The Impact of Active Learning on Availability Data Poisoning for Android Malware Classifiers

Shae McFadden[1,2,3], Zeliang Kan[1,3], Lorenzo Cavallaro[3], Fabio Pierazzi[3,1]
[1]King's College London, [2]The Alan Turing Institute, [3]University College London

ARTMAN Workshop, Waikiki, December 9th 2024

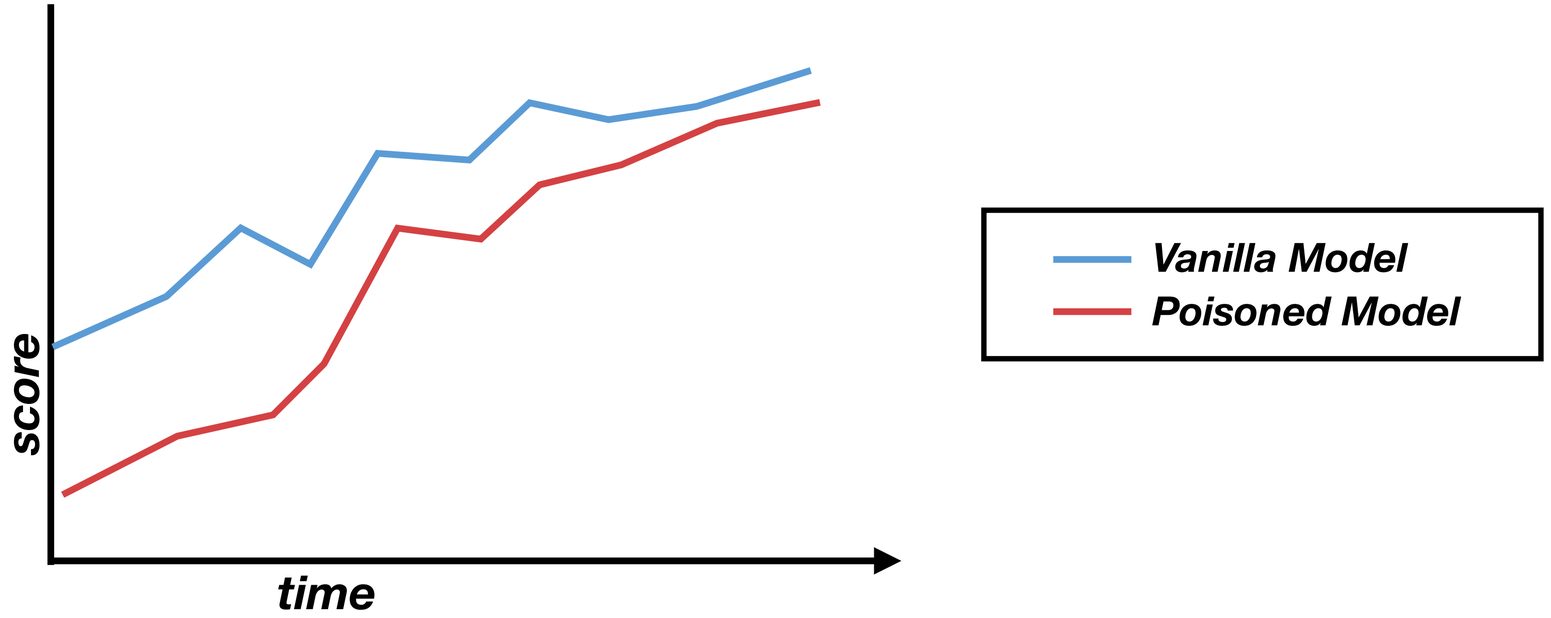
The Alan Turing Institute

Incorrect or poisoned data reduces out of the box classifier performance

However, models are frequently retrained with new data

Therefore, what is the over time impact of poisoning?

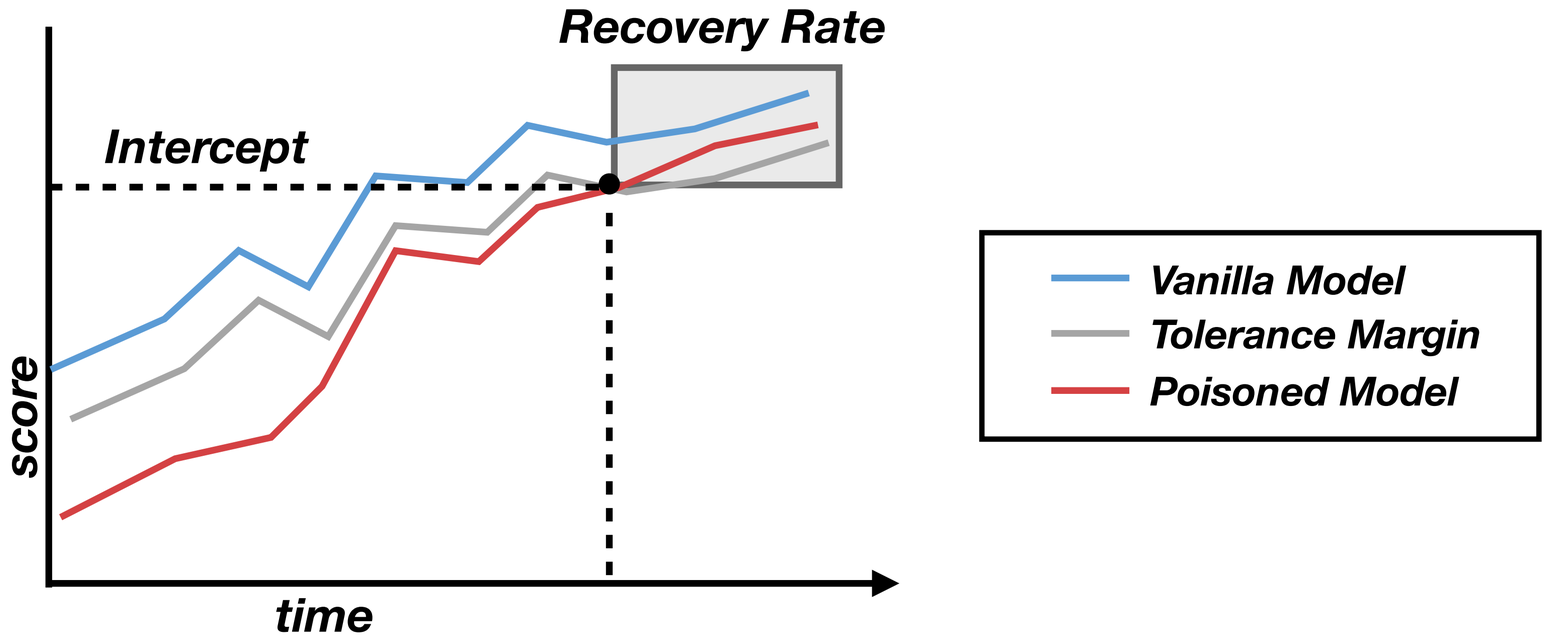
Passive Recovery



- ***Recovery***: converging the performance of a poisoned model with that of the *hypothetical model*, which was never poisoned.

- ***Passive Recovery*** refers to recovery achieved as a byproduct of an approach designed for another *purpose*.

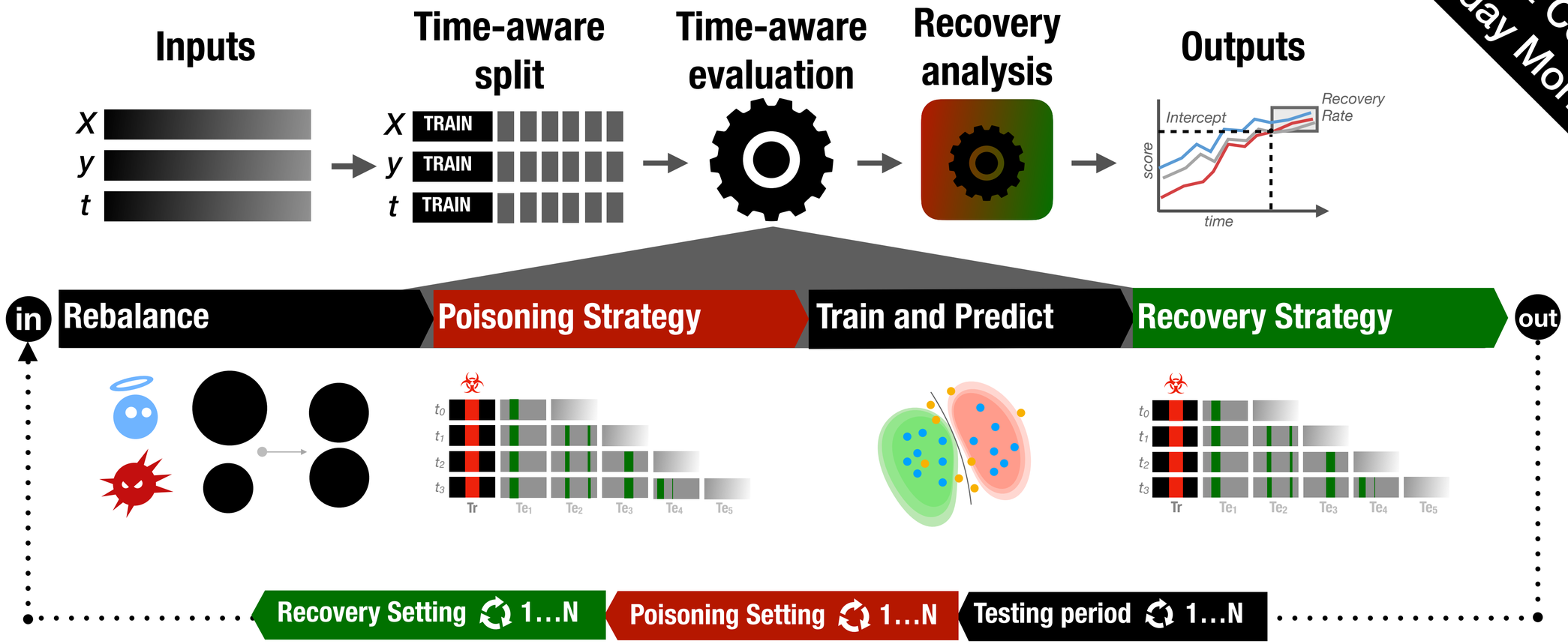
Measuring Passive Recovery



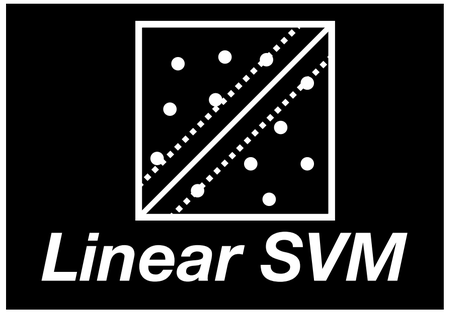
- ***Tolerance Margin*** defines the strictness of recovery
- ***Intercept*** measures the speed of recovery
- ***Recovery Rate*** measures the stability of recovery

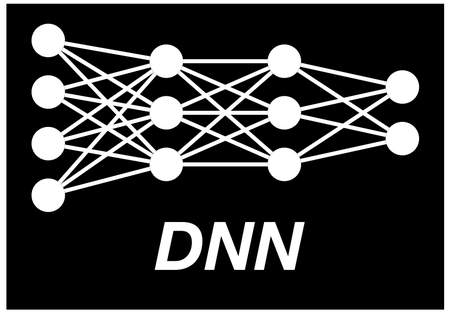
RPAL: Passive Recovery Evaluation Fra rk

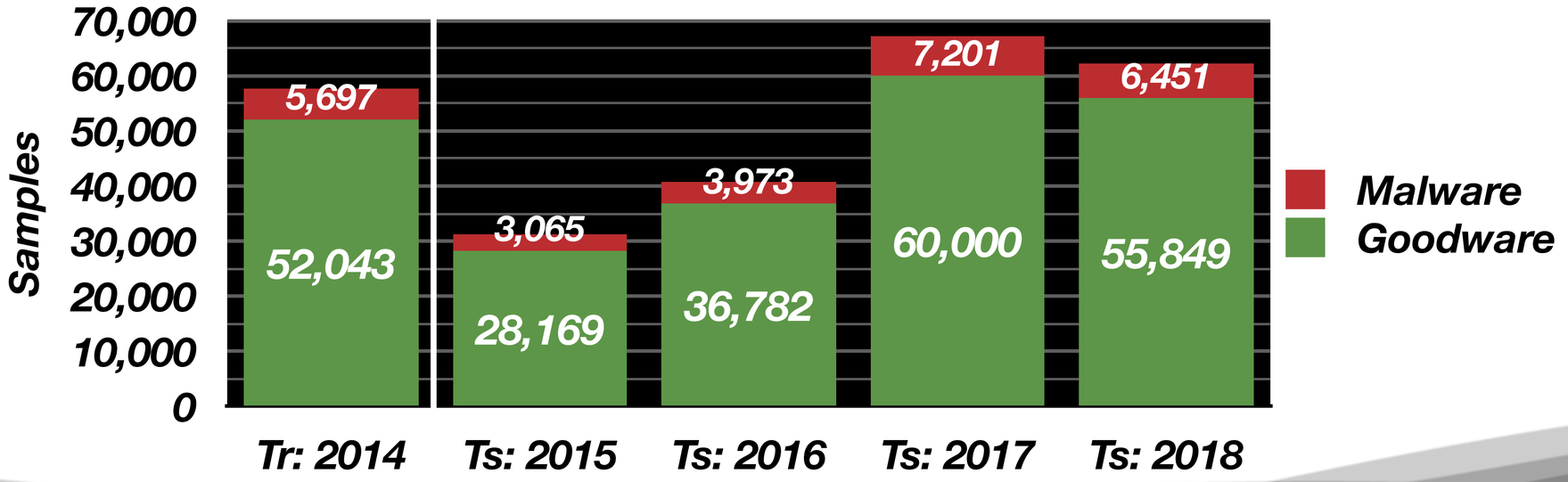
Tesseract @ Cybersecurity Artifact Impact Competition Thursday Morning



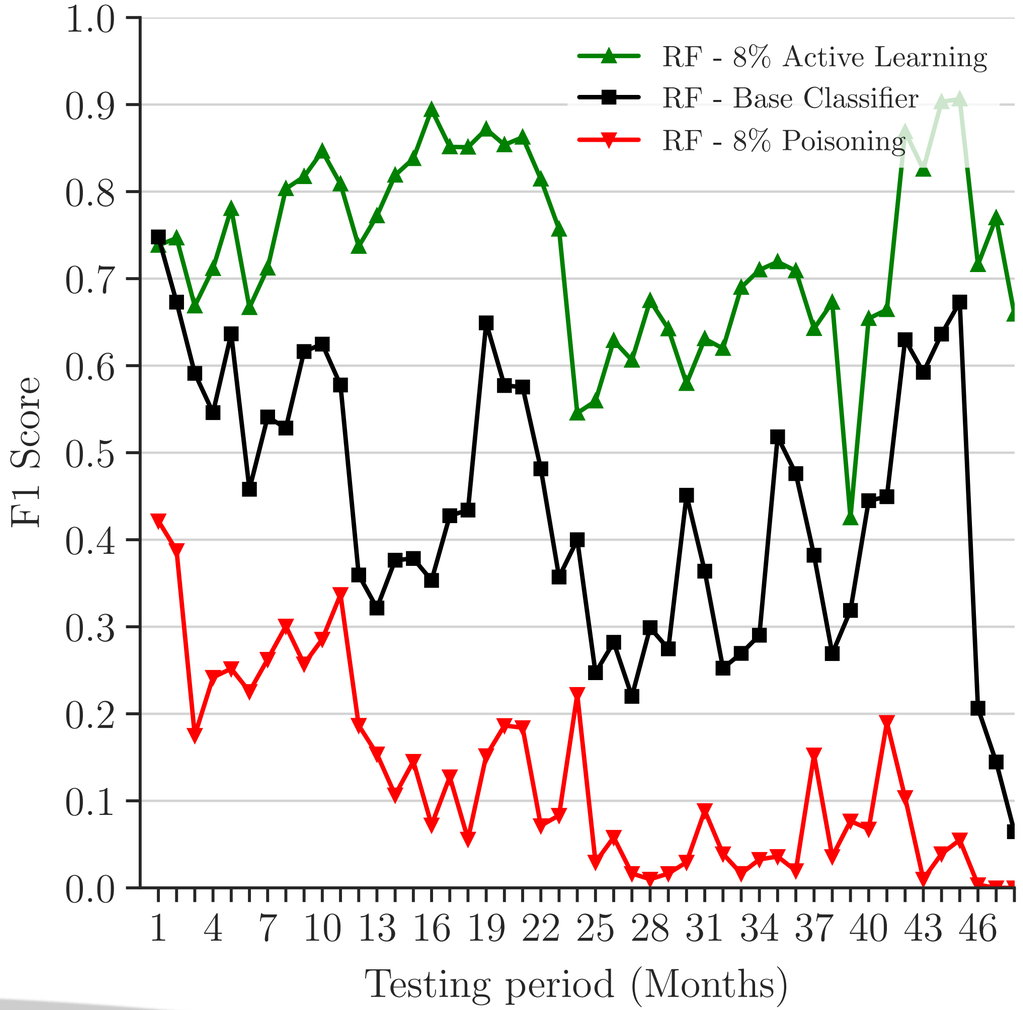
Experimental Settings





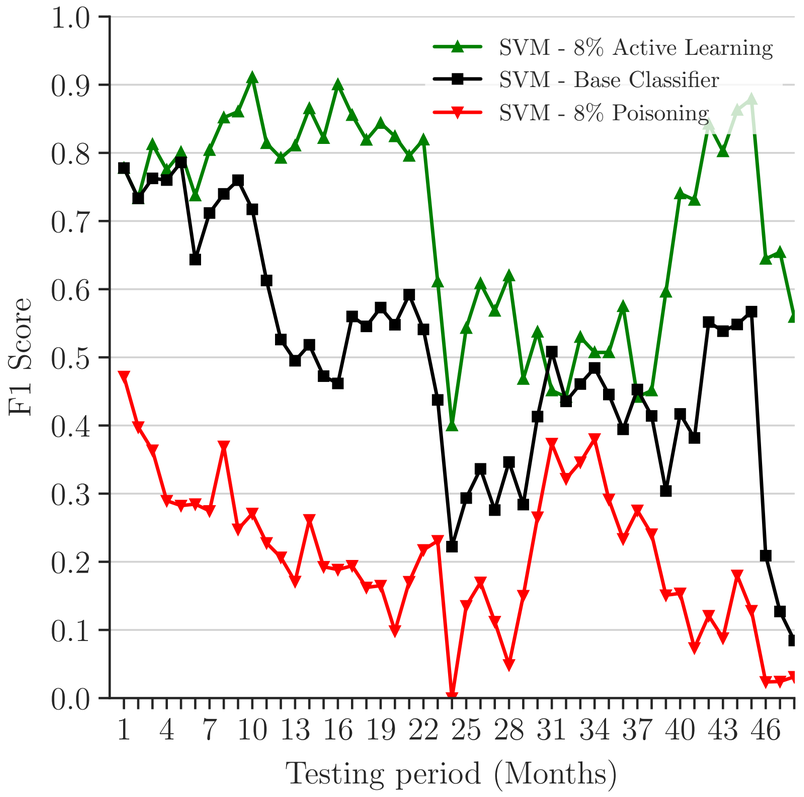


Recovery & Poisoning Mechanism (1/2)

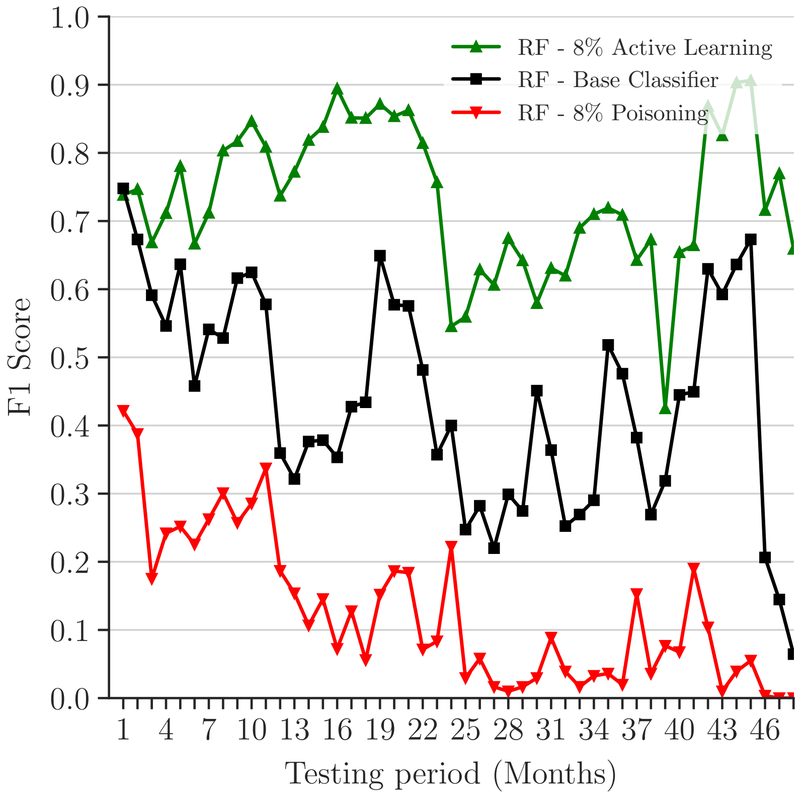


- ***Active learning***
  - ***Uncertainty Sampling*** selects the least certain samples for retraining
- ***Availability Data Poisoning***
  - ***Label-Flip Poisoning*** modifies a portion of training labels

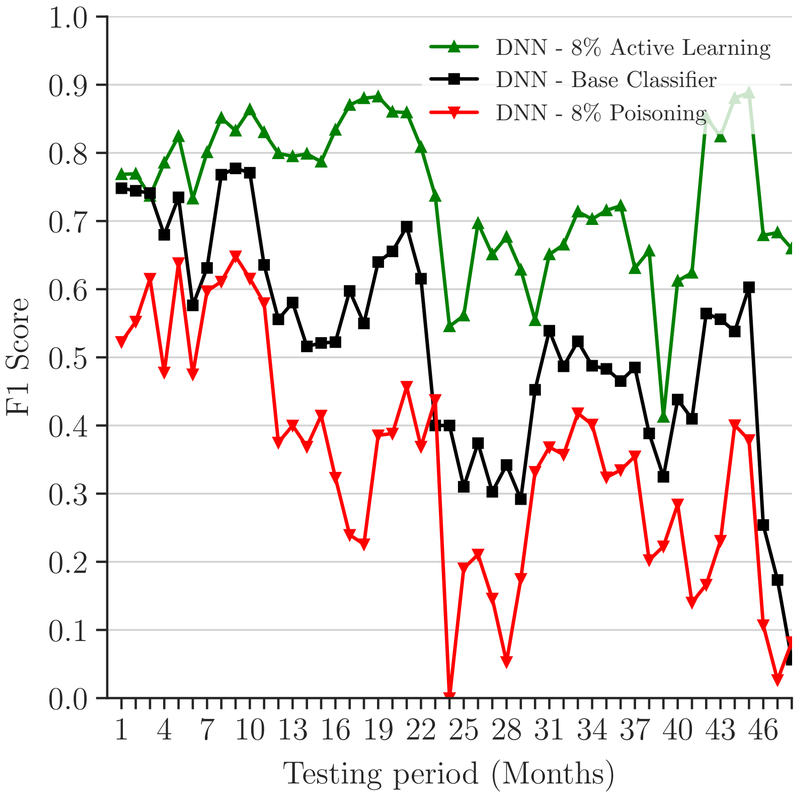
Recovery & Poisoning Mechanism (2/2)



SVM



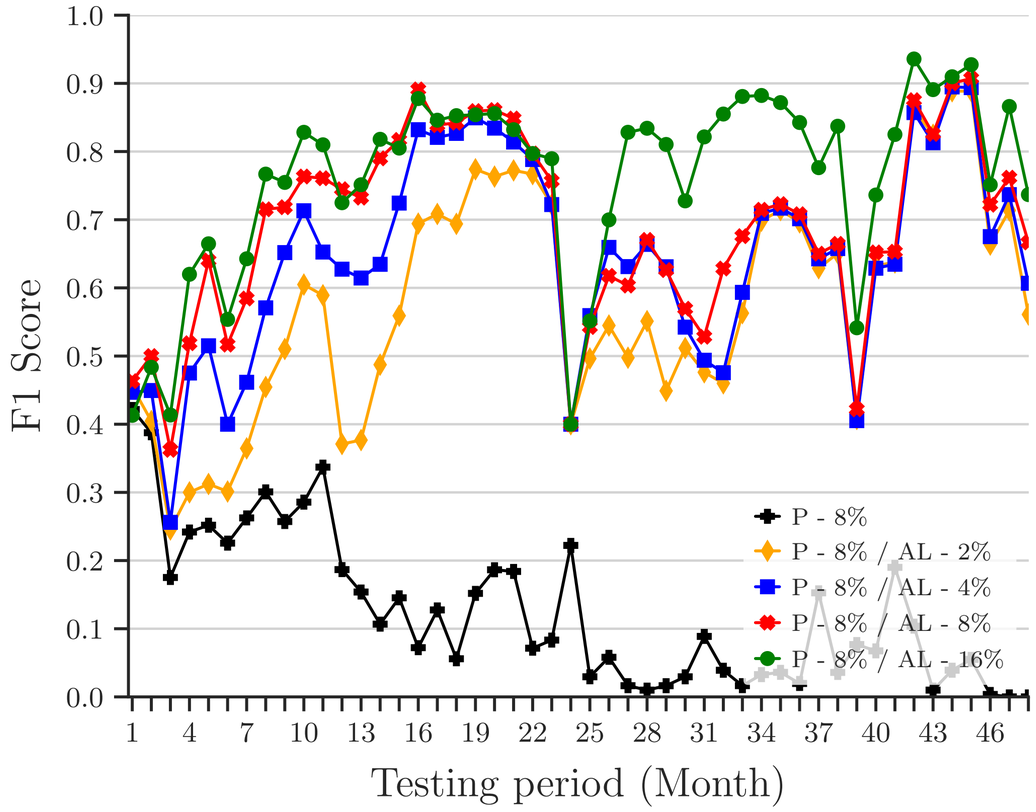
Random Forest



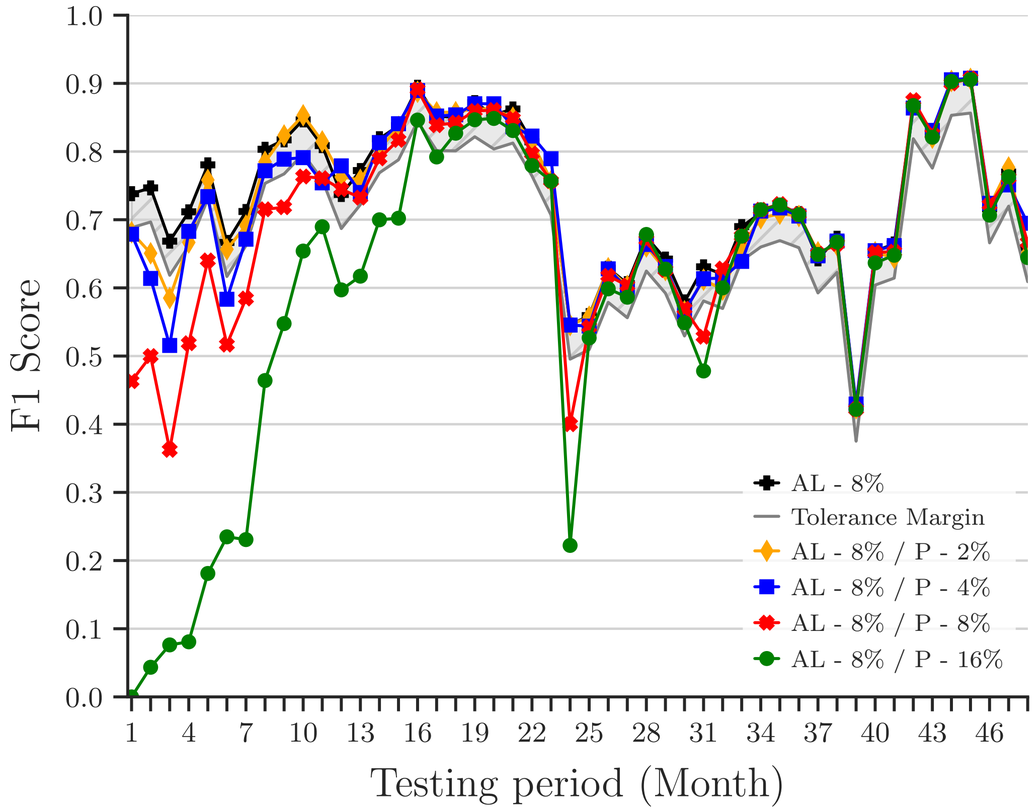
DNN

Results: How fast is passive recovery?

Fixed %P -> increasing %AL

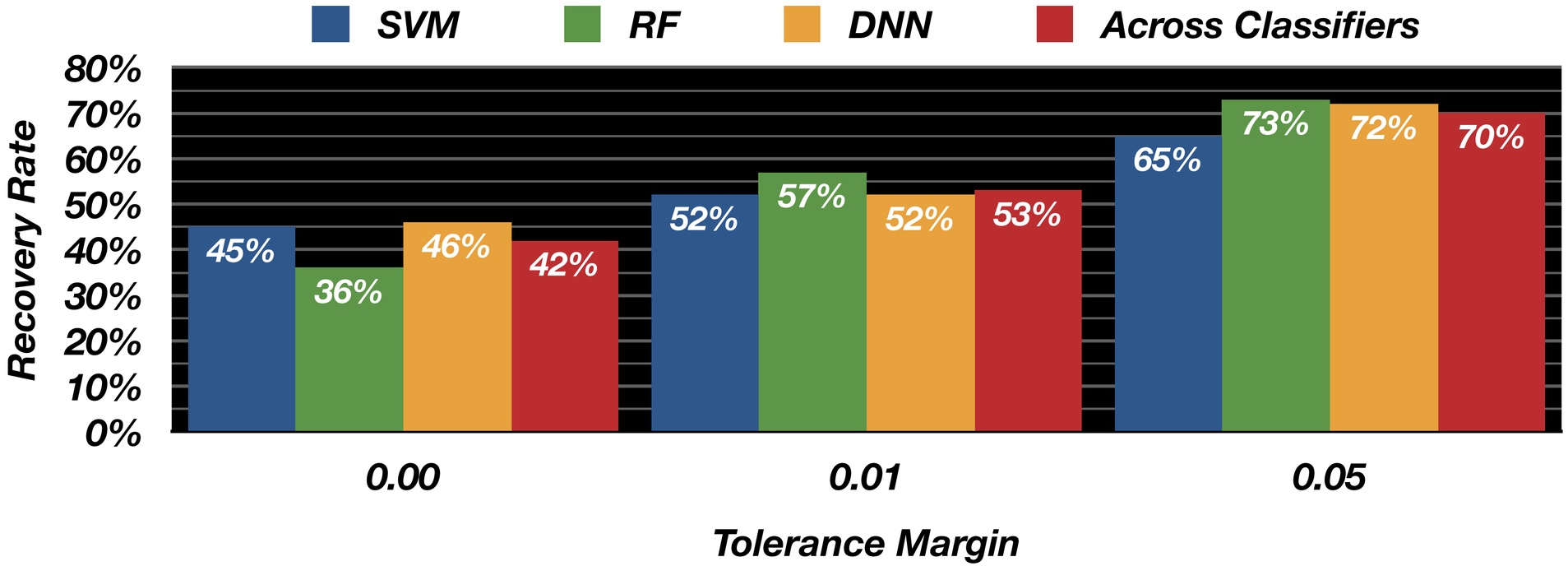


Fixed %AL -> increasing %P



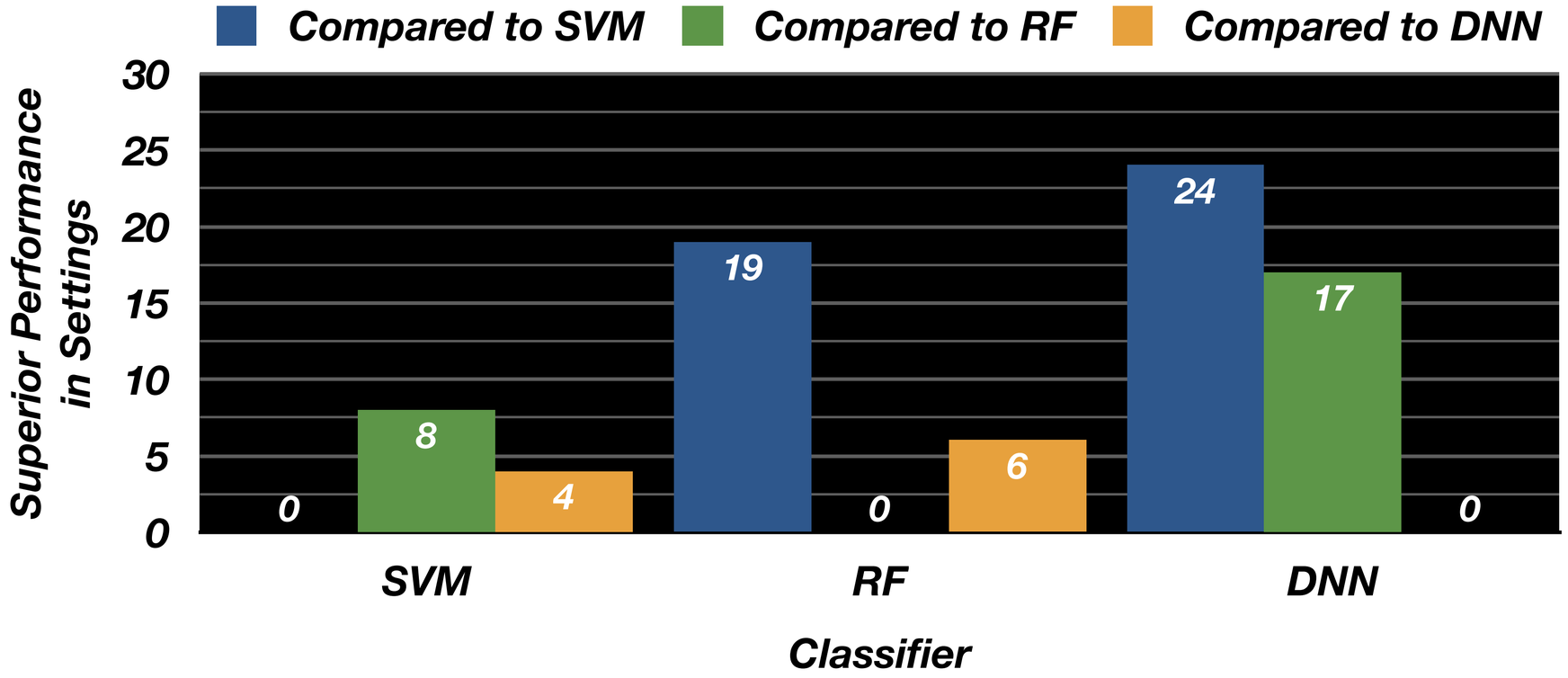
Higher poisoning rates —> delayed intercept (even for high %AL)

Results: *How stable is passive recovery?*



Recovery rate impacted more by %AL and %P than by the classifier

Results: How do classifiers impact passive recovery?



Higher model capacity -> improved overall recovery and performance

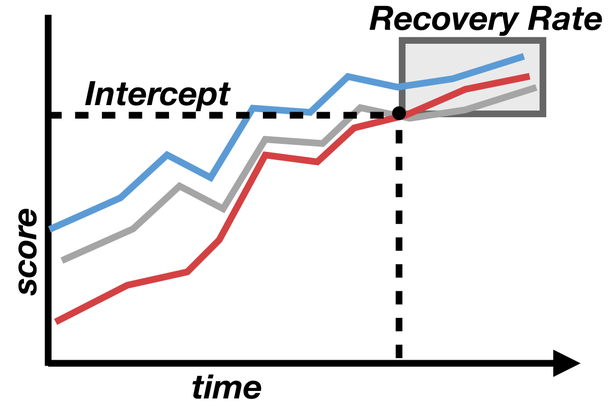
Conclusions

The Impact of Active Learning on Availability Data Poisoning for Android Malware Classifiers

Shae McFadden, Zeliang Kan, Lorenzo Cavallaro, Fabio Pierazzi

Code Repository

- ***Active Learning*** can facilitate passive recovery
- ***For a TM of 0.05***, the average I is 9 months and RR is 70%
- ***All Classifiers*** showed capability of passive recovery
- ***Choice of Classifier*** impacts the overall passive recovery



- ***Open Research Directions:*** *Problem Space Attacks, Time-Aware Poisoning, Relationship with Poison Mitigation*

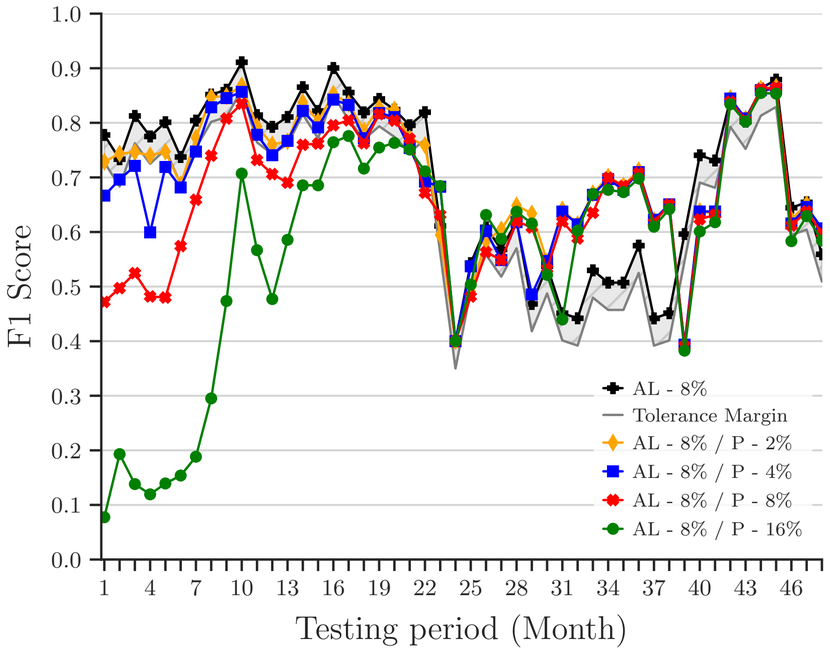
Additional Slides

TABLE 3: Recovery results table for the different tolerance margins (0.05, 0.01, and 0) and classifiers (SVM, DNN, RF). We report *intercept* (lower is better) and *recover rate* (higher is better) for each scenario, and use background gradients to provide a visual cue. The letter "X" is used if the intercept is never reached within the specified tolerance margin.

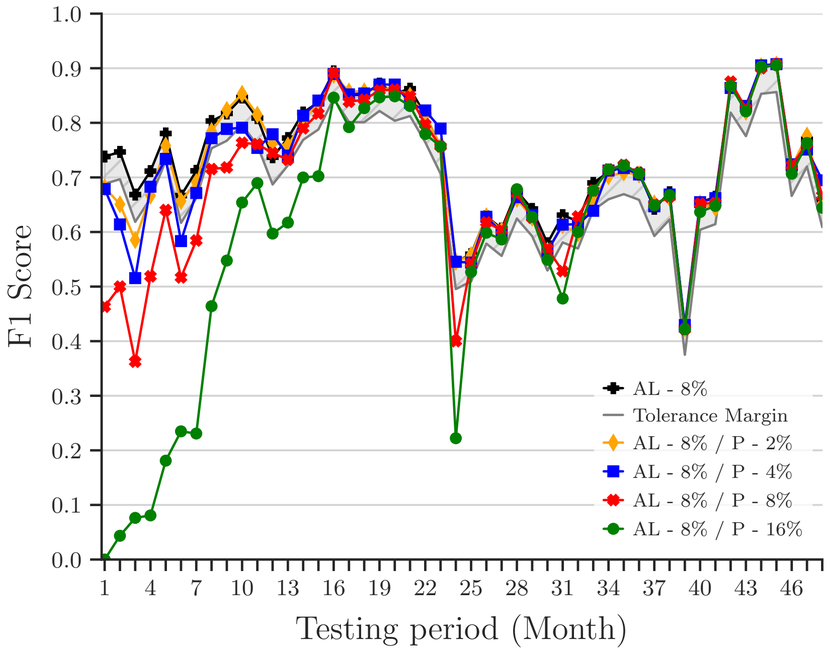
Recovery Results Table

Classifiers			SVM				DNN				RF			
Tolerance Margin	Active Learning Rate		Poisoning Rate				Poisoning Rate				Poisoning Rate			
			2%	4%	8%	16%	2%	4%	8%	16%	2%	4%	8%	16%
0.05	0%	Intercept (Month)	1	24	X	X	1	1	7	X	2	24	X	X
		Recovery Rate (%)	10%	12%	0%	0%	67%	56%	7%	0%	62%	20%	0%	0%
	2%	Intercept (Month)	1	16	22	22	1	1	3	22	1	5	16	21
		Recovery Rate (%)	75%	67%	67%	52%	90%	85%	72%	89%	94%	73%	70%	56%
	4%	Intercept (Month)	1	9	15	22	1	1	5	23	2	9	13	21
		Recovery Rate (%)	90%	88%	79%	67%	77%	75%	61%	81%	87%	90%	89%	75%
	8%	Intercept (Month)	1	2	19	21	1	2	7	16	4	4	11	16
		Recovery Rate (%)	83%	74%	83%	82%	94%	87%	83%	73%	100%	91%	95%	91%
	16%	Intercept (Month)	1	2	9	21	2	1	4	15	1	4	10	14
		Recovery Rate (%)	98%	87%	82%	96%	98%	90%	87%	74%	90%	93%	95%	86%
0.01	0%	Intercept (Month)	X	X	X	X	5	5	23	X	8	24	X	X
		Recovery Rate (%)	0%	0%	0%	0%	20%	23%	8%	0%	27%	4%	0%	0%
	2%	Intercept (Month)	9	22	23	22	1	4	3	22	5	8	23	34
		Recovery Rate (%)	33%	59%	46%	30%	60%	57%	52%	41%	70%	44%	69%	53%
	4%	Intercept (Month)	2	22	22	22	3	6	6	31	9	10	17	33
		Recovery Rate (%)	64%	74%	70%	59%	56%	37%	33%	100%	62%	64%	62%	88%
	8%	Intercept (Month)	2	23	23	23	9	3	15	28	9	12	12	20
		Recovery Rate (%)	45%	77%	65%	65%	65%	52%	53%	95%	70%	76%	59%	52%
	16%	Intercept (Month)	8	21	21	23	2	4	14	15	4	10	14	27
		Recovery Rate (%)	76%	86%	93%	96%	72%	67%	80%	65%	78%	74%	83%	95%
0	0%	Intercept (Month)	X	X	X	X	5	5	23	X	23	24	X	X
		Recovery Rate (%)	0%	0%	0%	0%	16%	18%	8%	0%	31%	4%	0%	0%
	2%	Intercept (Month)	9	22	23	22	10	15	15	22	5	8	23	34
		Recovery Rate (%)	20%	48%	27%	22%	51%	56%	41%	30%	48%	28%	54%	33%
	4%	Intercept (Month)	15	22	22	22	3	6	6	31	9	10	23	33
		Recovery Rate (%)	62%	63%	67%	52%	46%	21%	16%	89%	40%	38%	42%	62%
	8%	Intercept (Month)	2	23	23	23	9	9	15	28	9	12	12	23
		Recovery Rate (%)	38%	57%	50%	54%	56%	50%	44%	95%	40%	46%	32%	19%
	16%	Intercept (Month)	8	23	23	23	2	4	15	23	8	10	14	30
		Recovery Rate (%)	71%	88%	88%	92%	66%	57%	74%	81%	41%	33%	40%	84%

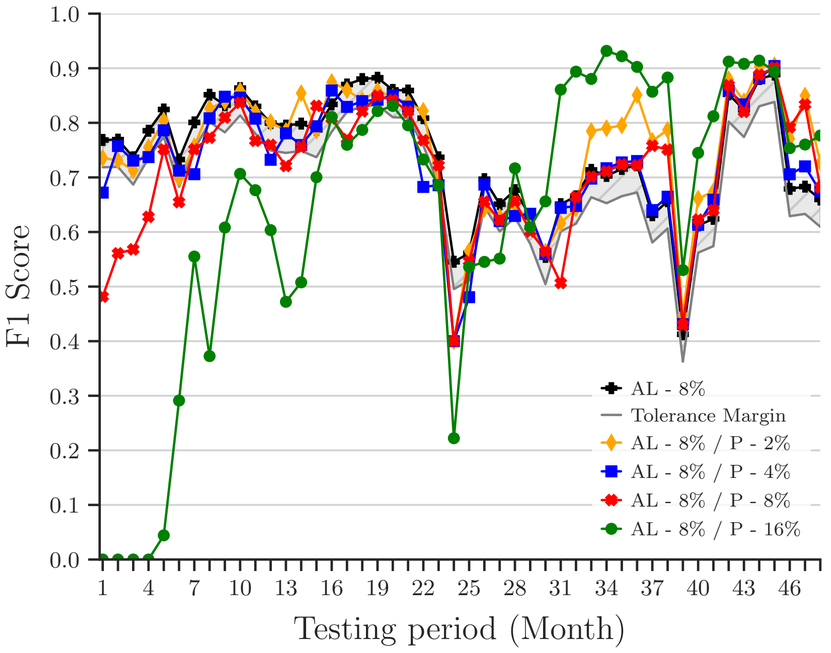
Fixed %AL



SVM

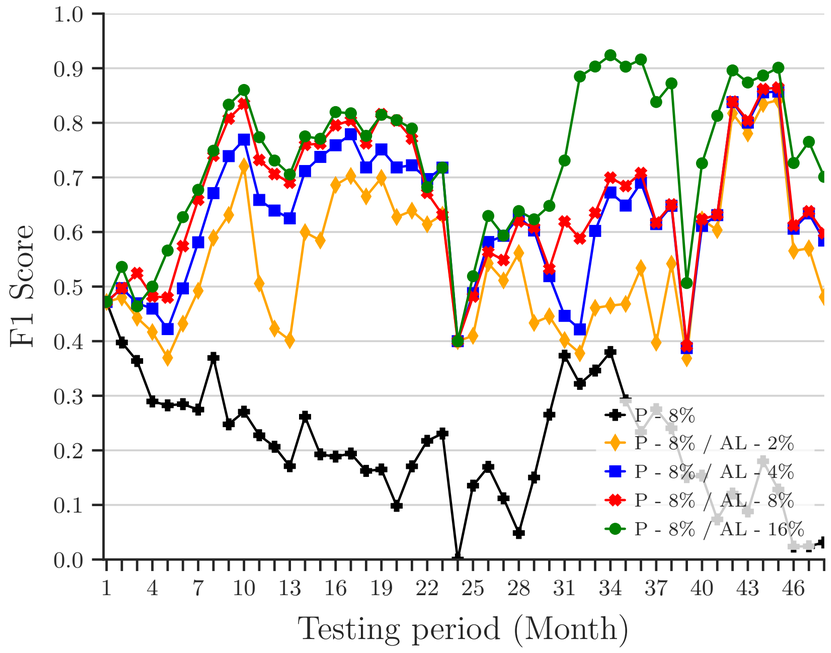


RF

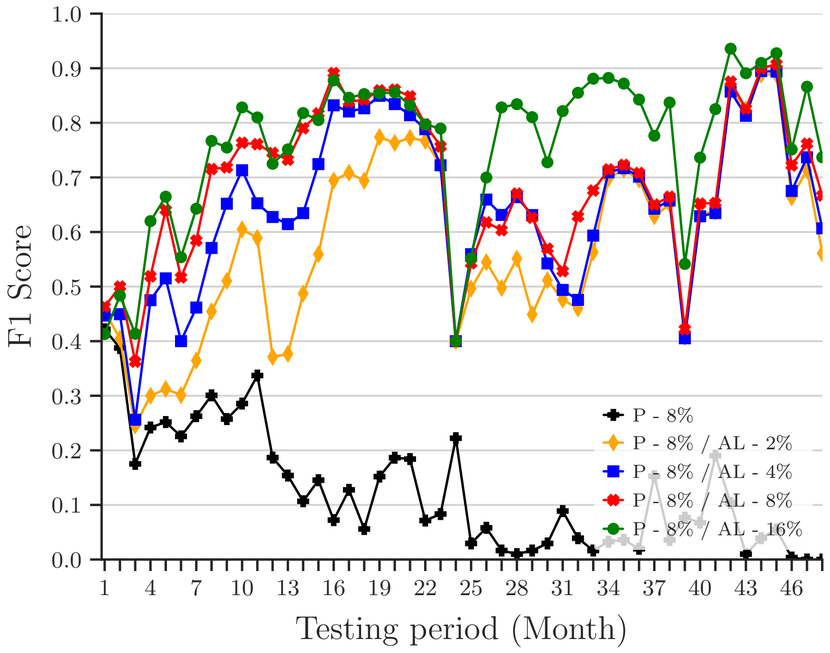


DNN

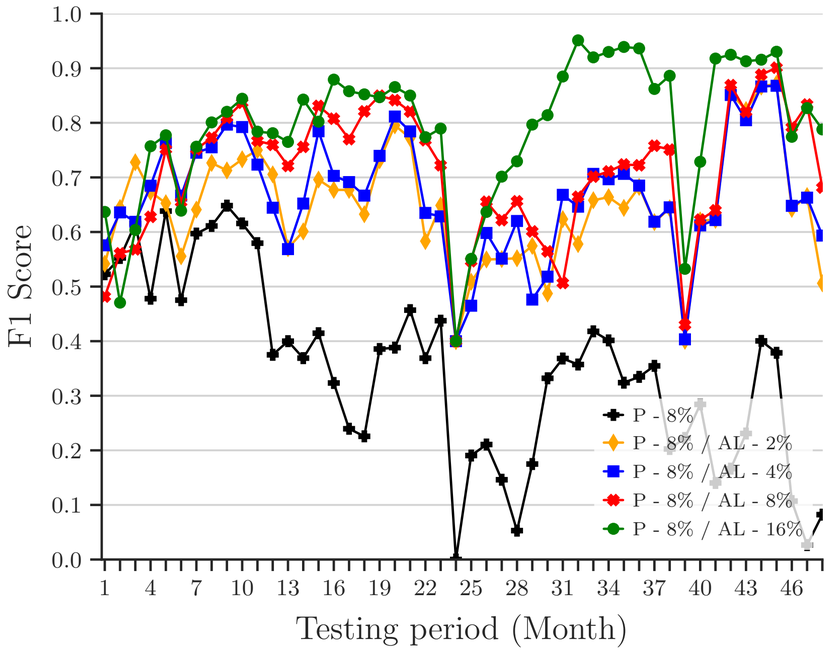
Fixed %P



SVM

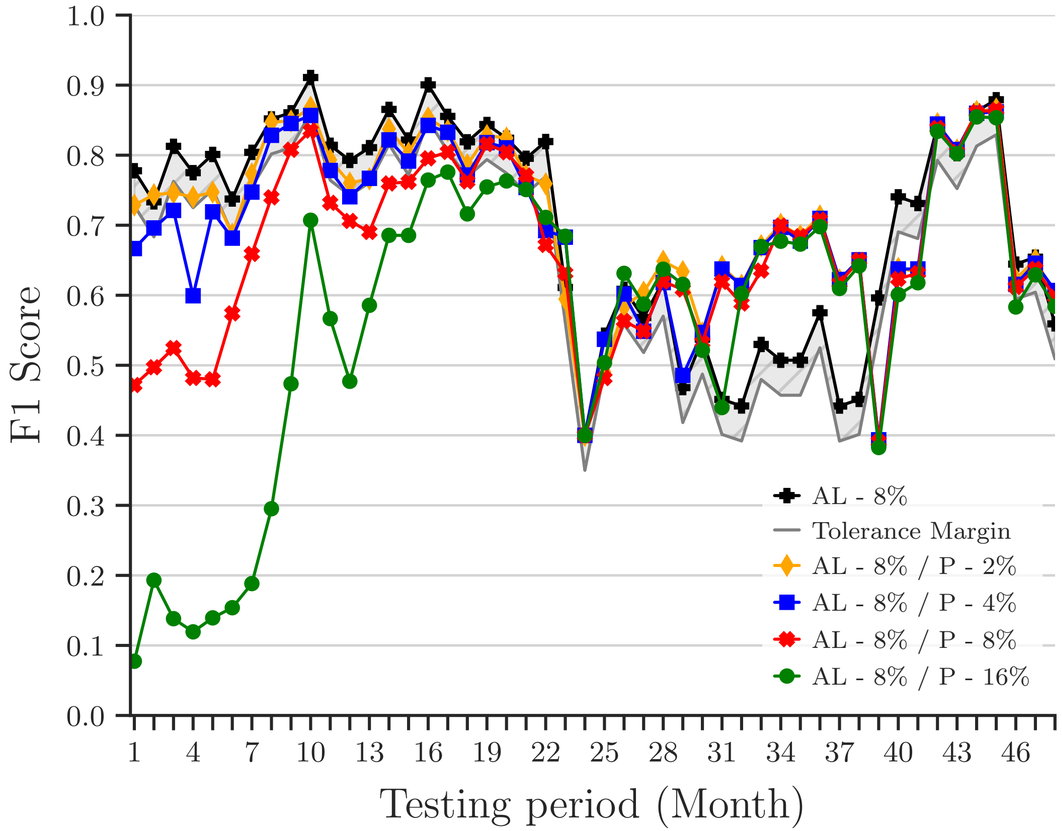


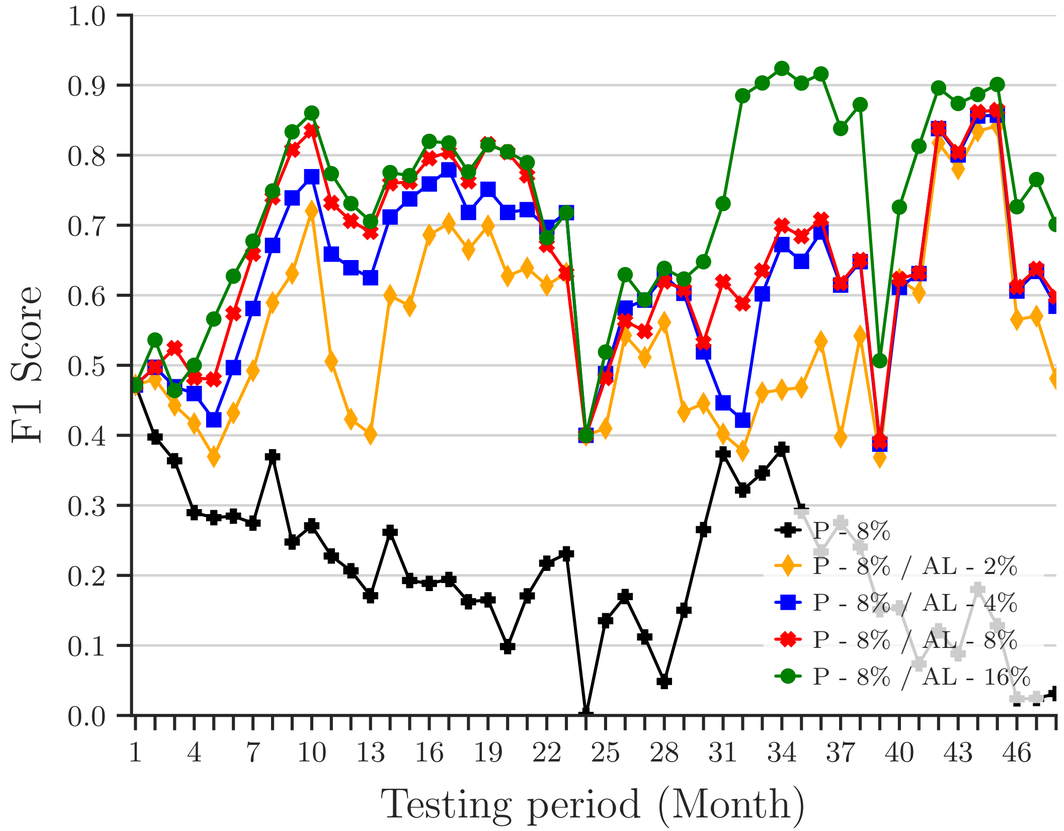
RF



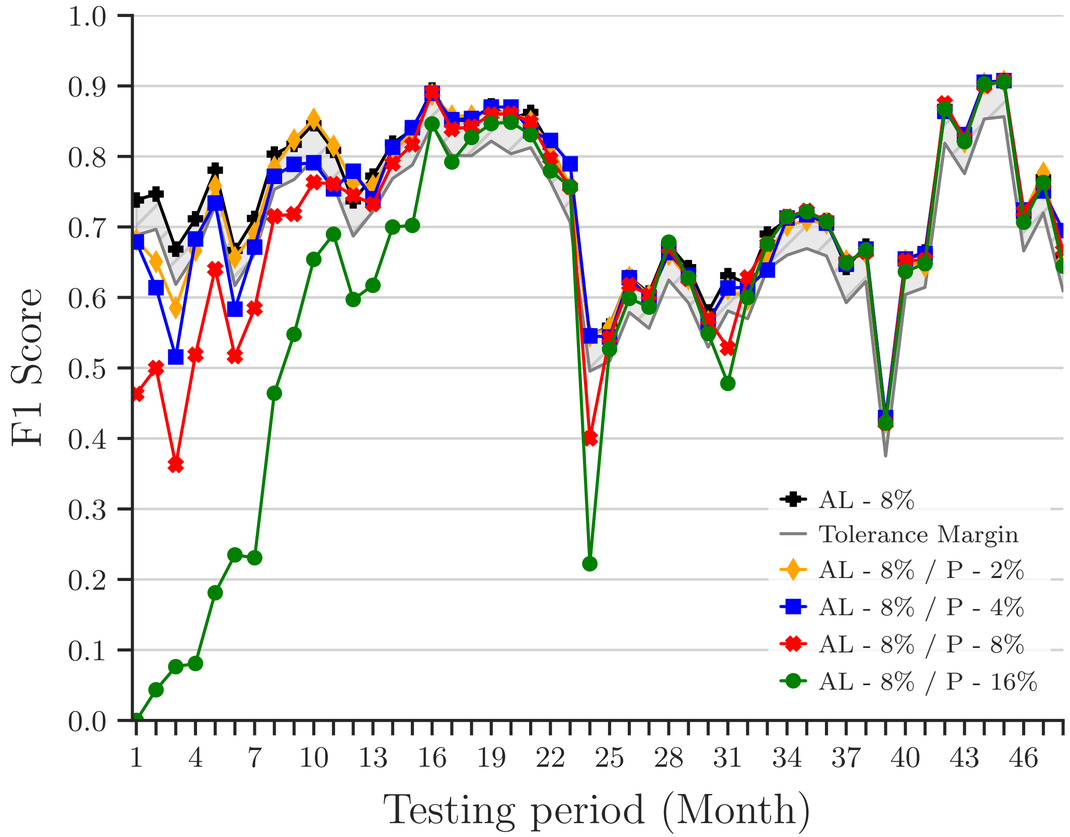
DNN

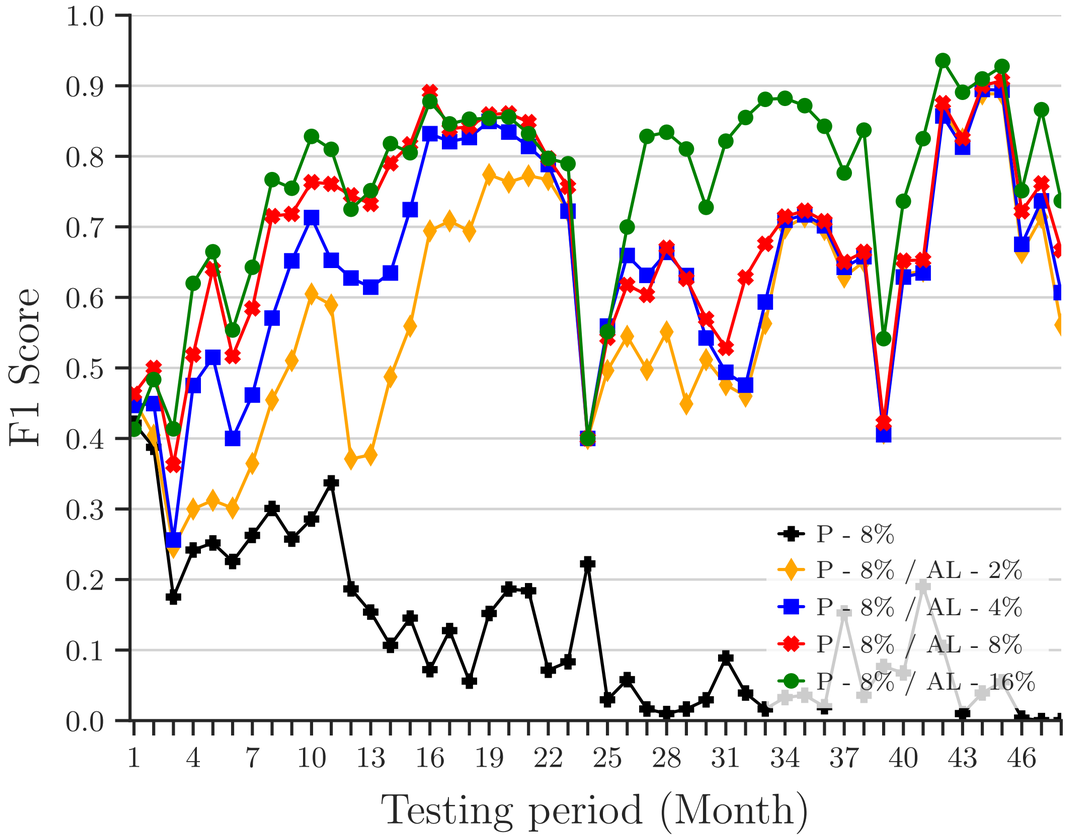
SVM Figures


F1 Score
Testing period (Month)
AL - 8%
Tolerance Margin
AL - 8% / P - 2%
AL - 8% / P - 4%
AL - 8% / P - 8%
AL - 8% / P - 16%

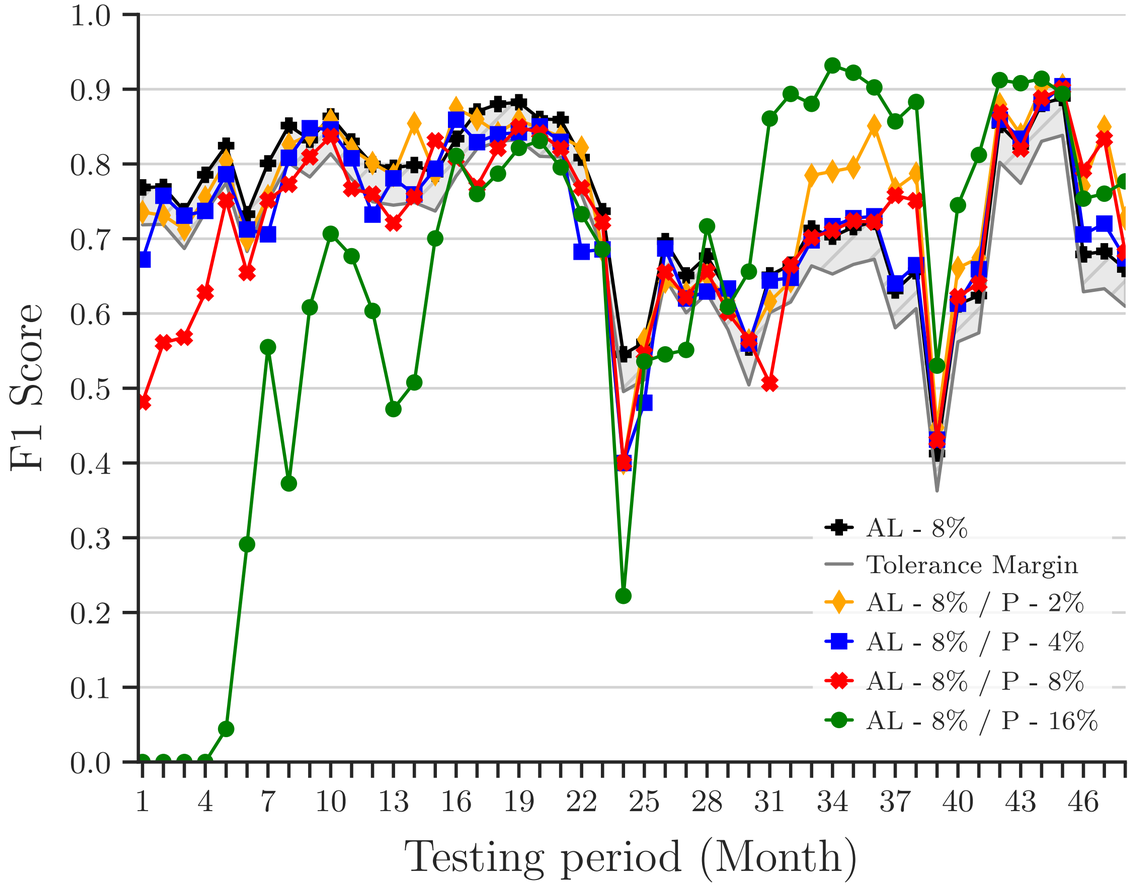

F1 Score
Testing period (Month)
P - 8%
P - 8% / AL - 2%
P - 8% / AL - 4%
P - 8% / AL - 8%
P - 8% / AL - 16%

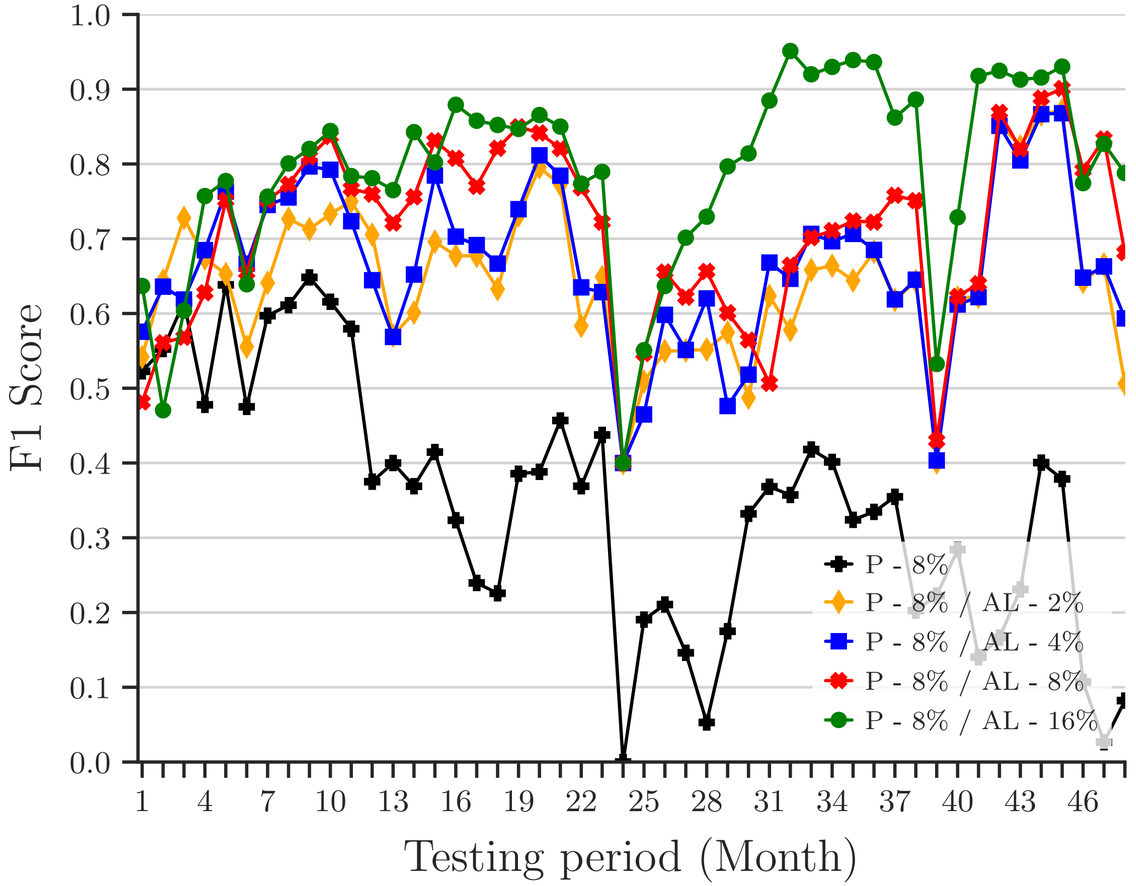
RF Figures


F1 Score
Testing period (Month)
AL - 8%
Tolerance Margin
AL - 8% / P - 2%
AL - 8% / P - 4%
AL - 8% / P - 8%
AL - 8% / P - 16%


F1 Score
Testing period (Month)
P - 8%
P - 8% / AL - 2%
P - 8% / AL - 4%
P - 8% / AL - 8%
P - 8% / AL - 16%

DNN Figures


F1 Score
Testing period (Month)
AL - 8%
Tolerance Margin
AL - 8% / P - 2%
AL - 8% / P - 4%
AL - 8% / P - 8%
AL - 8% / P - 16%


F1 Score
Testing period (Month)
P - 8%
P - 8% / AL - 2%
P - 8% / AL - 4%
P - 8% / AL - 8%
P - 8% / AL - 16%