Transcending TRANSCEND: Revisiting Malware Classification in the Presence of Concept Drift (Correction to Accepted Paper to IEEE S&P 2022)

Federico Barbero*[†‖], Feargus Pendlebury*[‡§¶], Fabio Pierazzi[†], Lorenzo Cavallaro[¶]
[†] King's College London, [‡] Royal Holloway, University of London, [§] The Alan Turing Institute,
[‖] University of Cambridge, [¶] University College London

*Equal contribution.

Dear Thorsten and Thomas,

We have identified a subtle bug in the implementation of our paper "Transcending Transcend: Revisiting Malware Classification in the Presence of Concept Drift" (https://s2lab.cs.ucl.ac.uk/downloads/transcending.pdf), which was accepted at IEEE S&P 2022.

The bug does not invalidate the paper's results. The bug introduced data snooping in one part of the pipeline, which produced unrealistic results. We have re-run all the main experiments of the paper (Figure 7 and Table 1) and verified that our approach is still state-of-the-art, as shown in the next pages.

We would like to issue an errata corrige (and a revised version, if possible) to reflect the corrected implementation and update the experimental results. Having the errata on IEEE eXplore Digital Library would be great, but anything would do.

We advocate and pursue the academic integrity of the scientific process and ensure that our work is reliable and trustworthy. We plan to reach out to all the institutions that received access to our codebase and share the lesson learned with our community, hoping to highlight the importance of research ethics and contribute to the research community.

Thank you for your understanding and support.

Sincerely,

Lorenzo Cavallaro (on behalf of all the authors)

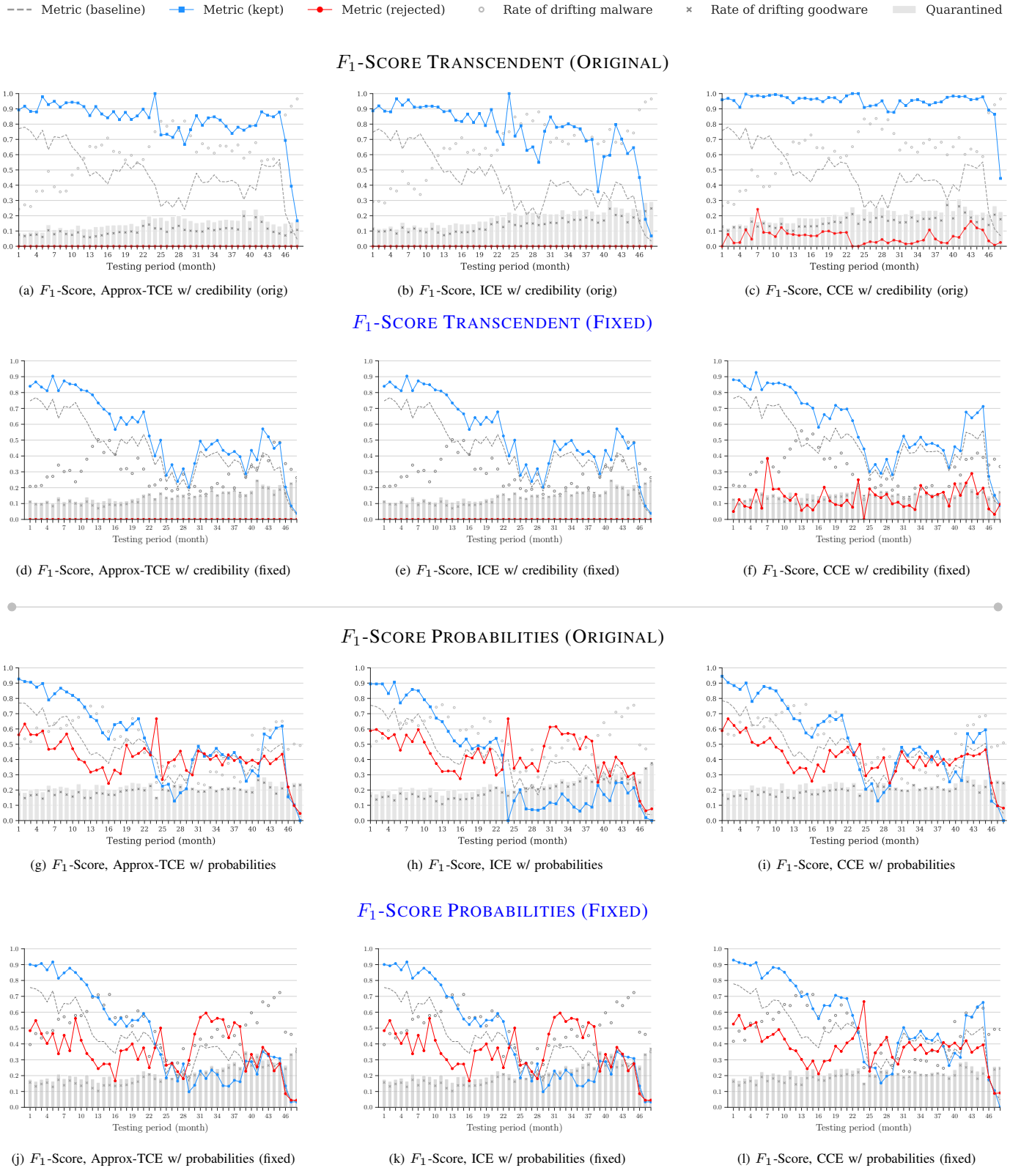
TABLE I: Area Under Time (AUT) of F_1 performance with respect to concept drift over the 48 month test period for different quality metrics: credibility, credibility with confidence, and probabilities (cf. Figure 1) as shown in the original paper (orig columns) and over the fixed data snooping issue (fixed columns). We aim to *maximize* the metrics of kept elements and *minimize* the metrics for rejected elements. Fixing the data snooping issue shows Transcendent on p-values still outperforms the baseline and thresholding on probabilities.

		Approx-TCE (orig)	**Approx-TCE (fixed)**	**ICE (orig)**	**ICE (fixed)**	**CCE (orig)**	**CCE (fixed)**
Baseline	AUT(F_1 w/ credibility, 48m)	.480	.440	.440	.440	.483	.484
	AUT(F_1 w/ cred + conf, 48m)	.480	.440	.440	.440	.483	.484
	AUT(F_1 w/ probability, 48m)	.456	.405	.405	.405	.455	.457
Kept Elements	AUT(F_1 w/ credibility, 48m)	.829	.546	.762	.546	.950	.590
	AUT(F_1 w/ cred + conf, 48m)	.822	.546	.887	.546	.962	.590
	AUT(F_1 w/ probability, 48m)	.531	.444	.388	.444	.532	.555
Rejected Elements	AUT(F_1 w/ credibility, 48m)	.000	.000	.000	.000	.064	.137
	AUT(F_1 w/ cred + conf, 48m)	.000	.000	.000	.000	.063	.137
	AUT(F_1 w/ probability, 48m)	.410	.368	.426	.368	.410	.370

I. Revised Results

This document reports the revised results corresponding to "Figure 7" (from the original paper) as follows:

- Figure 1 and Table I report the breakdown of the Transcendent and probabilities thresholding performance, before and after the fix. This shows that Transcendent is still state of the art, as it outperforms the probabilities, especially in the number of rejected samples and the number of FPs/FNs correctly rejected. A more detailed explanation can be found in the revised captions.
- Figure 2 and Figure 3 report additional breakdowns of Precision, Recall, and the scenario where both credibility and confidence are used.



(a) F_1-Score, Approx-TCE w/ credibility (orig)

(b) F_1-Score, ICE w/ credibility (orig)

(c) F_1-Score, CCE w/ credibility (orig)

(d) F_1-Score, Approx-TCE w/ credibility (fixed)

(e) F_1-Score, ICE w/ credibility (fixed)

(f) F_1-Score, CCE w/ credibility (fixed)

(g) F_1-Score, Approx-TCE w/ probabilities

(h) F_1-Score, ICE w/ probabilities

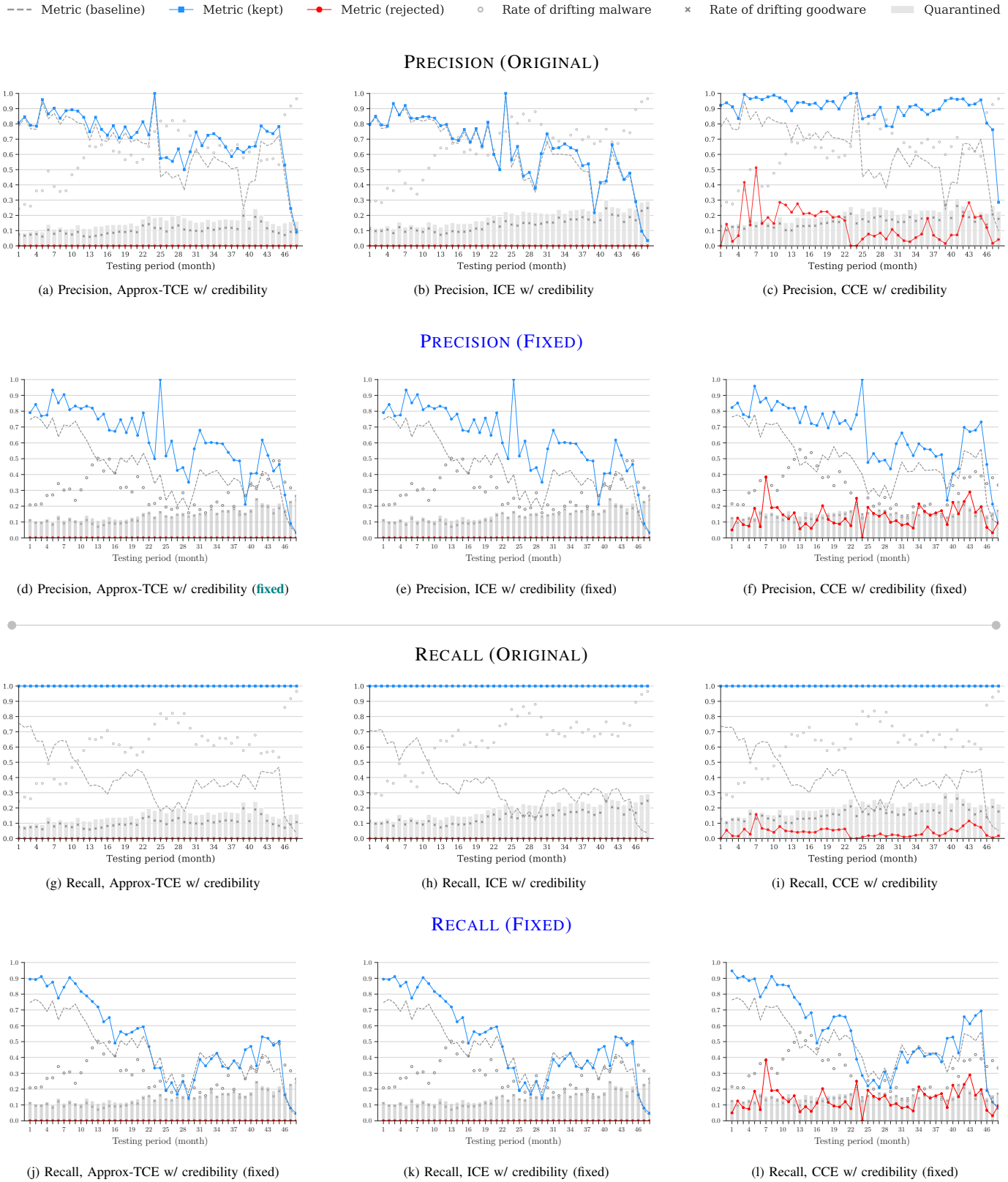
(i) F_1-Score, CCE w/ probabilities

(j) F_1-Score, Approx-TCE w/ probabilities (fixed)

(k) F_1-Score, ICE w/ probabilities (fixed)

(l) F_1-Score, CCE w/ probabilities (fixed)

Fig. 1: **Comparison of F_1-Score between Transcendent and probabilities for classification with rejection, before and after the data snooping fix.** Performance for the three proposed conformal evaluators (columns) using different quality metrics. The first row shows F_1-Score of the different evaluators using thresholding on credibility p-values as shown in the original paper. The second row depicts the same scenario over the fixed data snooping issue. The lower two rows show different conformal evaluators using thresholding on probabilities as shown in the original paper and over the fixed data snooping issue, respectively. The upper line (□ marker) shows the performance on kept examples whose classifications were accepted. The lower line (○ marker) shows the performance on rejected examples. These are the FPs/FNs that would have been made if the predictions were accepted by the degrading model. The bars show the proportion of rejected elements in each period, while the x and o markers show the proportion of *ground truth* malware and goodware that was identified as drifting and quarantined, respectively. We can observe that all the conformal evaluators with thresholding on credibility p-values (Transcendent) still outperform the baseline and the approach based on probabilities across all the experiments.



(a) Precision, Approx-TCE w/ credibility

(b) Precision, ICE w/ credibility

(c) Precision, CCE w/ credibility

(d) Precision, Approx-TCE w/ credibility (**fixed**)

(e) Precision, ICE w/ credibility (fixed)

(f) Precision, CCE w/ credibility (fixed)

(g) Recall, Approx-TCE w/ credibility

(h) Recall, ICE w/ credibility

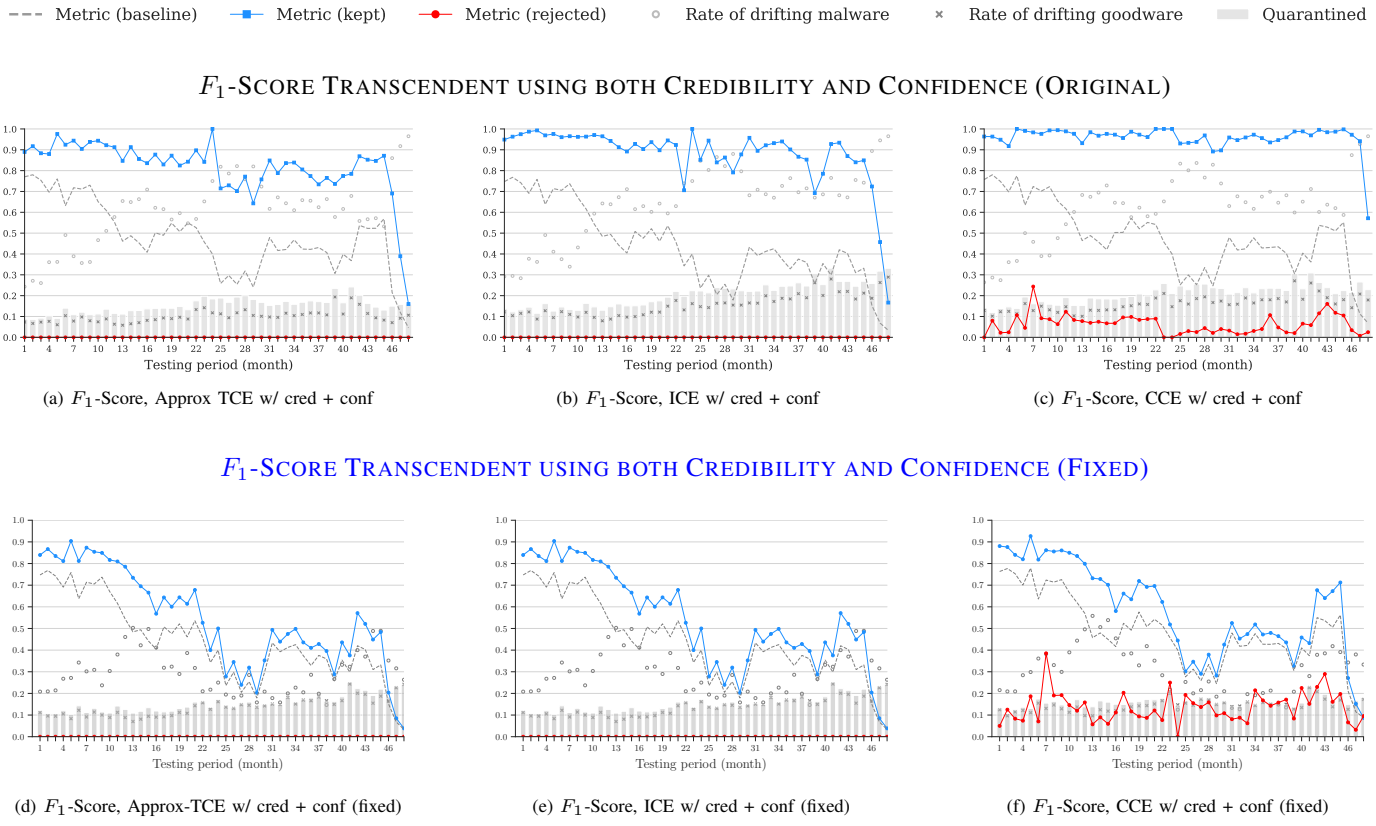
(i) Recall, CCE w/ credibility

(j) Recall, Approx-TCE w/ credibility (fixed)

(k) Recall, ICE w/ credibility (fixed)

(l) Recall, CCE w/ credibility (fixed)

Fig. 2: **Comparison of Precision and Recall in Transcendent, before and after the data snooping fix.** Performance for the three proposed conformal evaluators (columns) using different quality metrics. The four rows show the Precision and Recall of the different evaluators using thresholding on credibility p-values as shown in the original paper and over the fixed data snooping issue, respectively. The upper line (□ marker) shows the performance on kept examples whose classifications were accepted. The lower line (○ marker) shows the performance on rejected examples. These are the FPs/FNs that would have been made if the predictions were accepted by the degrading model. The bars show the proportion of rejected elements in each period, while the x and o markers show the proportion of *ground truth* malware and goodware that was identified as drifting and quarantined, respectively. We can observe that all the conformal evaluators with thresholding on credibility p-values (Transcendent) still outperform the baseline and the approach based on probabilities across all the experiments.



(a) F_1-Score, Approx TCE w/ cred + conf

(b) F_1-Score, ICE w/ cred + conf

(c) F_1-Score, CCE w/ cred + conf

(d) F_1-Score, Approx-TCE w/ cred + conf (fixed)

(e) F_1-Score, ICE w/ cred + conf (fixed)

(f) F_1-Score, CCE w/ cred + conf (fixed)

Fig. 3: **Comparison of F_1-Score on Transcendent with using both "credibility" and "confidence" scores, before and after the data snooping fix.** Performance for the three proposed conformal evaluators (columns) using different quality metrics. The two rows show the F_1-Score of the different evaluators using thresholding on credibility and confidence p-values as shown in the original paper and over the fixed data snooping issue, respectively. The upper line (□ marker) shows the performance on kept examples whose classifications were accepted. The lower line (○ marker) shows the performance on rejected examples. These are the FPs/FNs that would have been made if the predictions were accepted by the degrading model. The bars show the proportion of rejected elements in each period, while the x and o markers show the proportion of *ground truth* malware and goodware that was identified as drifting and quarantined, respectively. We can observe that all the conformal evaluators with thresholding on credibility p-values (Transcendent) still outperform the baseline and the approach based on probabilities across all the experiments.