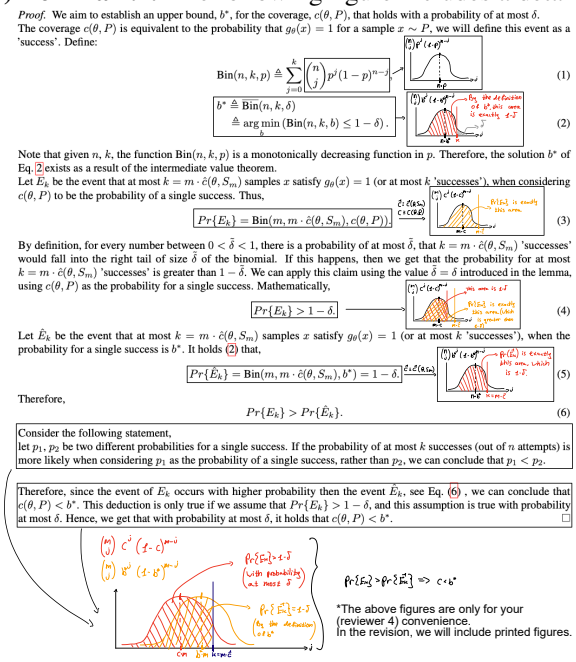
Reviewer 4: Thanks for your constructive feedback and detailed comments. We understand your technical concerns, this is particularly unfortunate because otherwise you mentioned that: "The paper would have been significant if the technical pieces were correct and the presentation was clear". Though we take full responsibility for the technical misunderstandings caused by insufficient explanation, we hope we can convince you that technically the paper is flawless. **1) Lemma 4.1.** The following figure includes a detailed


Proof. We aim to establish an upper bound, b^*, for the coverage, $c(\theta, P)$, that holds with a probability of at most δ.
The coverage $c(\theta, P)$ is equivalent to the probability that $g_\theta(x) = 1$ for a sample $x \sim P$, we will define this event as a 'success'. Define:

$$\text{Bin}(n,k,p) \triangleq \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} \quad (1)$$

$$b^* \triangleq \overline{\text{Bin}}(n,k,\delta) \triangleq \arg\min_b \left(\text{Bin}(n,k,b) \leq 1-\delta\right). \quad (2)$$

Note that given n, k, the function $\text{Bin}(n,k,p)$ is a monotonically decreasing function in p. Therefore, the solution b^* of Eq. 2 exists as a result of the intermediate value theorem.
Let E_k be the event that at most $k = m \cdot \hat{c}(\theta, S_m)$ samples x satisfy $g_\theta(x) = 1$ (or at most k 'successes'), when considering $c(\theta, P)$ to be the probability of a single success. Thus,

$$Pr\{E_k\} = \text{Bin}(m, m \cdot \hat{c}(\theta, S_m), c(\theta, P)) \quad (3)$$

By definition, for every number between $0 < \tilde{\delta} < 1$, there is a probability of at most $\tilde{\delta}$, that $k = m \cdot \hat{c}(\theta, S_m)$ 'successes' would fall into the right tail of size $\tilde{\delta}$ of the binomial. If this happens, then we get that the probability for at most $k = m \cdot \hat{c}(\theta, S_m)$ 'successes' is greater than $1 - \tilde{\delta}$. We can apply this claim using the value $\tilde{\delta} = \delta$ introduced in the lemma, using $c(\theta, P)$ as the probability for a single success. Mathematically,

$$Pr\{E_k\} > 1 - \delta. \quad (4)$$

Let $\hat{E}_k$ be the event that at most $k = m \cdot \hat{c}(\theta, S_m)$ samples x satisfy $g_\theta(x) = 1$ (or at most k 'successes'), when the probability for a single success is b^*. It holds (2) that,

$$Pr\{\hat{E}_k\} = \text{Bin}(m, m \cdot \hat{c}(\theta, S_m), b^*) = 1 - \delta. \quad (5)$$

Therefore,

$$Pr\{E_k\} > Pr\{\hat{E}_k\}. \quad (6)$$

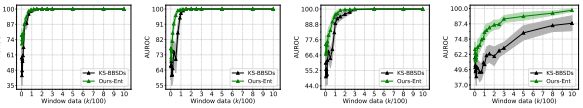
Consider the following statement,
let p_1, p_2 be two different probabilities for a single success. If the probability of at most k successes (out of n attempts) is more likely when considering p_1 as the probability of a single success, rather than p_2, we can conclude that $p_1 < p_2$.

Therefore, since the event of E_k occurs with higher probability then the event $\hat{E}_k$, see Eq. (6) , we can conclude that $c(\theta, P) < b^*$. This deduction is only true if we assume that $Pr\{E_k\} > 1 - \delta$, and this assumption is true with probability at most δ. Hence, we get that with probability at most δ, it holds that $c(\theta, P) < b^*$. □

*The above figures are only for your (reviewer 4) convenience. In the revision, we will include printed figures.

proof with figures we have added to make your reading more convenient. **2) misuse of t-test?** Our explanation indeed lacks in this regard. A well known rule of thumb is when the number of random variables (RVs) is ≥ 30, the summation is sufficiently close to being normally distributed; see 'An Introduction to Statistical Learning' (James *et al.*, p. 67). In the vast majority of our applications (Algorithm 2) this number is way larger. Specifically, we apply the t-test over $\hat{\mu}$, which is an average of $n = k\lfloor \log m \rfloor$ RVs (the window size $k \geq 3$ and $\lfloor \log m \rfloor \geq 13$). **3) Missing out-of-distribution refs.** Thanks for pointing out the missing refs, they will be included. **4) Improve on notations and presentation.** We appreciate your suggestions on how to improve the notations and provide intuition before presenting the lemma and theorem. Corrections will be made in the revised version.

Reviewer 5: Thank you for your thorough and constructive feedback. Due to limited space, we will discuss below the more important points you mentioned, but in the revised version, we will address each and every point you raised. **1) 'sliced' analysis.** A 'sliced' analysis will be included in the revision. For example, among others, the following four graphs will be included. The following four graphs demonstrate the AUROC metric where $p = 2/3$, for the following diverse shifts separately: ImageNet C-severity 1, rotation of 180 degrees, rotation of 40 degrees and the best considered adversarial attack, CW. For all window sizes, the green line



dominates the black line, namely, our method outperforms the best baseline. The same is true for the vast majority of other shifts as well as for the two other contamination values, $p = 1, 1/3$. All these other results will be included in the revised version. **2) Related work, including out-of-distribution (OOD) papers.** We agree that the OOD papers you mentioned should be included in our paper, and they will definitely be included in the revision. However, even though an OOD detector can be applied easily on a window, by checking if any instance appears to be OOD, it isn't yet clear how to aggregate the results to determine if a shift has occurred, and how severe it is, at a window level. Therefore, window level detection has its own motivation and rights of existence, as argued, e.g., by Rabanser *et al.* (Neurips 2019). To elaborate, Fig.3 is showing that a detection over a window may be more sensitive than single instance detection. Whenever $0 < k < 50$ the AUROC is less than 90, but as the window size increases, the AUROC reaches 100; a similar case (using the accuracy metric) is shown by Rabanser *et al.* **3) A few baselines.** Window detection is a novel problem in deep learning (as you stated: 'The problem of *windowed* detection is novel'), and the only methods to detect a distribution shift over a window, are those presented in Rabanser *et al.*, which are mentioned, and compared against. **4) Held out set.** Although we require a held-out set for an additional training step, the training step is quick (no weights to train), and the baselines (Rabanser *et al.*) also require such a set for the two samples test. However, we only use it once, and do not reference it again (only the statistics we extract from it). This is in contrast to baselines, which must refer to this held-out set in every window detection. **5) Small reported boost in one metric.** The AUPR metric is label dependant, because precision is more focused in the positive class. Therefore, it is common to report AUPR for the two possible label choices (AUPR-Tr, AUPR-Te, in our paper), see Liang *et al.*, (ICLR 2018). However, in our case, since the data is imbalanced, one metric (AUPR-Tr) is much more informative than the other (AUPR-Te). In any case, thanks for your note, we will remove the AUPR-Te metric from the revision, which is not informative in our case. **6) An 'atypical' experimental setup.** We wanted to demonstrate the results on ImageNet, which lacks a validating set, so we had to simulate one. **7) Writing and presentation + Estimating prediction quality.** Due to limited space, both of these topics will be discussed together. We will follow your recommendations (section ordering, metrics introduction, clarifying figures/tables/definitions/notations etc.) and make the necessary corrections. There was an omission concerning the topic of estimating prediction quality. We will add a discussion with references to the papers you cited.