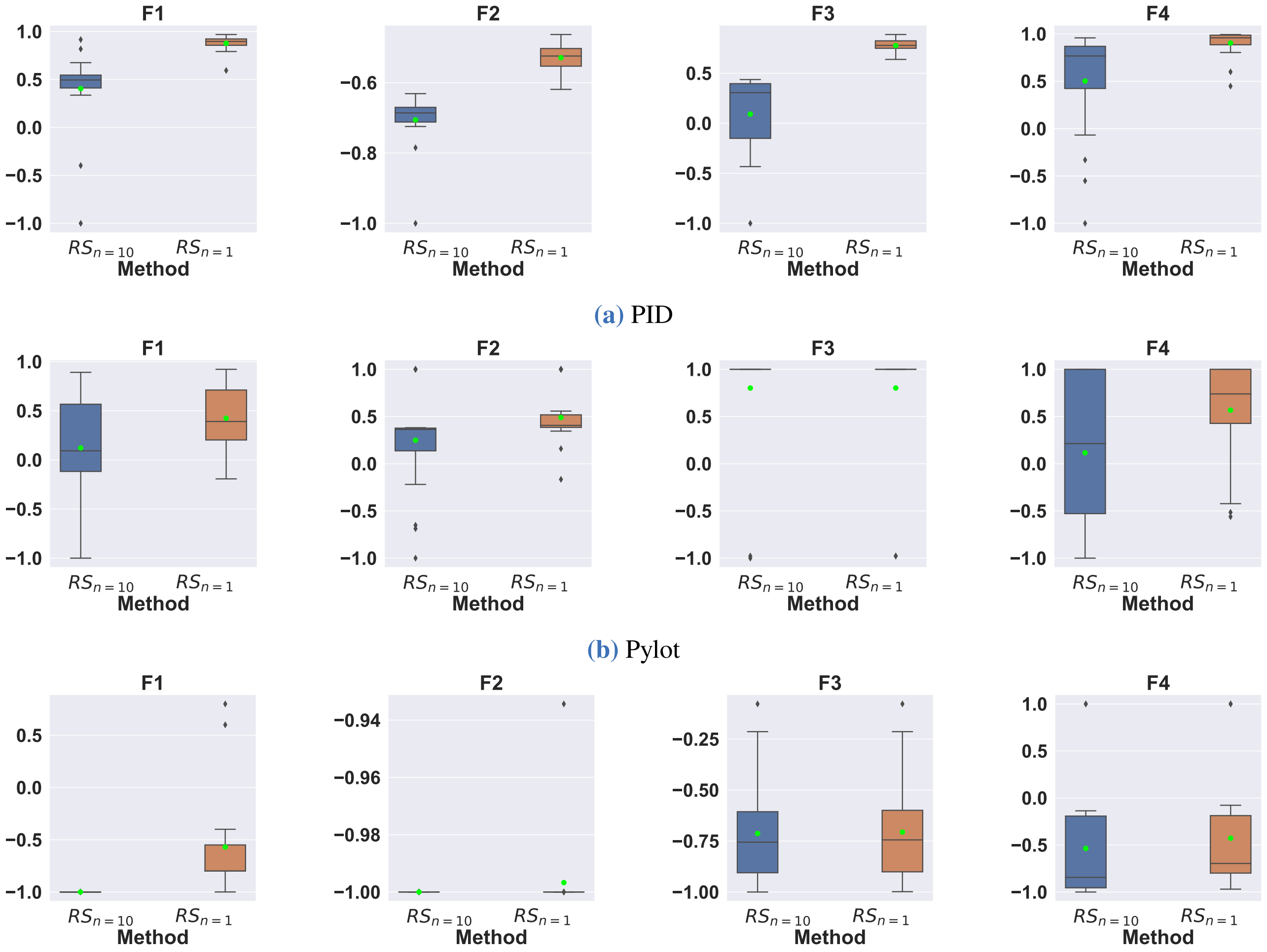


(a) PID

(b) PYLOT

(c) BEAMNG

Figure 1: The average and 95% interval of the best fitness values obtained by 20 runs of $RS_{n=1}$ and $RS_{n=10}$ over 50 iterations for four fitness functions of PID, PYLOT and BEAMNG. (Related to RQ1-3)



(a) PID

(b) Pylot

(c) BeamNG

Figure 2: Comparing the distributions and averages of the best fitness values obtained from 20 runs of $RS(n = 1)$ and $RS(n = 10)$ at the last iteration from Figure 1. (Related to RQ1-3)

RQ2

PID

Threshold: 5%

Table 1: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
MLP	With delta fitnesses	0.96	0.94	0.95
MLP	With delta fitnesses wøblueprints	0.96	0.87	0.92
MLP	With delta fitnesses wøweather	0.97	0.86	0.91
MLP	With delta fitnesses wøweather, blueprints	0.97	0.77	0.86
Random Forest	With delta fitnesses wøblueprints	0.97	0.75	0.85
Random Forest	With delta fitnesses	0.99	0.74	0.85
Decision Tree	With delta fitnesses wøblueprints	0.99	0.73	0.84
Random Forest	With delta fitnesses wøweather, blueprints	0.98	0.73	0.84
Decision Tree	With delta fitnesses wøweather, blueprints	0.99	0.73	0.84
Random Forest	With delta fitnesses wøweather	0.99	0.73	0.84
Decision Tree	With delta fitnesses wøweather	0.97	0.74	0.84
Decision Tree	With delta fitnesses	0.97	0.74	0.84
MLP	With 1 set of fitnesses wøweather	0.72	0.82	0.77
MLP	With 1 set of fitnesses wøblueprints	0.72	0.74	0.73
Decision Tree	With 1 set of fitnesses	0.82	0.65	0.72
Random Forest	With 1 set of fitnesses wøweather, blueprints	0.77	0.66	0.71
MLP	With 1 set of fitnesses	0.79	0.64	0.71
Decision Tree	With 1 set of fitnesses wøweather	0.80	0.62	0.70
Random Forest	With 1 set of fitnesses wøblueprints	0.79	0.63	0.70
Decision Tree	With 1 set of fitnesses wøweather, blueprints	0.83	0.60	0.69
Random Forest	With 1 set of fitnesses	0.80	0.61	0.69
Decision Tree	With 1 set of fitnesses wøblueprints	0.83	0.59	0.69
Random Forest	With 1 set of fitnesses wøweather	0.79	0.60	0.68
MLP	With 1 set of fitnesses wøweather, blueprints	0.72	0.65	0.68
SVM	With delta fitnesses	0.99	0.40	0.57
SVM	With delta fitnesses wøweather, blueprints	0.99	0.40	0.57
SVM	With delta fitnesses wøblueprints	0.99	0.40	0.57
SVM	With delta fitnesses wøweather	0.99	0.40	0.57
SVM	With 1 set of fitnesses wøblueprints	0.70	0.37	0.49
SVM	With 1 set of fitnesses	0.71	0.37	0.48
SVM	With 1 set of fitnesses wøweather, blueprints	1.00	0.06	0.12
SVM	With 1 set of fitnesses wøweather	1.00	0.06	0.12

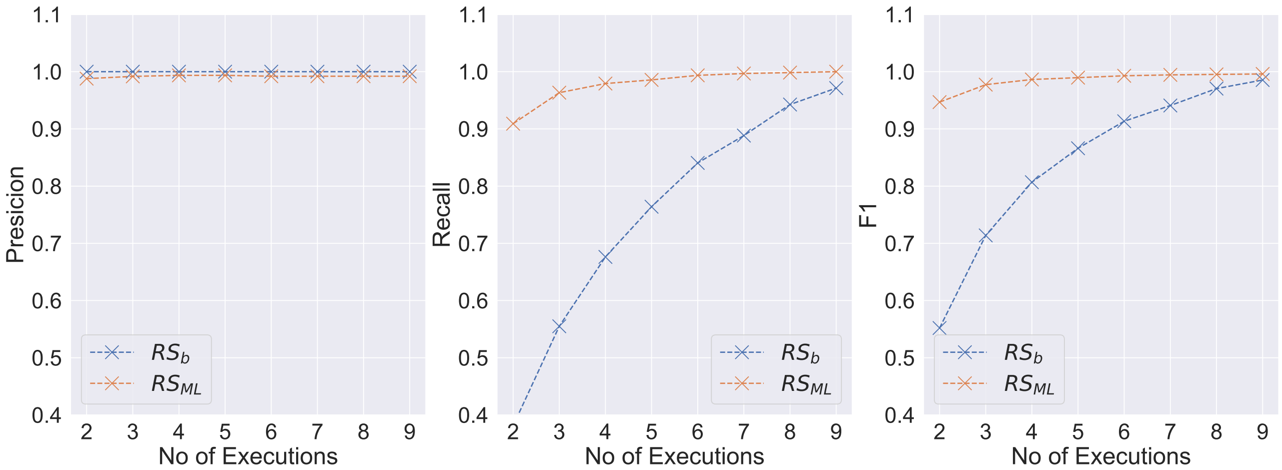


Figure 3: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 2: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.381180	0.551963	0.993056	0.912281	0.950956
3	1.0	0.555024	0.713846	0.990148	0.961722	0.975728
4	1.0	0.676236	0.806851	0.990307	0.977671	0.983949
5	1.0	0.763955	0.866184	0.990385	0.985646	0.988010
6	1.0	0.840510	0.913345	0.988871	0.992026	0.990446
7	1.0	0.888357	0.940878	0.988889	0.993620	0.991249
8	1.0	0.942584	0.970443	0.987362	0.996810	0.992063
9	1.0	0.971292	0.985437	0.985827	0.998405	0.992076

Table 3: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{12}	F2 A^{12}	F3 A^{12}	F4 A^{12}
RS_b	5323	1.421085e-14	1.421085e-14	1.421085e-14	1.421085e-14	0.100951	0.081032	0.105025	0.086012
RS_{ML}	5338	1.421085e-14	1.421085e-14	1.421085e-14	1.421085e-14	0.091444	0.053871	0.082390	0.078316

Table 4: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{12}	F2 A^{12}	F3 A^{12}	F4 A^{12}
RS_b	5323	0.001359	1.421085e-14	0.01252	1.421085e-14	0.55636	0.779538	0.537347	0.60593

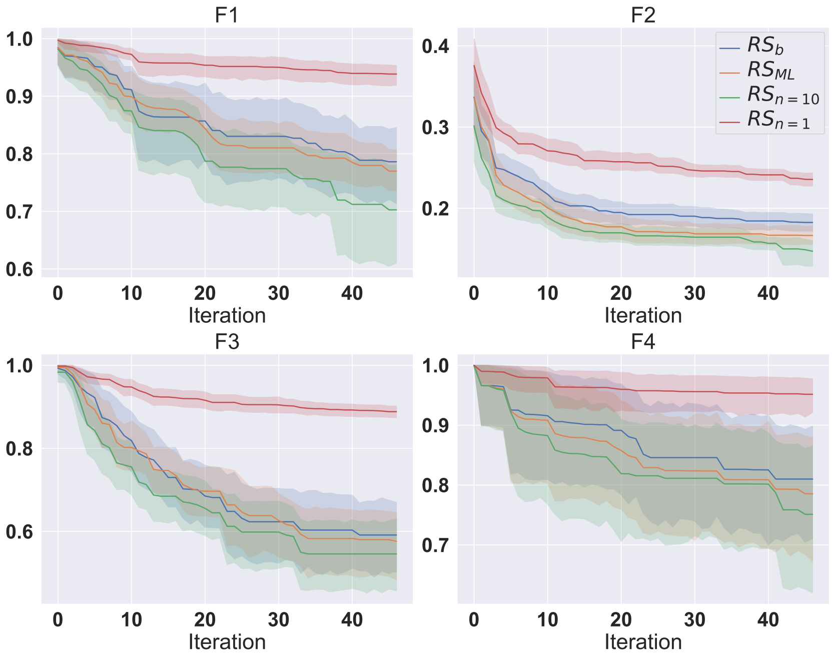


Figure 4: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PID (Related to RQ2-2)

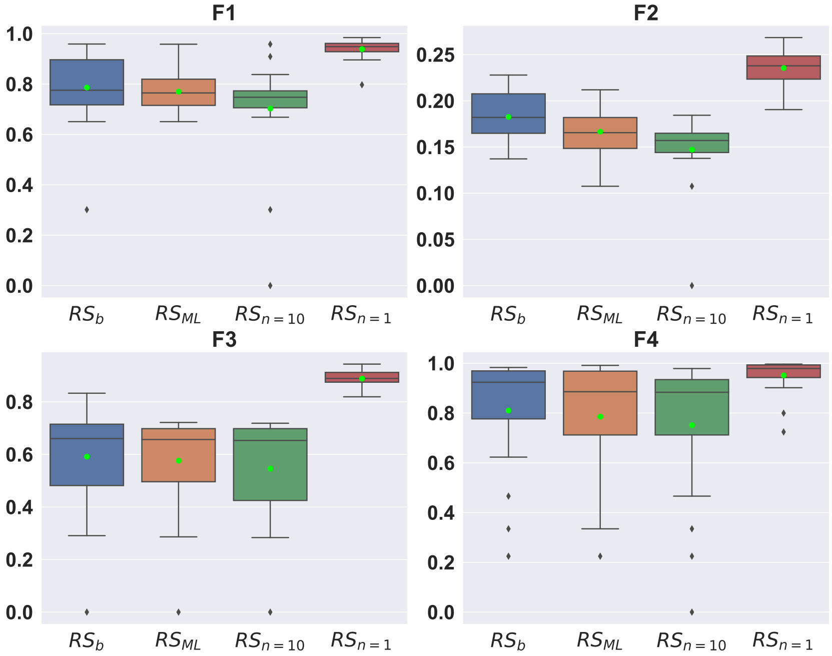


Figure 5: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PID (Related to RQ2-2).

Threshold: 10%

Table 5: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
MLP	With delta fitnesses	0.98	0.86	0.90
MLP	With delta fitnesses wøblueprints	0.95	0.80	0.84
MLP	With delta fitnesses wøweather	0.96	0.80	0.84
Random Forest	With delta fitnesses	0.99	0.69	0.81
Random Forest	With delta fitnesses wøweather	0.99	0.68	0.81
Random Forest	With delta fitnesses wøweather, blueprints	0.99	0.69	0.80
Random Forest	With delta fitnesses wøblueprints	0.99	0.69	0.80
MLP	With delta fitnesses wøweather, blueprints	0.95	0.71	0.80
Decision Tree	With delta fitnesses wøblueprints	1.00	0.68	0.80
Decision Tree	With delta fitnesses wøweather, blueprints	0.99	0.67	0.79
Decision Tree	With delta fitnesses	1.00	0.67	0.79
Decision Tree	With delta fitnesses wøweather	0.99	0.67	0.79
MLP	With 1 set of fitnesses wøweather	0.92	0.96	0.74
MLP	With 1 set of fitnesses wøweather, blueprints	1.00	0.86	0.71
MLP	With 1 set of fitnesses	0.79	0.87	0.71
MLP	With 1 set of fitnesses wøblueprints	0.74	0.78	0.68
Random Forest	With 1 set of fitnesses	0.76	0.63	0.66
Random Forest	With 1 set of fitnesses wøweather	0.75	0.59	0.64
Random Forest	With 1 set of fitnesses wøweather, blueprints	0.71	0.61	0.63
Decision Tree	With 1 set of fitnesses wøweather, blueprints	0.77	0.60	0.63
Random Forest	With 1 set of fitnesses wøblueprints	0.75	0.59	0.63
Decision Tree	With 1 set of fitnesses wøweather	0.78	0.57	0.63
Decision Tree	With 1 set of fitnesses wøblueprints	0.77	0.59	0.62
Decision Tree	With 1 set of fitnesses	0.78	0.57	0.62
SVM	With delta fitnesses wøweather, blueprints	0.97	0.45	0.62
SVM	With delta fitnesses wøblueprints	0.97	0.45	0.62
SVM	With delta fitnesses wøweather	0.98	0.45	0.62
SVM	With delta fitnesses	0.97	0.45	0.62
SVM	With 1 set of fitnesses wøblueprints	1.00	0.15	0.26
SVM	With 1 set of fitnesses	1.00	0.15	0.26
SVM	With 1 set of fitnesses wøweather, blueprints	1.00	0.07	0.13
SVM	With 1 set of fitnesses wøweather	1.00	0.07	0.13

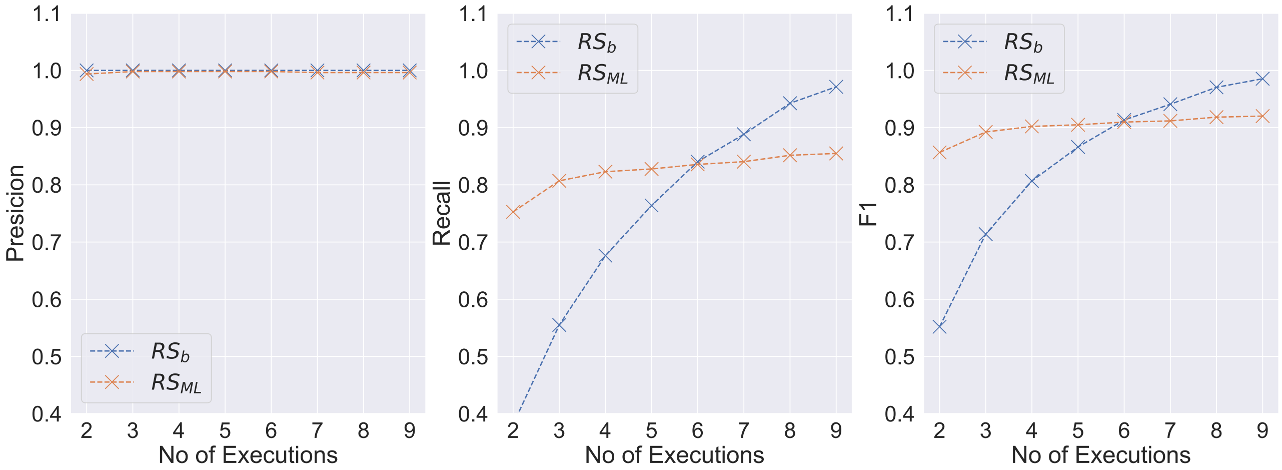


Figure 6: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 6: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.381180	0.551963	0.989858	0.778309	0.871429
3	1.0	0.555024	0.713846	0.990253	0.810207	0.891228
4	1.0	0.676236	0.806851	0.992308	0.822967	0.899738
5	1.0	0.763955	0.866184	0.992410	0.834131	0.906412
6	1.0	0.840510	0.913345	0.992467	0.840510	0.910190
7	1.0	0.888357	0.940878	0.992495	0.843700	0.912069
8	1.0	0.942584	0.970443	0.992593	0.854864	0.918595
9	1.0	0.971292	0.985437	0.992593	0.854864	0.918595

Table 7: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^12	F2 A^12	F3 A^12	F4 A^12
Baseline-Manual	6018	1.421085e-14	1.421085e-14	1.421085e-14	1.421085e-14	0.097329	0.059303	0.102309	0.080127
fastFitness	3517	1.421085e-14	1.421085e-14	1.421085e-14	2.842171e-14	0.111363	0.092349	0.087823	0.099593

Table 8: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^12	F2 A^12	F3 A^12	F4 A^12
Baseline-Manual	6018	1.946887e-12	0.013751	0.000113	1.421085e-14	0.328203	0.676777	0.422816	0.273427

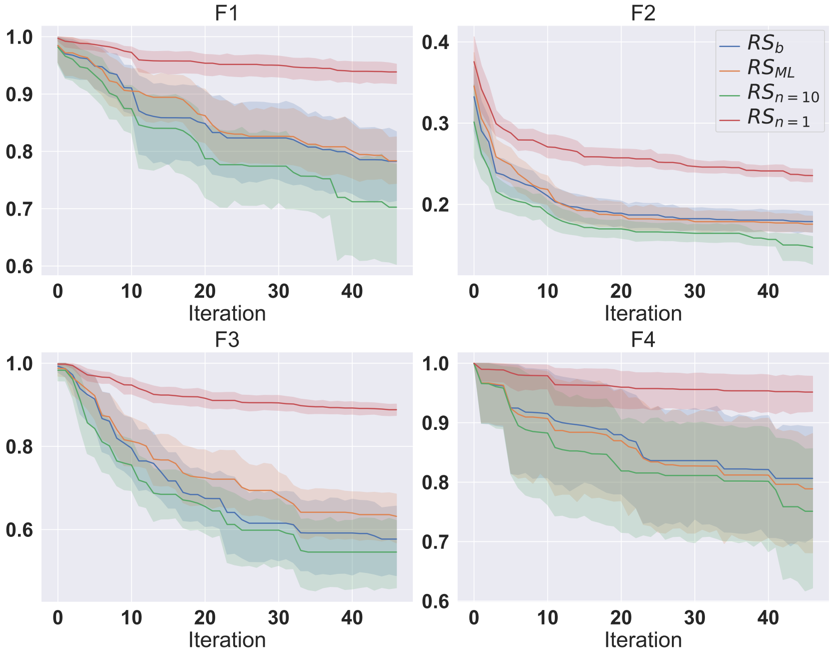


Figure 7: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PID (Related to RQ2-2)

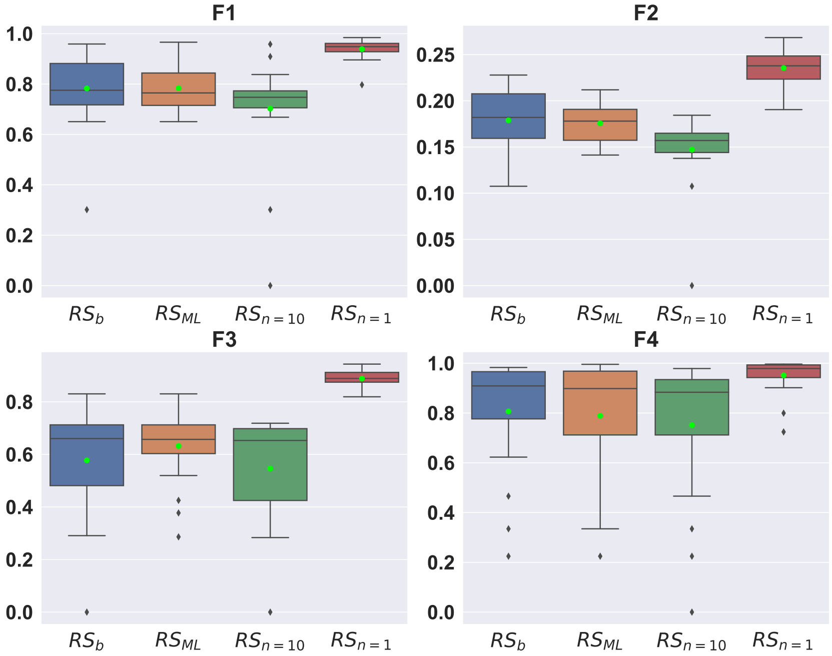


Figure 8: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PID (Related to RQ2-2).

Threshold: 20%

Table 9: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
MLP	With delta fitnesses	0.95	0.90	0.85
Random Forest	With delta fitnesses wøblueprints	1.00	0.61	0.75
Random Forest	With delta fitnesses	1.00	0.60	0.75
Decision Tree	With delta fitnesses wøweather, blueprints	1.00	0.62	0.74
Random Forest	With delta fitnesses wøweather	1.00	0.59	0.74
Random Forest	With delta fitnesses wøweather, blueprints	1.00	0.59	0.74
MLP	With delta fitnesses wøweather	0.86	0.79	0.74
Decision Tree	With delta fitnesses wøweather	1.00	0.58	0.73
Decision Tree	With delta fitnesses	1.00	0.58	0.73
Decision Tree	With delta fitnesses wøblueprints	1.00	0.58	0.73
MLP	With delta fitnesses wøblueprints	0.88	0.69	0.72
MLP	With delta fitnesses wøweather, blueprints	0.90	0.63	0.70
Decision Tree	With 1 set of fitnesses wøweather	0.65	0.76	0.68
Decision Tree	With 1 set of fitnesses	0.63	0.76	0.68
SVM	With delta fitnesses wøweather	0.98	0.49	0.66
SVM	With delta fitnesses wøweather, blueprints	0.98	0.49	0.65
SVM	With delta fitnesses	0.98	0.49	0.65
SVM	With delta fitnesses wøblueprints	0.98	0.49	0.65
MLP	With 1 set of fitnesses wøweather	0.90	0.87	0.63
MLP	With 1 set of fitnesses	0.58	0.84	0.63
Random Forest	With 1 set of fitnesses	0.66	0.62	0.62
MLP	With 1 set of fitnesses wøweather, blueprints	1.00	0.93	0.59
MLP	With 1 set of fitnesses wøblueprints	0.62	0.67	0.59
Random Forest	With 1 set of fitnesses wøblueprints	0.65	0.52	0.58
Random Forest	With 1 set of fitnesses wøweather	0.62	0.52	0.56
Decision Tree	With 1 set of fitnesses wøblueprints	0.63	0.67	0.55
Decision Tree	With 1 set of fitnesses wøweather, blueprints	0.62	0.62	0.54
Random Forest	With 1 set of fitnesses wøweather, blueprints	0.60	0.43	0.49
SVM	With 1 set of fitnesses	0.85	0.33	0.43
SVM	With 1 set of fitnesses wøblueprints	0.86	0.23	0.33
SVM	With 1 set of fitnesses wøweather, blueprints	0.86	0.14	0.24
SVM	With 1 set of fitnesses wøweather	0.86	0.14	0.24

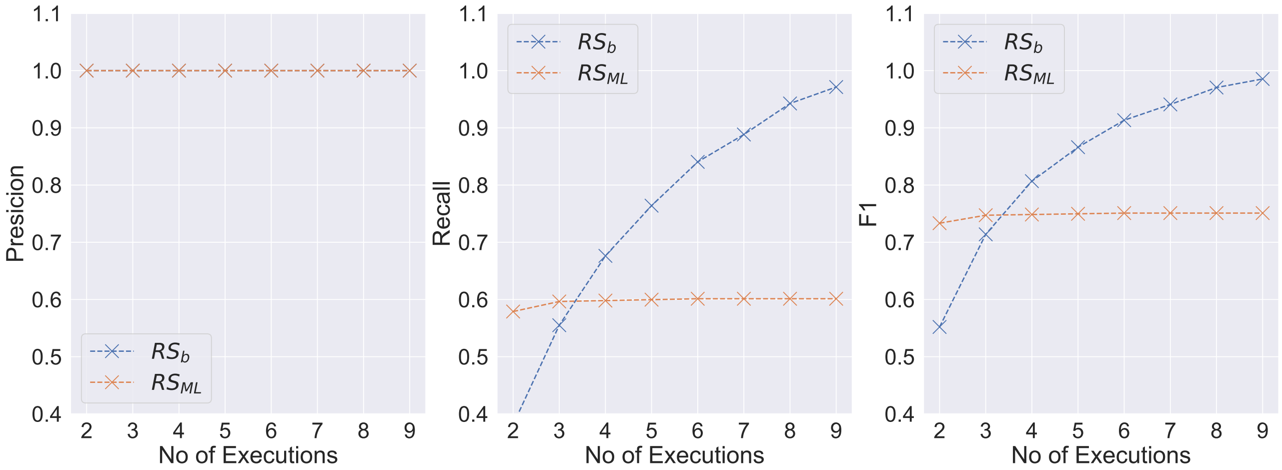


Figure 9: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 10: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.381180	0.551963	1.0	0.577352	0.732053
3	1.0	0.555024	0.713846	1.0	0.594896	0.746000
4	1.0	0.676236	0.806851	1.0	0.601276	0.750996
5	1.0	0.763955	0.866184	1.0	0.601276	0.750996
6	1.0	0.840510	0.913345	1.0	0.601276	0.750996
7	1.0	0.888357	0.940878	1.0	0.601276	0.750996
8	1.0	0.942584	0.970443	1.0	0.601276	0.750996
9	1.0	0.971292	0.985437	1.0	0.601276	0.750996

Table 11: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^12	F2 A^12	F3 A^12	F4 A^12
Baseline-Manual	6852	1.421085e-14	1.421085e-14	1.421085e-14	1.421085e-14	0.093708	0.045722	0.092802	0.079674
fastFitness	3526	1.421085e-14	1.421085e-14	1.421085e-14	2.842171e-14	0.096424	0.089633	0.114984	0.080579

Table 12: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^12	F2 A^12	F3 A^12	F4 A^12
Baseline-Manual	6852	1.989520e-13	1.421085e-14	1.421085e-14	0.074876	0.395201	0.260299	0.386148	0.453825

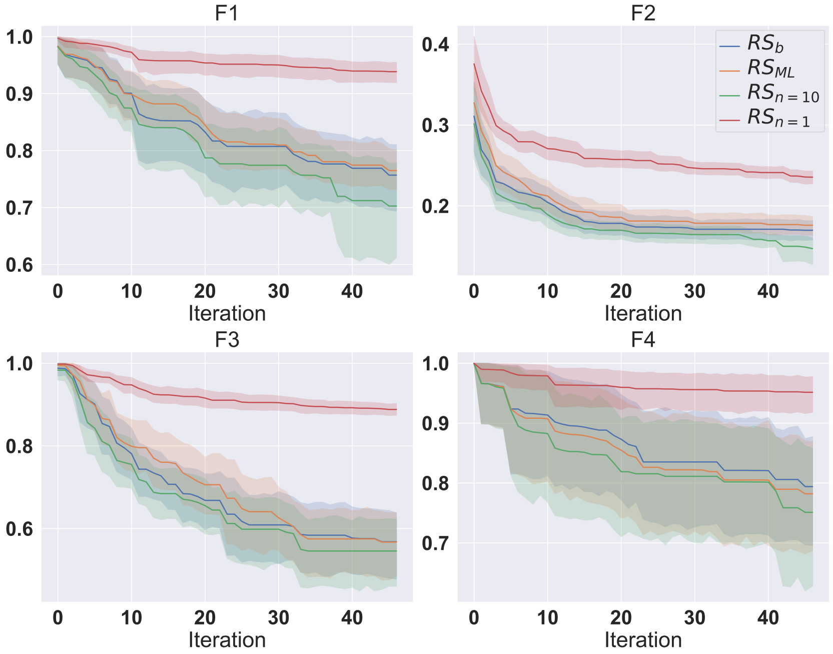


Figure 10: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PID (Related to RQ2-2)

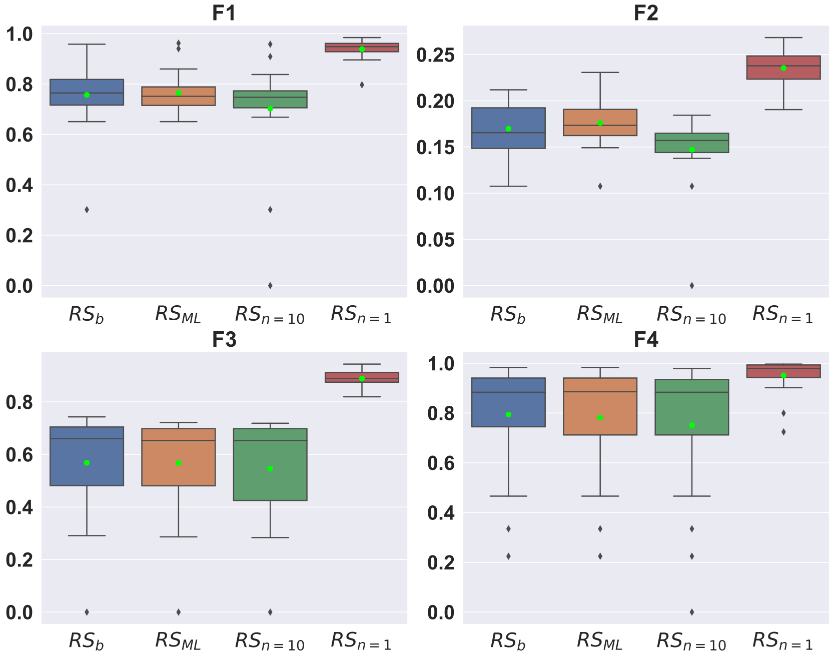


Figure 11: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PID (Related to RQ2-2).

Pylot

Threshold: 5%

Table 13: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
MLP	With delta fitnesses	0.99	0.95	0.97
MLP	With delta fitnesses wøweather	0.98	0.93	0.95
MLP	With delta fitnesses wøblueprints	0.92	0.93	0.92
MLP	With delta fitnesses wøweather, blueprints	0.90	0.88	0.89
Random Forest	With delta fitnesses	0.94	0.83	0.88
Random Forest	With delta fitnesses wøweather	0.94	0.82	0.88
Random Forest	With delta fitnesses wøblueprints	0.93	0.82	0.87
Random Forest	With delta fitnesses wøweather, blueprints	0.93	0.81	0.87
Decision Tree	With delta fitnesses	0.94	0.80	0.87
Decision Tree	With delta fitnesses wøweather	0.95	0.79	0.86
Decision Tree	With delta fitnesses wøblueprints	0.95	0.79	0.86
Decision Tree	With delta fitnesses wøweather, blueprints	0.95	0.79	0.86
SVM	With delta fitnesses wøweather	0.86	0.83	0.84
SVM	With delta fitnesses	0.81	0.81	0.81
Random Forest	With 1 set of fitnesses wøweather	0.78	0.81	0.79
Random Forest	With 1 set of fitnesses wøblueprints	0.71	0.76	0.73
Random Forest	With 1 set of fitnesses	0.71	0.76	0.73
MLP	With 1 set of fitnesses wøweather	0.71	0.73	0.72
SVM	With 1 set of fitnesses wøweather	0.62	0.82	0.71
Decision Tree	With 1 set of fitnesses	0.75	0.66	0.70
SVM	With delta fitnesses wøblueprints	0.72	0.67	0.70
MLP	With 1 set of fitnesses	0.63	0.78	0.69
Decision Tree	With 1 set of fitnesses wøweather	0.64	0.75	0.69
Random Forest	With 1 set of fitnesses wøweather, blueprints	0.64	0.75	0.69
Decision Tree	With 1 set of fitnesses wøweather, blueprints	0.76	0.63	0.69
Decision Tree	With 1 set of fitnesses wøblueprints	0.76	0.63	0.69
SVM	With delta fitnesses wøweather, blueprints	0.74	0.63	0.68
SVM	With 1 set of fitnesses wøweather, blueprints	0.61	0.72	0.66
MLP	With 1 set of fitnesses wøweather, blueprints	0.62	0.69	0.65
SVM	With 1 set of fitnesses	0.57	0.75	0.65
MLP	With 1 set of fitnesses wøblueprints	0.59	0.57	0.58
SVM	With 1 set of fitnesses wøblueprints	0.50	0.63	0.56

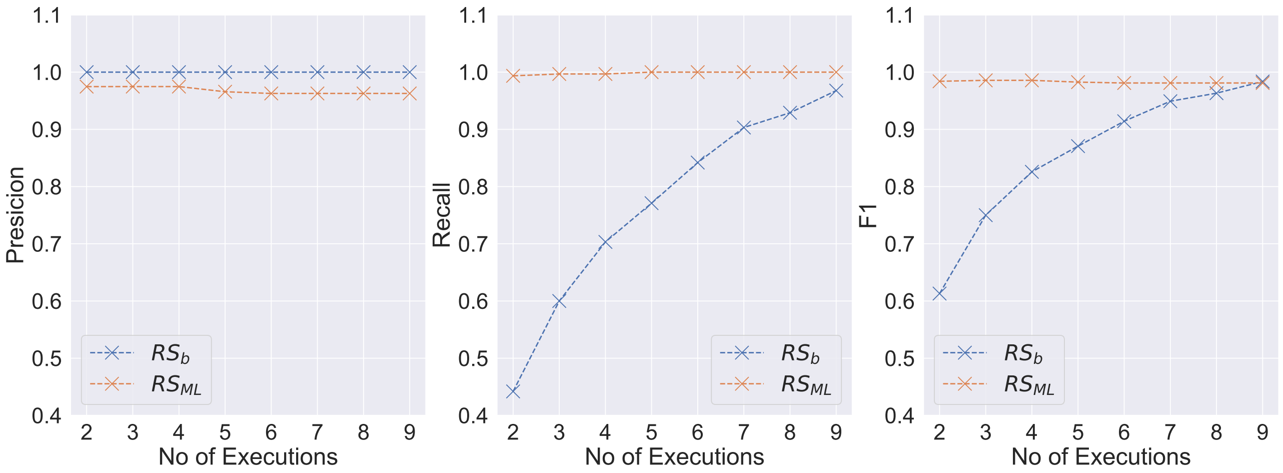


Figure 12: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 14: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.441935	0.612975	0.980645	0.980645	0.980645
3	1.0	0.600000	0.750000	0.980831	0.990323	0.985554
4	1.0	0.703226	0.825758	0.980892	0.993548	0.987179
5	1.0	0.770968	0.870674	0.980952	0.996774	0.988800
6	1.0	0.841935	0.914186	0.981013	1.000000	0.990415
7	1.0	0.903226	0.949153	0.977918	1.000000	0.988836
8	1.0	0.929032	0.963211	0.977918	1.000000	0.988836
9	1.0	0.967742	0.983607	0.974843	1.000000	0.987261

Table 15: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	7391	6.527722e-01	4.610039e-02	1.399457e-02	1.776357e-15	0.5164	0.5120	0.2330	0.0468
fastFitness	4030	1.776357e-15	1.776357e-15	9.652696e-11	1.776357e-15	0.1172	0.2868	0.0198	0.0588

Table 16: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	7391	3.552714e-15	1.059160e-09	0.000185	0.038082	0.8552	0.7102	0.6666	0.4616

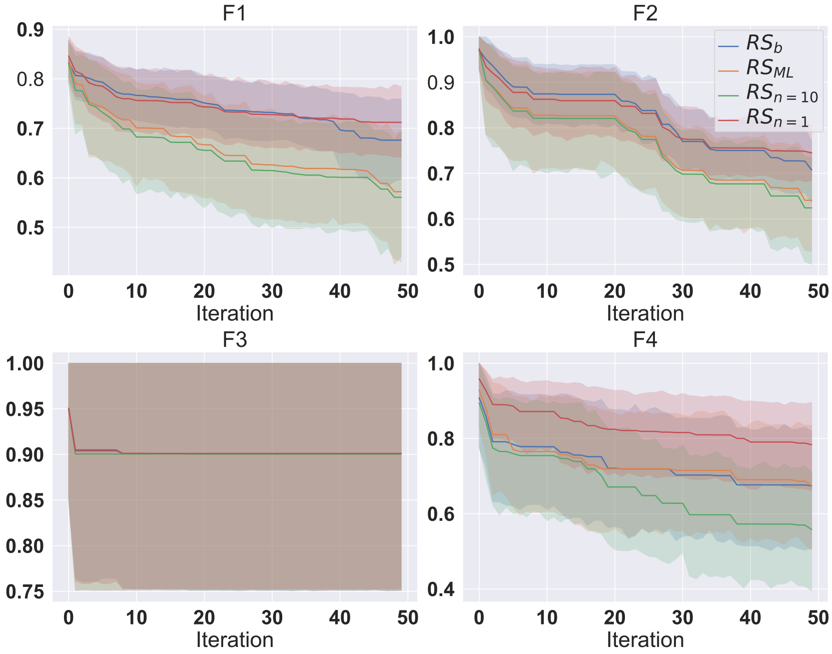


Figure 13: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PYLOT (Related to RQ2-2)

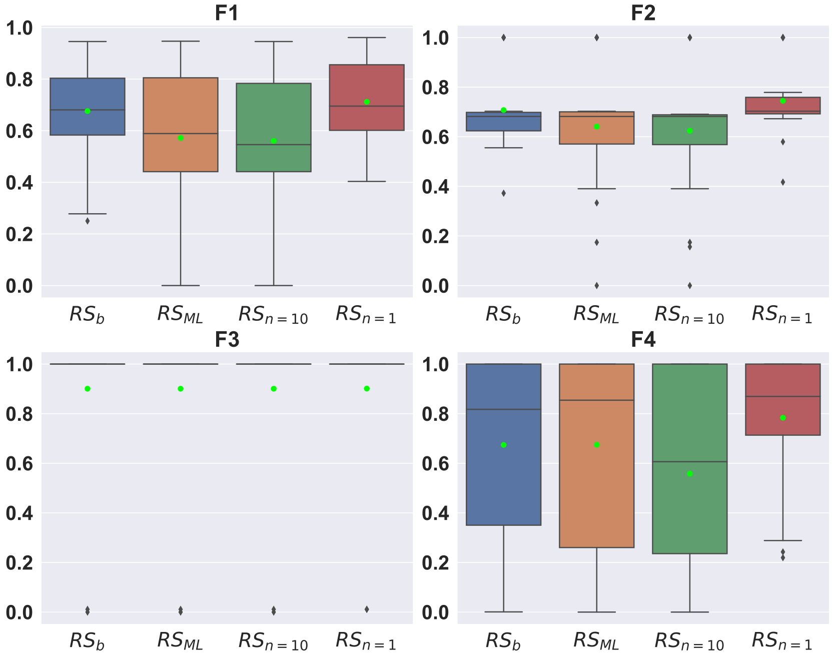


Figure 14: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PYLOT (Related to RQ2-2).

Threshold: 10%

Table 17: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
MLP	With delta fitnesses	0.99	0.95	0.97
MLP	With delta fitnesses wøweather	0.98	0.95	0.97
MLP	With delta fitnesses wøweather, blueprints	0.95	0.91	0.93
MLP	With delta fitnesses wøblueprints	0.96	0.90	0.93
Random Forest	With delta fitnesses wøweather	0.86	0.90	0.88
Random Forest	With delta fitnesses	0.90	0.86	0.88
Random Forest	With delta fitnesses wøblueprints	0.88	0.88	0.88
Random Forest	With delta fitnesses wøweather, blueprints	0.85	0.88	0.87
Decision Tree	With delta fitnesses	0.91	0.82	0.87
Decision Tree	With delta fitnesses wøweather	0.91	0.82	0.87
Decision Tree	With delta fitnesses wøblueprints	0.91	0.82	0.87
Decision Tree	With delta fitnesses wøweather, blueprints	0.91	0.82	0.87
SVM	With delta fitnesses wøweather	0.80	0.85	0.82
SVM	With delta fitnesses	0.74	0.84	0.79
Random Forest	With 1 set of fitnesses	0.75	0.77	0.76
Random Forest	With 1 set of fitnesses wøweather	0.75	0.77	0.76
SVM	With delta fitnesses wøweather, blueprints	0.78	0.70	0.74
Random Forest	With 1 set of fitnesses wøweather, blueprints	0.75	0.70	0.72
SVM	With delta fitnesses wøblueprints	0.73	0.70	0.71
Decision Tree	With 1 set of fitnesses	0.61	0.84	0.71
Decision Tree	With 1 set of fitnesses wøweather	0.61	0.84	0.71
Random Forest	With 1 set of fitnesses wøblueprints	0.72	0.67	0.70
MLP	With 1 set of fitnesses wøweather	0.56	0.91	0.69
SVM	With 1 set of fitnesses wøweather	0.51	0.88	0.65
MLP	With 1 set of fitnesses	0.54	0.74	0.63
MLP	With 1 set of fitnesses wøweather, blueprints	0.56	0.65	0.60
SVM	With 1 set of fitnesses wøweather, blueprints	0.53	0.70	0.60
Decision Tree	With 1 set of fitnesses wøweather, blueprints	0.51	0.70	0.59
Decision Tree	With 1 set of fitnesses wøblueprints	0.51	0.70	0.59
SVM	With 1 set of fitnesses	0.47	0.79	0.59
SVM	With 1 set of fitnesses wøblueprints	0.43	0.65	0.52
MLP	With 1 set of fitnesses wøblueprints	0.45	0.58	0.51

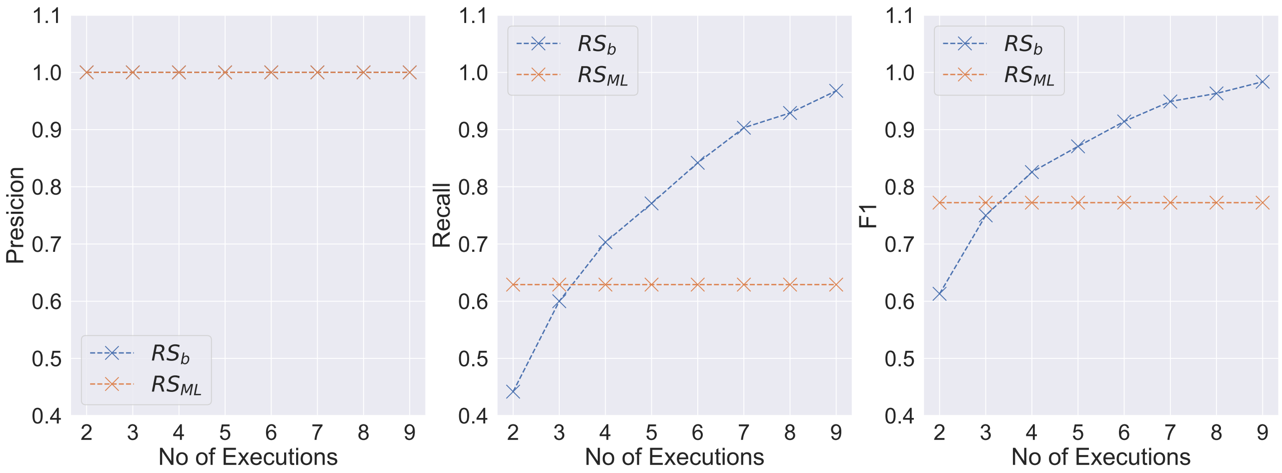


Figure 15: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 18: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.441935	0.612975	1.0	0.629032	0.772277
3	1.0	0.600000	0.750000	1.0	0.629032	0.772277
4	1.0	0.703226	0.825758	1.0	0.629032	0.772277
5	1.0	0.770968	0.870674	1.0	0.629032	0.772277
6	1.0	0.841935	0.914186	1.0	0.629032	0.772277
7	1.0	0.903226	0.949153	1.0	0.629032	0.772277
8	1.0	0.929032	0.963211	1.0	0.629032	0.772277
9	1.0	0.967742	0.983607	1.0	0.629032	0.772277

Table 19: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^12	F2 A^12	F3 A^12	F4 A^12
Baseline-Manual	7998	2.211422e-04	1.719142e-01	1.399457e-02	1.776357e-15	0.4172	0.496	0.2330	0.0336
fastFitness	3358	1.776357e-15	1.776357e-15	9.652696e-11	1.776357e-15	0.1228	0.288	0.0198	0.0328

Table 20: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^12	F2 A^12	F3 A^12	F4 A^12
Baseline-Manual	7998	3.552714e-15	1.058158e-09	0.000185	1.776357e-15	0.8132	0.7078	0.6666	0.6164

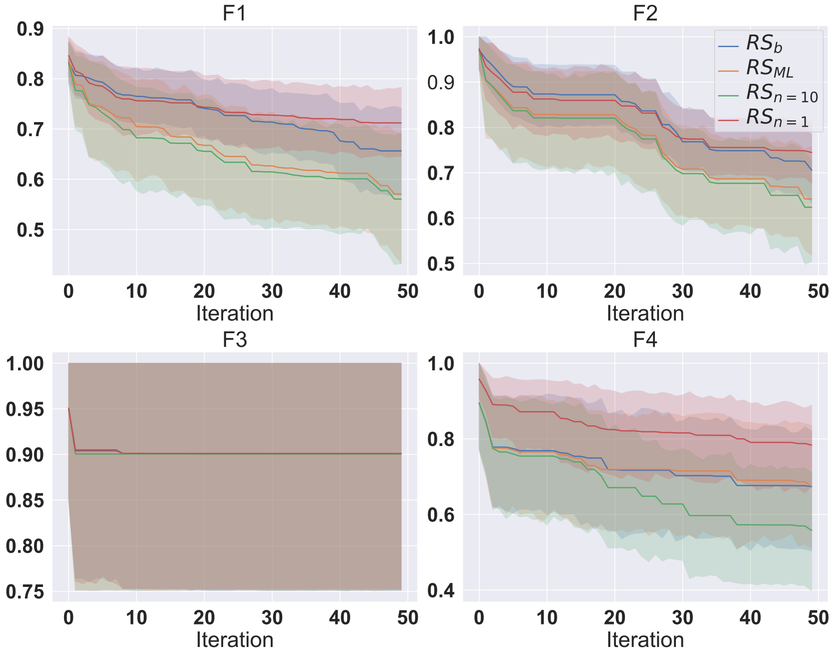


Figure 16: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PYLOT (Related to RQ2-2)

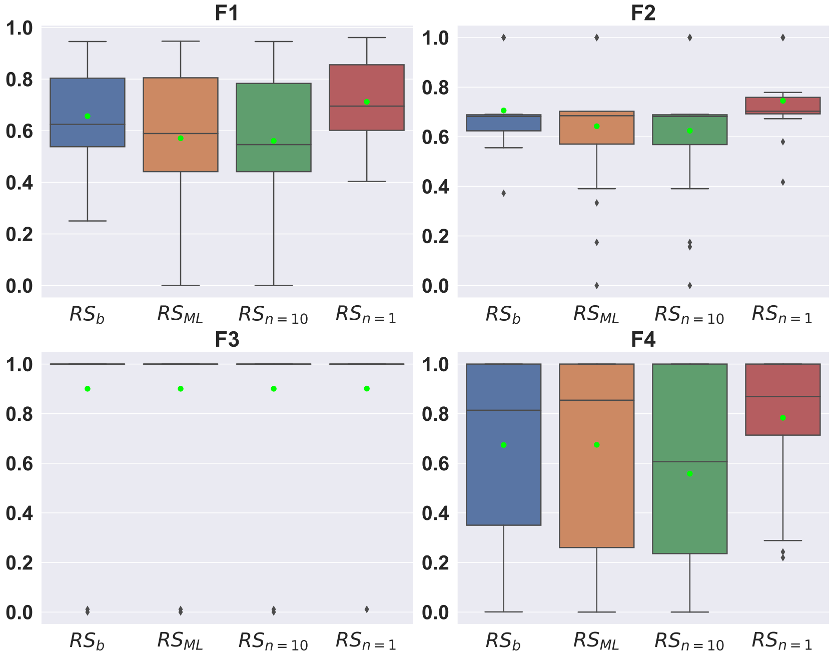


Figure 17: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PYLOT (Related to RQ2-2).

Threshold: 20%

Table 21: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
MLP	With delta fitnesses wøweather	0.99	0.93	0.96
MLP	With delta fitnesses	0.98	0.93	0.96
MLP	With delta fitnesses wøblueprints	1.00	0.89	0.94
MLP	With delta fitnesses wøweather, blueprints	0.93	0.90	0.92
Random Forest	With delta fitnesses wøweather	0.84	0.87	0.86
Random Forest	With delta fitnesses wøweather, blueprints	0.85	0.86	0.86
Random Forest	With delta fitnesses wøblueprints	0.84	0.86	0.85
Random Forest	With delta fitnesses	0.81	0.87	0.84
Decision Tree	With delta fitnesses	0.86	0.82	0.84
Decision Tree	With delta fitnesses wøblueprints	0.86	0.82	0.84
SVM	With delta fitnesses wøweather	0.77	0.92	0.84
Decision Tree	With delta fitnesses wøweather	0.86	0.80	0.83
Decision Tree	With delta fitnesses wøweather, blueprints	0.86	0.80	0.83
SVM	With delta fitnesses wøweather, blueprints	0.85	0.77	0.81
SVM	With delta fitnesses	0.73	0.83	0.78
SVM	With delta fitnesses wøblueprints	0.67	0.75	0.71
MLP	With 1 set of fitnesses wøweather	0.59	0.81	0.68
Random Forest	With 1 set of fitnesses wøweather, blueprints	0.61	0.69	0.65
Random Forest	With 1 set of fitnesses	0.56	0.69	0.62
Random Forest	With 1 set of fitnesses wøblueprints	0.55	0.66	0.60
Random Forest	With 1 set of fitnesses wøweather	0.52	0.69	0.59
SVM	With 1 set of fitnesses wøweather	0.46	0.81	0.58
MLP	With 1 set of fitnesses	0.47	0.72	0.57
MLP	With 1 set of fitnesses wøweather, blueprints	0.51	0.59	0.55
Decision Tree	With 1 set of fitnesses wøblueprints	0.46	0.69	0.55
Decision Tree	With 1 set of fitnesses wøweather, blueprints	0.44	0.69	0.54
MLP	With 1 set of fitnesses wøblueprints	0.44	0.66	0.53
Decision Tree	With 1 set of fitnesses	0.41	0.72	0.52
Decision Tree	With 1 set of fitnesses wøweather	0.41	0.69	0.51
SVM	With 1 set of fitnesses	0.38	0.75	0.51
SVM	With 1 set of fitnesses wøweather, blueprints	0.38	0.72	0.50
SVM	With 1 set of fitnesses wøblueprints	0.34	0.72	0.46

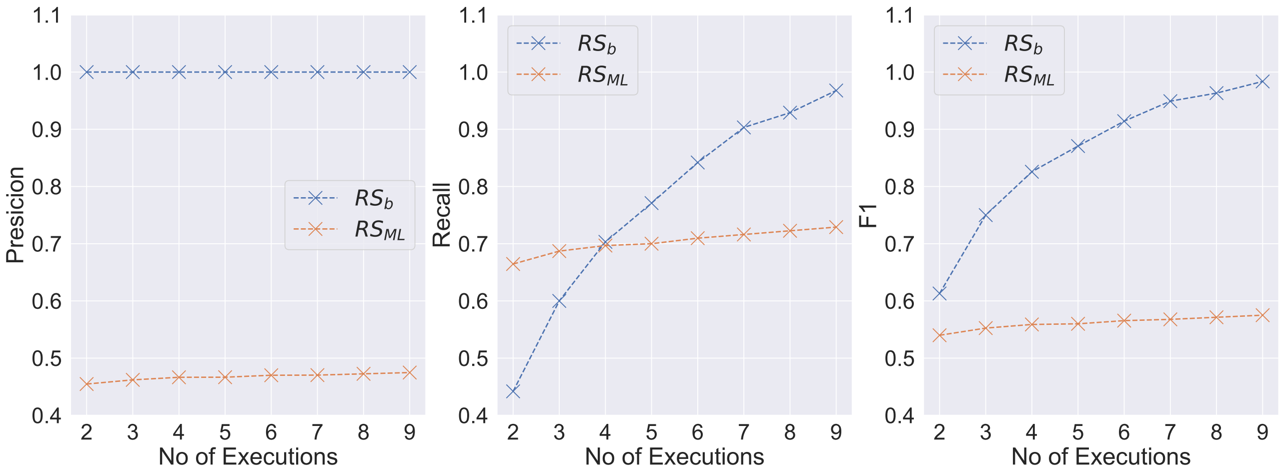


Figure 18: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 22: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.441935	0.612975	0.469697	0.500000	0.484375
3	1.0	0.600000	0.750000	0.488506	0.548387	0.516717
4	1.0	0.703226	0.825758	0.502793	0.580645	0.538922
5	1.0	0.770968	0.870674	0.509537	0.603226	0.552437
6	1.0	0.841935	0.914186	0.512064	0.616129	0.559297
7	1.0	0.903226	0.949153	0.517241	0.629032	0.567686
8	1.0	0.929032	0.963211	0.523560	0.645161	0.578035
9	1.0	0.967742	0.983607	0.527273	0.654839	0.584173

Table 23: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	8278	4.218847e-12	3.089085e-12	0.013995	1.776357e-15	0.3388	0.3872	0.233	0.0336
fastFitness	3078	1.776357e-15	1.776357e-15	0.013995	1.776357e-15	0.1396	0.3008	0.233	0.1896

Table 24: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	8278	2.486900e-14	1.065945e-09	0.0	1.776357e-15	0.692	0.6486	0.5	0.238

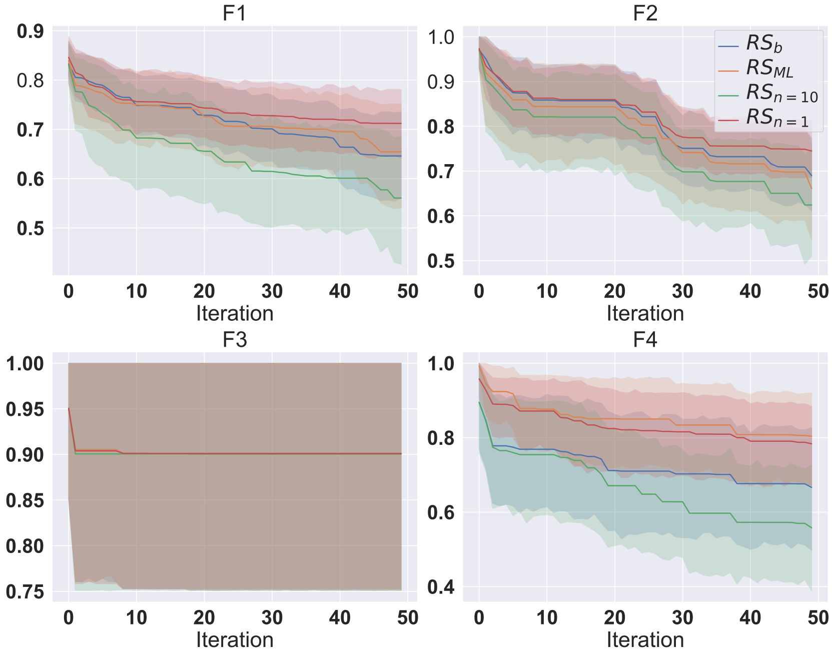


Figure 19: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PYLOT (Related to RQ2-2)

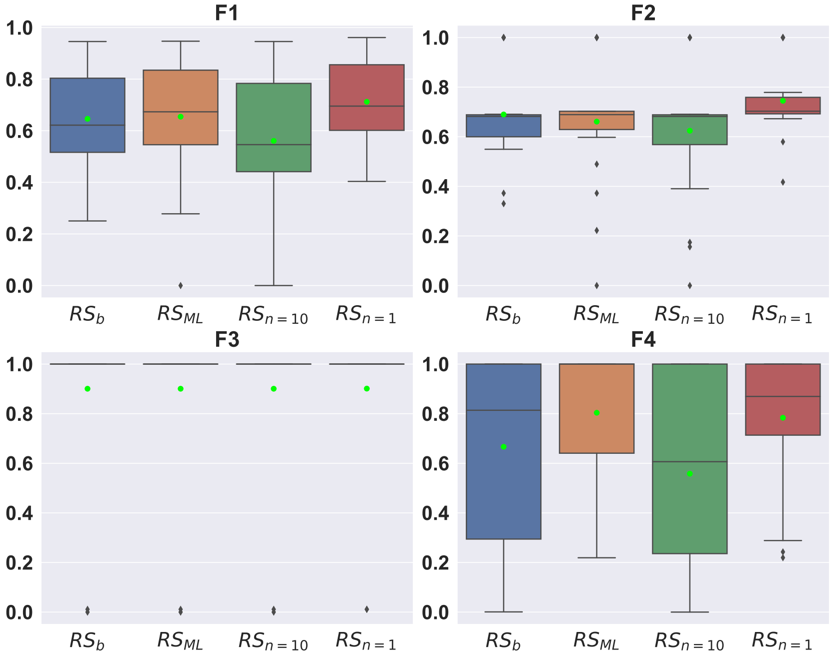


Figure 20: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of PYLOT (Related to RQ2-2).

BeamNG

Threshold: 5%

Table 25: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
Random Forest	With delta fitnesses	1.00	0.97	0.98
Random Forest	With delta fitnesses wøweather, blueprints	1.00	0.97	0.98
Random Forest	With delta fitnesses wøblueprints	1.00	0.96	0.98
Random Forest	With delta fitnesses wøweather	1.00	0.96	0.98
MLP	With delta fitnesses wøblueprints	1.00	0.96	0.98
MLP	With delta fitnesses	0.98	0.98	0.98
Decision Tree	With delta fitnesses	1.00	0.96	0.98
Decision Tree	With delta fitnesses wøweather	1.00	0.96	0.98
Decision Tree	With delta fitnesses wøblueprints	1.00	0.96	0.98
Decision Tree	With delta fitnesses wøweather, blueprints	1.00	0.96	0.98
MLP	With delta fitnesses wøweather	0.99	0.97	0.98
MLP	With delta fitnesses wøweather, blueprints	0.99	0.96	0.98
Random Forest	With 1 set of fitnesses wøblueprints	0.99	0.93	0.96
Decision Tree	With 1 set of fitnesses wøweather	0.97	0.95	0.96
Decision Tree	With 1 set of fitnesses wøblueprints	0.97	0.94	0.96
Random Forest	With 1 set of fitnesses wøweather	0.99	0.93	0.95
Random Forest	With 1 set of fitnesses wøweather, blueprints	0.99	0.92	0.95
Random Forest	With 1 set of fitnesses	0.99	0.92	0.95
Decision Tree	With 1 set of fitnesses wøweather, blueprints	0.96	0.95	0.95
Decision Tree	With 1 set of fitnesses	0.97	0.93	0.95
SVM	With delta fitnesses	0.98	0.92	0.95
SVM	With delta fitnesses wøweather	0.98	0.92	0.95
SVM	With delta fitnesses wøblueprints	0.98	0.92	0.95
SVM	With delta fitnesses wøweather, blueprints	0.98	0.92	0.95
MLP	With 1 set of fitnesses	0.97	0.90	0.94
MLP	With 1 set of fitnesses wøweather	0.98	0.88	0.93
MLP	With 1 set of fitnesses wøblueprints	0.89	0.96	0.92
MLP	With 1 set of fitnesses wøweather, blueprints	0.98	0.87	0.92
SVM	With 1 set of fitnesses wøweather, blueprints	0.98	0.83	0.90
SVM	With 1 set of fitnesses wøblueprints	0.98	0.83	0.90
SVM	With 1 set of fitnesses wøweather	0.98	0.83	0.90
SVM	With 1 set of fitnesses	0.98	0.83	0.90

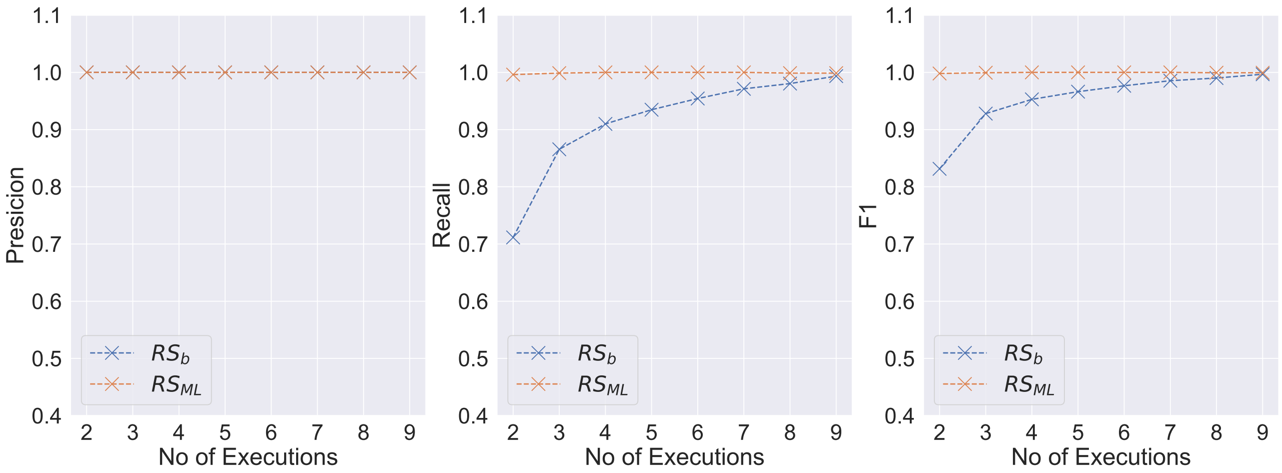


Figure 21: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 26: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.711488	0.831426	1.0	0.993473	0.996726
3	1.0	0.865535	0.927922	1.0	0.998695	0.999347
4	1.0	0.909922	0.952837	1.0	1.000000	1.000000
5	1.0	0.934726	0.966262	1.0	1.000000	1.000000
6	1.0	0.954308	0.976620	1.0	1.000000	1.000000
7	1.0	0.971279	0.985430	1.0	1.000000	1.000000
8	1.0	0.980418	0.990112	1.0	1.000000	1.000000
9	1.0	0.993473	0.996726	1.0	1.000000	1.000000

Table 27: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	4153	1.776357e-15	1.776357e-15	1.776357e-15	1.776357e-15	0.0524	0.0	0.0	0.3168
fastFitness	7229	1.776357e-15	1.776357e-15	1.776357e-15	1.776357e-15	0.0124	0.0	0.0	0.2076

Table 28: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	4153	0.211485	1.776357e-15	1.776357e-15	1.776357e-15	0.6796	0.9312	0.6072	0.6708

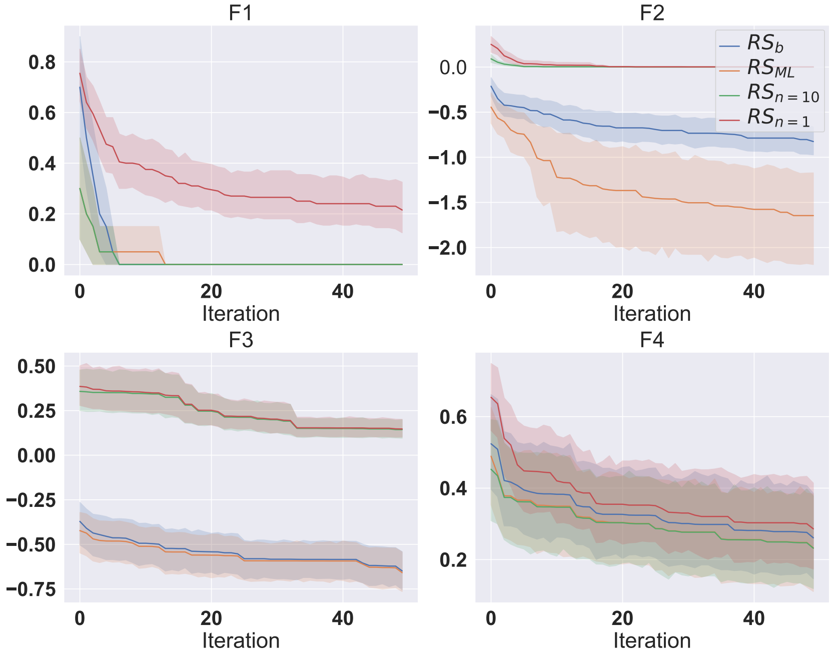


Figure 22: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of BeamNG (Related to RQ2-2)

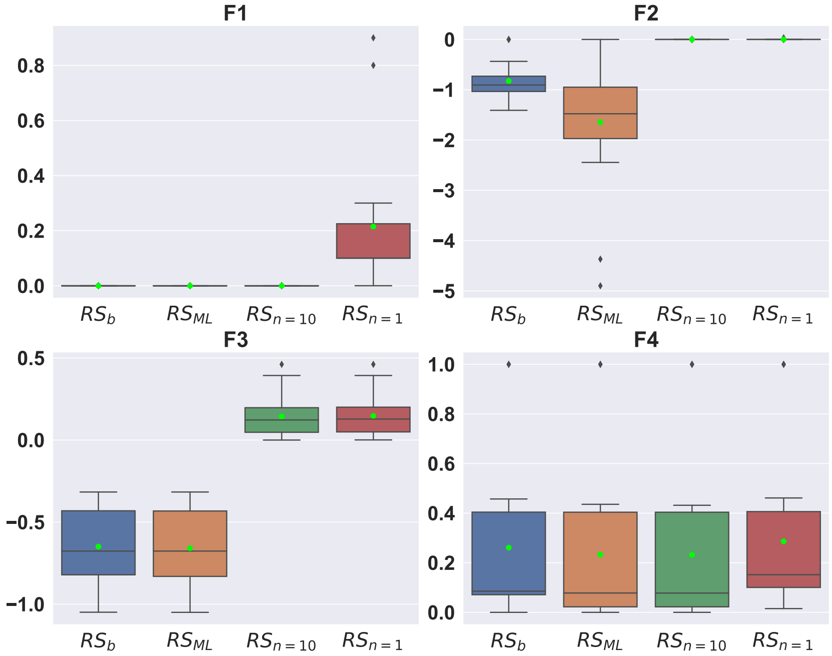


Figure 23: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of BeamNG (Related to RQ2-2).

Threshold: 10%

Table 29: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
Random Forest	With delta fitnesses wøweather	1.00	0.97	0.99
Random Forest	With delta fitnesses	1.00	0.97	0.99
Random Forest	With delta fitnesses wøweather, blueprints	1.00	0.97	0.98
Random Forest	With delta fitnesses wøblueprints	1.00	0.97	0.98
Decision Tree	With delta fitnesses	1.00	0.96	0.98
Decision Tree	With delta fitnesses wøweather	1.00	0.96	0.98
Decision Tree	With delta fitnesses wøblueprints	1.00	0.96	0.98
Decision Tree	With delta fitnesses wøweather, blueprints	1.00	0.96	0.98
Decision Tree	With 1 set of fitnesses wøweather, blueprints	1.00	0.95	0.97
Random Forest	With 1 set of fitnesses	1.00	0.95	0.97
Random Forest	With 1 set of fitnesses wøblueprints	1.00	0.95	0.97
Random Forest	With 1 set of fitnesses wøweather, blueprints	1.00	0.95	0.97
MLP	With delta fitnesses wøweather	1.00	0.95	0.97
MLP	With delta fitnesses wøweather, blueprints	1.00	0.95	0.97
MLP	With delta fitnesses wøblueprints	1.00	0.95	0.97
MLP	With delta fitnesses	0.99	0.95	0.97
Decision Tree	With 1 set of fitnesses	1.00	0.94	0.97
Decision Tree	With 1 set of fitnesses wøweather	1.00	0.94	0.97
Decision Tree	With 1 set of fitnesses wøblueprints	1.00	0.94	0.97
Random Forest	With 1 set of fitnesses wøweather	1.00	0.94	0.97
SVM	With delta fitnesses	1.00	0.91	0.95
SVM	With delta fitnesses wøweather	1.00	0.91	0.95
SVM	With delta fitnesses wøblueprints	1.00	0.91	0.95
SVM	With delta fitnesses wøweather, blueprints	1.00	0.91	0.95
MLP	With 1 set of fitnesses wøblueprints	0.95	0.90	0.92
MLP	With 1 set of fitnesses wøweather	0.92	0.90	0.91
MLP	With 1 set of fitnesses wøweather, blueprints	1.00	0.82	0.90
SVM	With 1 set of fitnesses wøweather	1.00	0.82	0.90
SVM	With 1 set of fitnesses wøblueprints	1.00	0.82	0.90
SVM	With 1 set of fitnesses wøweather, blueprints	1.00	0.82	0.90
MLP	With 1 set of fitnesses	1.00	0.82	0.90
SVM	With 1 set of fitnesses	1.00	0.82	0.90

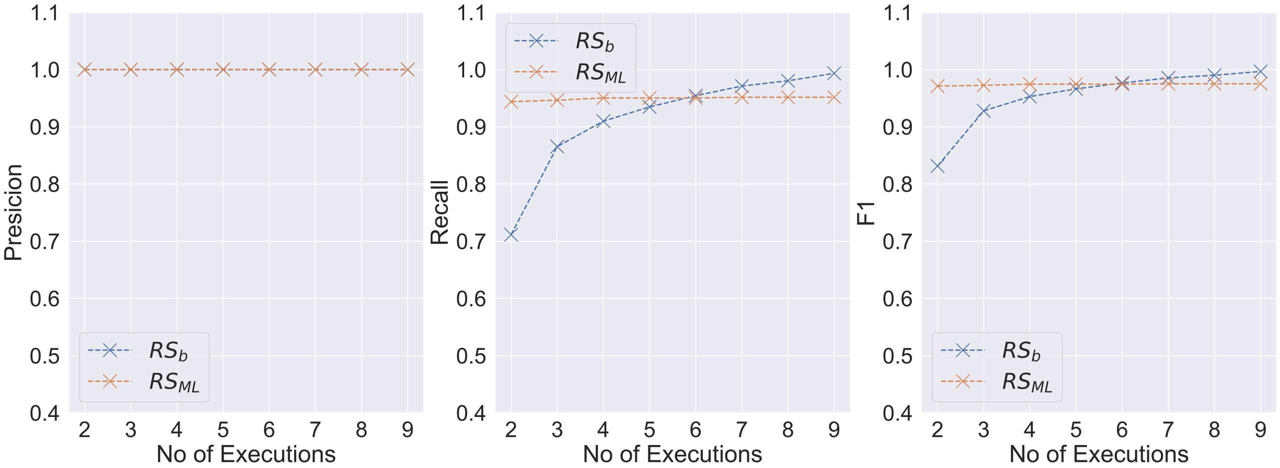


Figure 24: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 30: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.711488	0.831426	1.0	0.946475	0.972502
3	1.0	0.865535	0.927922	1.0	0.950392	0.974565
4	1.0	0.909922	0.952837	1.0	0.950392	0.974565
5	1.0	0.934726	0.966262	1.0	0.949086	0.973878
6	1.0	0.954308	0.976620	1.0	0.951697	0.975251
7	1.0	0.971279	0.985430	1.0	0.951697	0.975251
8	1.0	0.980418	0.990112	1.0	0.951697	0.975251
9	1.0	0.993473	0.996726	1.0	0.951697	0.975251

Table 31: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	4379	1.776357e-15	1.776357e-15	1.776357e-15	1.776357e-15	0.0492	0.0	0.0	0.3168
fastFitness	7023	1.776357e-15	1.776357e-15	1.776357e-15	1.776357e-15	0.0180	0.0	0.0	0.2116

Table 32: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	4379	3.403677e-11	1.776357e-15	1.776357e-15	1.776357e-15	0.8828	0.9264	0.6072	0.6532

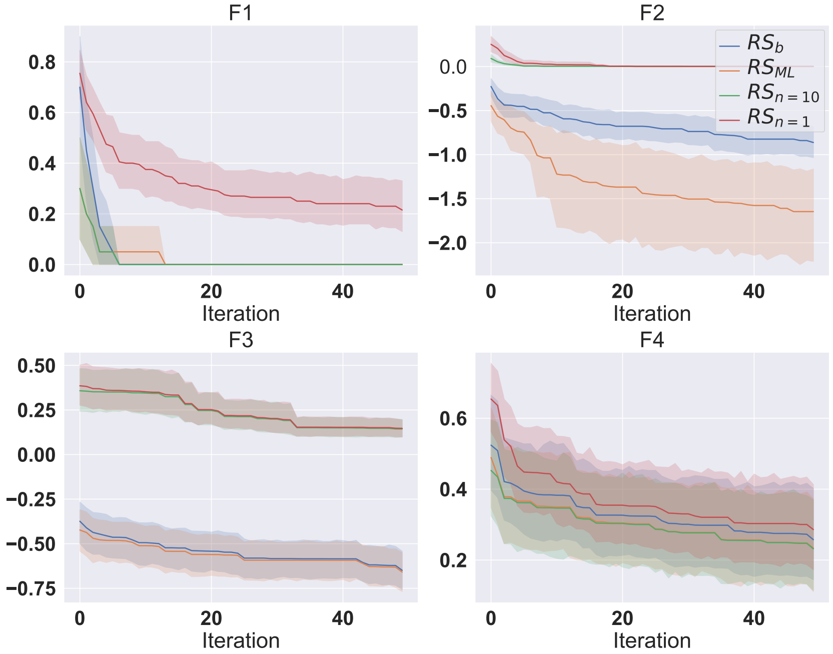


Figure 25: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of BeamNG (Related to RQ2-2)

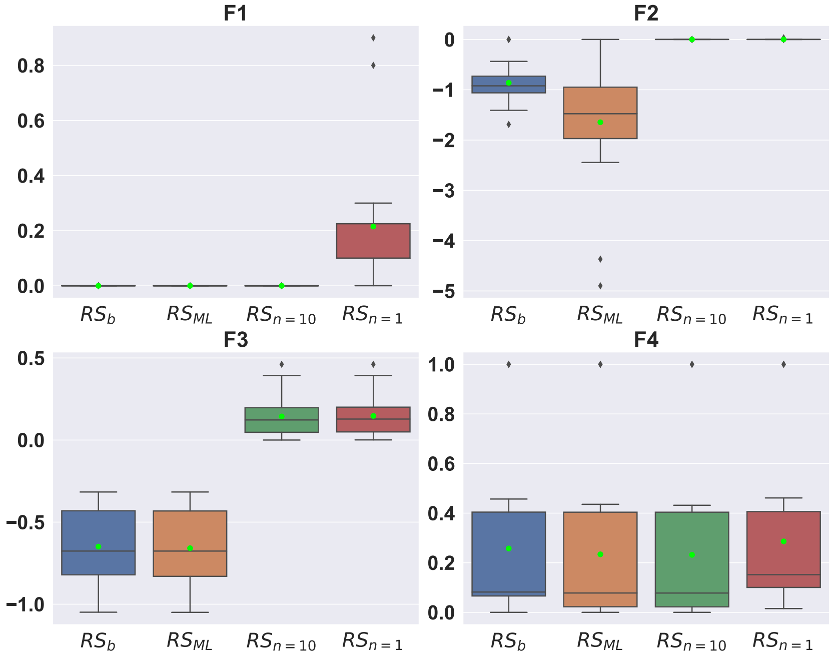


Figure 26: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of BeamNG (Related to RQ2-2).

Threshold: 20%

Table 33: Models trained for STEC and MTEC classifiers (Related to RQ2-1)

Method	Input	Precision	Recall	F1
Random Forest	With delta fitnesses wøblueprints	1.00	0.98	0.99
Random Forest	With delta fitnesses wøweather	1.00	0.98	0.99
Decision Tree	With delta fitnesses	0.99	0.99	0.99
Decision Tree	With delta fitnesses wøweather	0.99	0.99	0.99
Decision Tree	With delta fitnesses wøblueprints	0.99	0.99	0.99
Decision Tree	With delta fitnesses wøweather, blueprints	0.99	0.99	0.99
Random Forest	With delta fitnesses wøweather, blueprints	1.00	0.98	0.99
Random Forest	With delta fitnesses	1.00	0.97	0.99
MLP	With delta fitnesses wøweather, blueprints	1.00	0.98	0.99
MLP	With delta fitnesses wøblueprints	1.00	0.97	0.98
MLP	With delta fitnesses	0.99	0.97	0.98
MLP	With delta fitnesses wøweather	1.00	0.97	0.98
Random Forest	With 1 set of fitnesses wøweather	1.00	0.96	0.98
Random Forest	With 1 set of fitnesses	1.00	0.96	0.98
Random Forest	With 1 set of fitnesses wøblueprints	1.00	0.96	0.98
Random Forest	With 1 set of fitnesses wøweather, blueprints	1.00	0.96	0.98
Decision Tree	With 1 set of fitnesses	1.00	0.94	0.97
Decision Tree	With 1 set of fitnesses wøweather	1.00	0.94	0.97
Decision Tree	With 1 set of fitnesses wøweather, blueprints	1.00	0.94	0.97
Decision Tree	With 1 set of fitnesses wøblueprints	1.00	0.94	0.97
SVM	With delta fitnesses	1.00	0.93	0.96
SVM	With delta fitnesses wøweather	1.00	0.93	0.96
SVM	With delta fitnesses wøblueprints	1.00	0.93	0.96
SVM	With delta fitnesses wøweather, blueprints	1.00	0.93	0.96
MLP	With 1 set of fitnesses wøweather, blueprints	0.95	0.88	0.92
MLP	With 1 set of fitnesses wøweather	0.91	0.92	0.92
MLP	With 1 set of fitnesses wøblueprints	0.99	0.85	0.91
MLP	With 1 set of fitnesses	0.98	0.85	0.91
SVM	With 1 set of fitnesses wøweather	1.00	0.83	0.91
SVM	With 1 set of fitnesses wøblueprints	1.00	0.83	0.91
SVM	With 1 set of fitnesses wøweather, blueprints	1.00	0.83	0.91
SVM	With 1 set of fitnesses	1.00	0.83	0.91

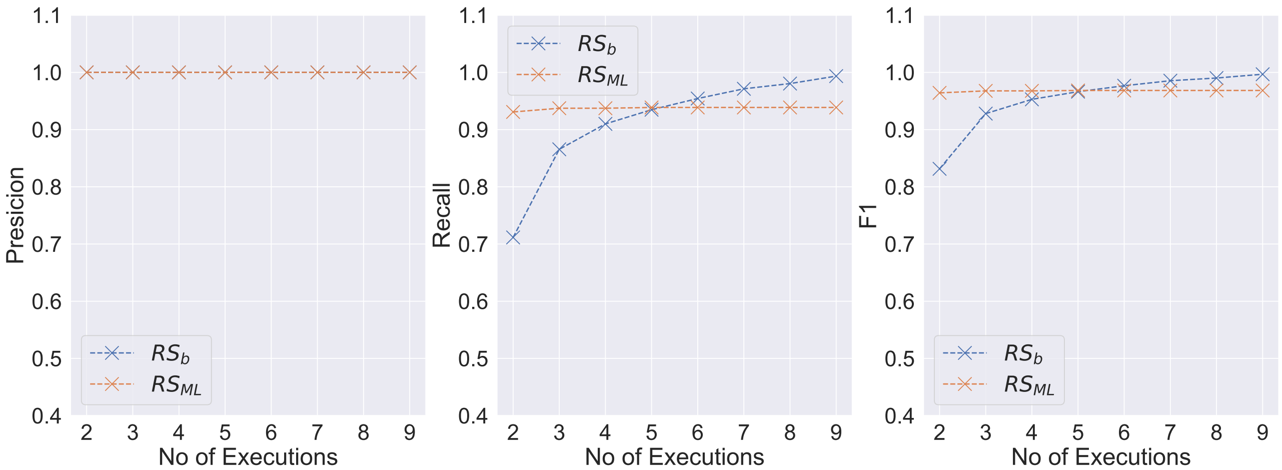


Figure 27: Precision, Recall, and F1-Score of RS_b and RS_{ML} based on inputs using different re-executions of each test input(Related to RQ2-1)

Table 34: Comparing the best MTEC classifiers with our non-ML-based baseline for different ADS test setups (Related to RQ2-1)

Timestep	Baseline Presicion	Baseline Recall	Baseline F1	Model Presicion	Model Recall	Model F1
2	1.0	0.711488	0.831426	1.0	0.932115	0.964865
3	1.0	0.865535	0.927922	1.0	0.937337	0.967655
4	1.0	0.909922	0.952837	1.0	0.937337	0.967655
5	1.0	0.934726	0.966262	1.0	0.938642	0.968350
6	1.0	0.954308	0.976620	1.0	0.938642	0.968350
7	1.0	0.971279	0.985430	1.0	0.938642	0.968350
8	1.0	0.980418	0.990112	1.0	0.938642	0.968350
9	1.0	0.993473	0.996726	1.0	0.938642	0.968350

Table 35: Statistical tests between RS_b and RS_{ML} with $RS_{n=10}$ (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	4530	1.776357e-15	1.776357e-15	1.776357e-15	1.776357e-15	0.0460	0.0	0.0	0.3132
fastFitness	6957	1.776357e-15	1.776357e-15	1.776357e-15	1.776357e-15	0.0124	0.0	0.0	0.2276

Table 36: Statistical tests between RS_b and RS_{ML} (Related to RQ2-2)

Method	Simulations	F1 p-value	F2 p-value	F3 p-value	F4 p-value	F1 A^{1}2	F2 A^{1}2	F3 A^{1}2	F4 A^{1}2
Baseline-Manual	4530	0.261209	1.776357e-15	1.776357e-15	1.776357e-15	0.6784	0.8812	0.604	0.6392

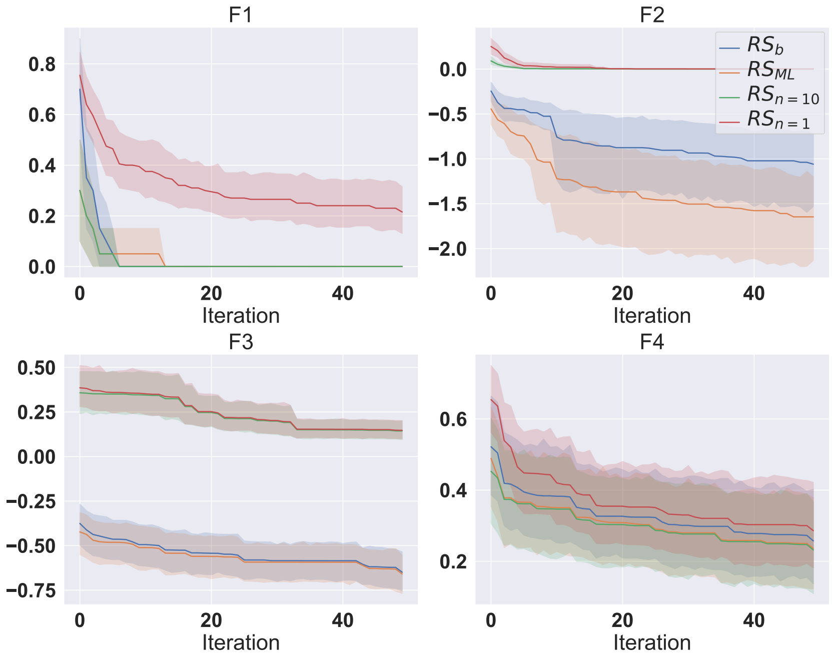


Figure 28: The average and 95% interval of the best fitness values obtained by 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of BeamNG (Related to RQ2-2)

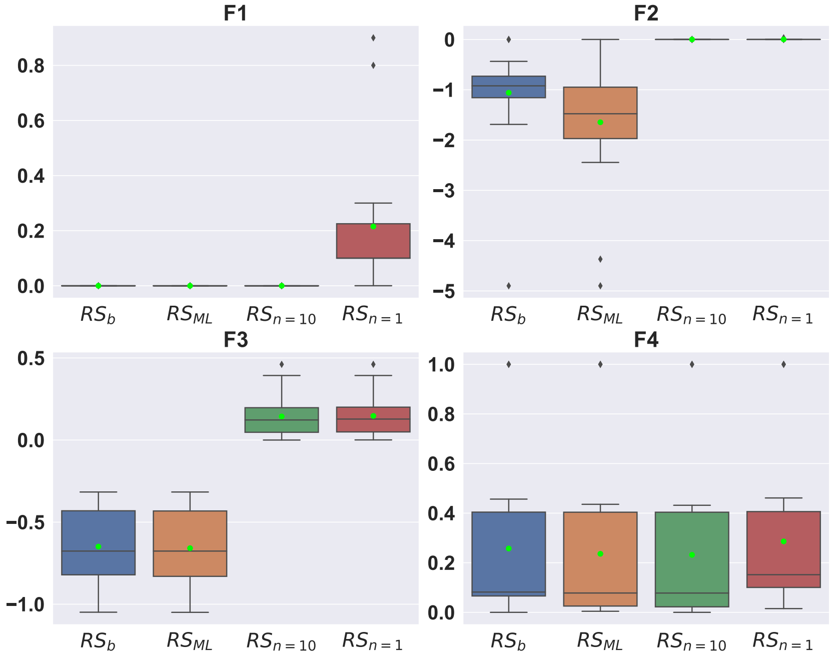


Figure 29: Distributions and averages of the best fitness values obtained from 20 runs of RS_b, RS_{ML}, $RS_{n=1}$, and $RS_{n=10}$ over 50 iterations for four fitness functions of BeamNG (Related to RQ2-2).