International Journal of Research and Development in Engineering Sciences
ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/
Regular Issue, Vol. 8, Issue. 6 (November – December) , 2026 , Pages: 20 – 30
Received: 21 April 2026 ; *Accepted:* 31 October 2026 ; *Published:* 11 November 2026
https://doi.org/10.63328/IJRDES-V8RI6P3
https://www.youtube.com/watch?v=mgSK92VrOAs
https://archive.org/details/jsp-ijrdes-v8ri6p4
Original Paper : Cross-Sectional Research – Experimental Research Paper

Computational and Statistical Reliability of Machine Learning Performance Metrics: AUC Variance and Confidence Interval Failure

Sruthi Pushadapu

Department of Department of Data Science, New Jersey Institute of Technology, Newark, New Jersey, USA.

** Corresponding Author : Sruthi Pushadapu; sruthipushadapu22@gmail.com*

Abstract: The area under the ROC curve (AUC) is reported with a confidence interval, usually from a variance formula or a bootstrap, and both are trusted at sample sizes where they are not reliable. We study the Mann–Whitney estimate of the AUC in the binormal model, where the exact variance is a function of two integrals. We show that the Hanley–McNeil approximation overstates the variance by up to a factor of 3.1, that the variance is governed by the number of positives (extra negatives beyond about ten per positive reduce it by less than 10%), and that a 95% interval of half-width 0.05 at AUC 0.9 needs 138 cases with balanced classes but 3,478 with 1% positives. We also prove a bound on coverage: any interval that collapses to a point when the estimate equals 1 has coverage at most $1 - p_1$, where p_1 is the probability that every positive outscores every negative; a one-dimensional integral gives p_1, and it matched simulation within 0.006. At AUC 0.98 with 20 positives and 380 negatives, the best of five methods covered 98% instead of 95%. On the breast cancer and digits datasets, with the AUC near 0.995, coverage with 20 positives and 180 negatives ranged from 45% to 68%, while for a weak feature (AUC 0.72) the same methods were close to nominal. Synthetic results are simulations; the real data are two small benchmark sets sampled with replacement.

Keywords: AUC, ROC curve, confidence interval, Mann–Whitney statistic, class imbalance, bootstrap.

1. Introduction

A classifier is often summarised by one number, the area under its ROC curve, and a careful report adds a confidence interval [1, 8]. The interval usually comes from a software default: a normal approximation with a variance from DeLong and colleagues [2], from Hanley and McNeil [1], or a bootstrap [6, 9]. These are trusted at the sample sizes where medical, fraud and defect-detection studies are run, where there may be a few dozen positives. Two features of that setting are easy to miss. The variance of the AUC estimate is governed mainly by the number of positives, not the total sample size. And when the AUC is close to one, a finite sample can separate the classes perfectly, in which case the estimated variance is zero. We study how many positives are needed and which interval to use, in the binormal model [11] where the exact variance can be computed, and then on real data. The binormal model is only a convenient reference. It lets us separate the error of an approximation from the error of the sampling, which a real dataset cannot. We make five contributions.

- The exact variance of the Mann–Whitney AUC estimate in the binormal model, with a measured size of the Hanley–McNeil error (Section 4).
- A result on the value of extra negatives, which gives diminishing returns beyond about ten per positive, and the total sample size for a given interval width as a function of prevalence.
- A coverage bound for intervals that degenerate when the estimate equals one, with a closed-form integral for the probability of that event.
- A simulation of five interval methods over AUC from 0.7 to 0.98, balanced and 5% prevalence, and from 10 to 200 positives (Section 6).
- A real-data check on the breast cancer and digits datasets, including a weak-feature case.

All synthetic experiments are simulations from a known model. The real-data experiment treats each dataset as the population and samples from it with replacement.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

2. Background and Related Work

The AUC as a U-statistic. The AUC equals the probability that a random positive scores above a random negative, and its empirical version is the Mann–Whitney statistic [4]. This is a two-sample U-statistic, and its variance and asymptotic normality follow from Hoeffding's theory [5, 15]. Bamber related it to the area under the ROC curve [3]. Hanley and McNeil gave a variance formula by approximating two probabilities with the values for exponential scores [1]. DeLong and colleagues estimated the variance with structural components, which are one-sample placements of each score among the other class [2].

Intervals. Cortes and Mohri studied intervals for the AUC and the variance of the estimate as a function of the class sizes [7]. The bootstrap is common in software [6, 9]. For a proportion near the boundary, the normal (Wald) interval is known to behave badly, which motivated transformed and score-based intervals [10]; the AUC is a proportion-like quantity with a similar boundary.

What we add. The ingredients are known. We contribute the exact binormal variance and the size of the Hanley–McNeil error; the diminishing value of negatives and the resulting sample size by prevalence; a bound on coverage near one with a computable probability; and a comparison of five intervals in conditions where the positives are few.

3. Setting and Methods

3.1. The binormal model

Scores of negatives are $N(0, 1)$ and scores of positives are $N(d, 1)$. The AUC is $\theta = \Phi(d/\sqrt{2})$. A sample has n_1 positives and n_0 negatives, so the prevalence is $\pi = n_1/(n_1 + n_0)$ in a balanced-by-design or random sample. The estimate is

$$\hat{\theta} = (1/(n_1 n_0)) \Sigma_i \Sigma_j 1[X_i > Y_j], \quad (1)$$

the fraction of positive–negative pairs that are correctly ordered. It is unbiased for θ.

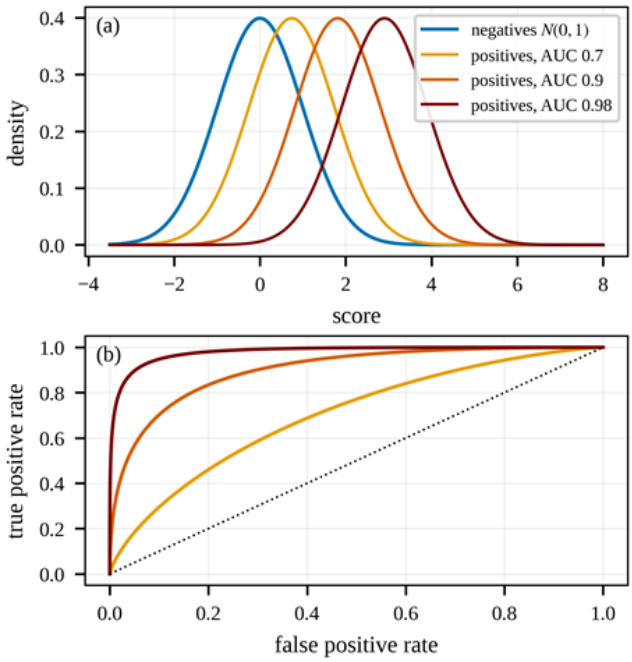

Figure 1: The binormal model. (a) Score densities for negatives and for positives at three AUC values. (b) The corresponding ROC curves. The AUC is the probability that a random positive scores above a random negative.

3.2. Interval Methods

All intervals have nominal level 95%. **Wald, DeLong:** $\hat{\theta} \pm 1.96\,\hat{s}$, with $\hat{s}^2$ from the placements of each score among the other class [2].

Wald, Hanley–McNeil: the same form with the variance of Eq. (2) evaluated at $\hat{\theta}$ with the exponential-score approximations [1].

Wald, binormal: the variance of Eq. (2) with the exact binormal integrals at $\hat{d} = \sqrt{2}\,\Phi^{-1}(\hat{\theta})$, read from a lookup table.

Logit, DeLong: the DeLong interval for logit θ, mapped back with the delta method, which keeps the limits inside (0, 1).

Bootstrap: the percentile interval from 200 stratified resamples of positives and negatives. When $\hat{\theta}$ is 0 or 1, the Wald and logit intervals are the single point $\hat{\theta}$, and the bootstrap interval is also a point because every resample gives the same value.

4. Proposed System

Proposition 1 (exact variance). *Let $Q_1 = Pr(X_1 > Y, X_2 > Y)$ and $Q_2 = Pr(X > Y_1, X > Y_2)$ for independent positives X, X_i and*

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

negatives Y, Y_j. Then

$$\mathrm{Var}(\theta) = [\theta(1-\theta) + (n_1-1)(Q_1-\theta^2) + (n_0-1)(Q_2-\theta^2)] / (n_1 n_0). \quad (2)$$

This is the variance of a two-sample U-statistic [3, 5]. In the binormal model, $Q_1 = \int \varphi(y)[1 - \Phi(y - d)]^2 \, dy$ and $Q_2 = \int \varphi(x - d)\Phi(x)^2 \, dx$, which are one-dimensional integrals. Hanley and McNeil replace them by $\theta/(2 - \theta)$ and $2\theta^2/(1 + \theta)$, which are exact for exponential scores but not for Gaussian ones [1].

Corollary 1 (negatives and positives). *For large samples,*

$$\mathrm{Var}(\theta) \approx (Q_2 - \theta^2)/n_1 + (Q_1 - \theta^2)/n_0. \quad (3)$$

The first term shrinks with the number of positives and the second with the number of negatives. As n_0 grows, the variance approaches $(Q_2 - \theta^2)/n_1$, so the number of positives sets a floor that no number of negatives can lower.

Corollary 2 (sample size). *For a 95% interval of half-width h in a sample with prevalence π, the total number of cases is about*

$$n \approx (1.96^2/h^2) \, [\, (Q_2 - \theta^2)/\pi + (Q_1 - \theta^2)/(1 - \pi) \,]. \quad (4)$$

The bracket grows like $1/\pi$ for rare positives, so the required sample size is roughly proportional to the inverse of the prevalence.

Proposition 2 (coverage bound). *Let $p_1 = Pr(\hat{\theta} = 1)$ be the probability that every positive outscores every negative. If an interval equals $\{\theta\}$ whenever $\hat{\theta} = 1$, as the Wald, logit and percentile bootstrap intervals do, then for $\theta < 1$ its coverage is at most $1 - p_1$, where in the binormal model*

$$p_1 = n_0 \int \varphi(y) \, \Phi(y)^{n_1 - 1} \, [1 - \Phi(y - d)]^{n_0} \, dy. \quad (5)$$

Proof. When $\hat{\theta} = 1$ the DeLong variance is zero, and every bootstrap resample also has $\hat{\theta} = 1$. The interval is then $\{1\}$, which misses $\theta < 1$. The probability of this event is the probability that the smallest positive exceeds the largest negative; conditioning on the largest negative, whose density is $n_0 \varphi(y) \Phi(y)^{n_1 - 1}$, gives Eq. (5).

The bound is not a property of a method but of the sample: when the classes can be perfectly separated by chance, no interval built from the data alone can reflect uncertainty.

5. Experimental and Implementation

Synthetic experiments use the binormal model with $\theta \in \{0.7, 0.9, 0.98\}$. The variance study uses 20,000 repetitions per case. The coverage study uses 3,000 repetitions per cell for the analytic intervals and the first 300 of them for the bootstrap, so the standard error of a coverage estimate is about 0.4 points for the analytic intervals and 1.3 points for the bootstrap. Classes are balanced ($n_0 = n_1$, with n_1 from 10 to 200) or have 5% positives ($n_0 = 19 \, n_1$, with n_1 from 10 to 100). The sample-size check draws samples at the sizes of Eq. (4) for $\theta = 0.9$ and $h = 0.05$, with 3,000 repetitions. The probability p_1 is simulated with 20,000 repetitions. Real data are the Wisconsin breast cancer set (569 cases; 212 malignant) [13] and the optical digits set (1,797 cases; 180 are the digit 9) [14], from scikit-learn [12]. Three scores are used. For "breast, all features" and "digits 9 vs rest" we use out-of-fold decision values from a logistic regression on standardised features with five folds. For "breast, one weak feature" we use mean smoothness. Each dataset is treated as the population, the population AUC is the AUC of all cases, and each repetition samples n_1 positives and n_0 negatives with replacement, with $n_1 \in \{10, 20, 50\}$ and n_0 equal to n_1 or 9 n_1. Each cell has 2,000 repetitions, and 300 for the bootstrap.

6. Results and Analysis

6.1. The variance of the estimate

Table 1 compares the exact variance of Eq. (2) with the Hanley–McNeil approximation and with simulation. The exact formula matches simulation within 3.3%, which is consistent with the sampling error of a variance estimated from 20,000 repetitions.

Table 1: Variance of the AUC estimate ($\times 10^4$): exact binormal, Hanley–McNeil approximation, and simulation (20,000 repetitions).

θ	n_1	n_0	Exact	H–M	Sim.
0.7	20	20	68.68	69.34	67.95
0.7	20	380	35.24	44.61	34.84
0.7	100	100	13.43	13.57	13.48
0.9	20	20	23.53	26.39	23.70
0.9	20	380	11.85	21.58	11.86
0.9	100	100	4.53	5.12	4.50
0.98	20	20	3.33	5.28	3.44
0.98	20	380	1.59	4.87	1.59
0.98	100	100	0.61	1.02	0.61

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

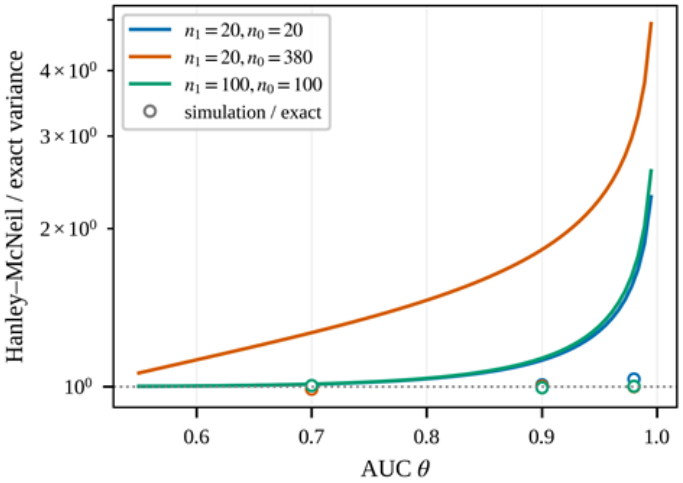

Figure 2: Variance of the AUC estimate. Lines: Hanley–McNeil formula divided by the exact binormal variance. Circles: simulated variance divided by the exact variance (20,000 repetitions).

The Hanley–McNeil approximation is accurate at AUC 0.7 but overstates the variance with the AUC (Figure 2): by 12% at AUC 0.9 with 20 and 20 cases, and by a factor of 3.1 at AUC 0.98 with 20 positives and 380 negatives. A Wald interval that uses it is therefore too wide in that regime, which we see as over-coverage in Section 6.4.

6.2. Extra Negatives

Eq. (3) says that negatives help only up to a point. Figure 3 and Table 2 hold the number of positives at 50, θ = 0.9, and increase the number of negatives. Equal numbers of both classes give a variance 2.04 times its floor. With three negatives per positive it is 1.35 times, with ten it is 1.10, and with twenty 1.05. Beyond ten negatives per positive, more negatives reduce the variance by less than 10%. The simulation agrees with the formula.

Table 2: Variance relative to its limit as $n_0 \to \infty$ ($\theta = 0.9$, $n_1 = 50$): exact, and simulation (8,000 repetitions).

n_0/n_1	Exact	Simulated
1	2.040	2.002
2	1.520	–
3	1.347	1.400
5	1.208	–
10	1.104	1.123
20	1.052	1.062
50	1.021	–
100	1.010	–

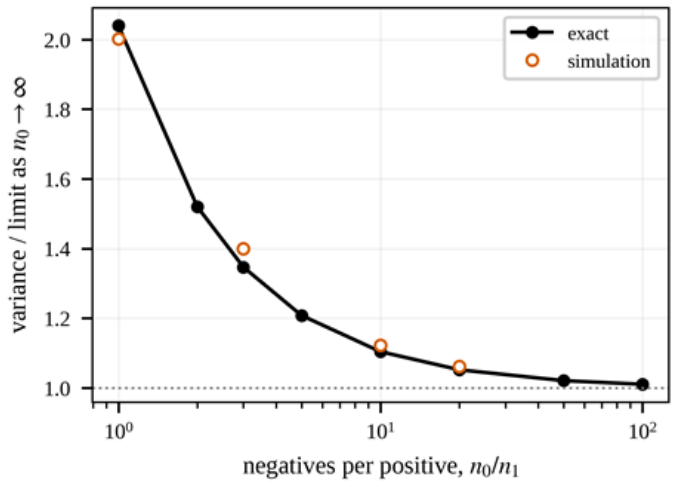

Figure 3: Diminishing returns from extra negatives ($\theta = 0.9$, $n_1 = 50$): variance of the AUC estimate relative to its limit as $n_0 \to \infty$.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

6.3. Sample size and prevalence

Eq. (4) gives the sample size for an interval of half-width 0.05 at θ = 0.9: 138 cases at prevalence 0.5, 383 at 0.1, 725 at 0.05 and 3,478 at 0.01. The 1% case needs 25.2 times as many cases as the balanced one (Figure 4), and in both the number of positives, 69 and 35, is of the same order. Table 3 checks the formula: at these sizes the simulated half-width, from 1.96 times the standard deviation of the estimate, is 0.049 to 0.051, close to the target 0.05. The coverage of the DeLong Wald interval at these sizes, however, is only 91.0% to 93.4%. The formula controls the width, not the coverage; with 35 to 70 positives, the normal approximation is not yet accurate.

Table 3: Total sample size for half-width 0.05 at θ = 0.9 (Eq. 4), with the simulated half-width (1.96 × sd), mean DeLong half-width, and DeLong Wald coverage (3,000 repetitions).

π	n	n_1	Half-width (1.96 sd)	Half-width (DeLong)	Coverage (%)
0.5	138	69	0.051	0.050	93.4
0.1	383	38	0.049	0.050	92.9
0.05	725	36	0.049	0.049	91.8
0.01	3478	35	0.051	0.049	91.0

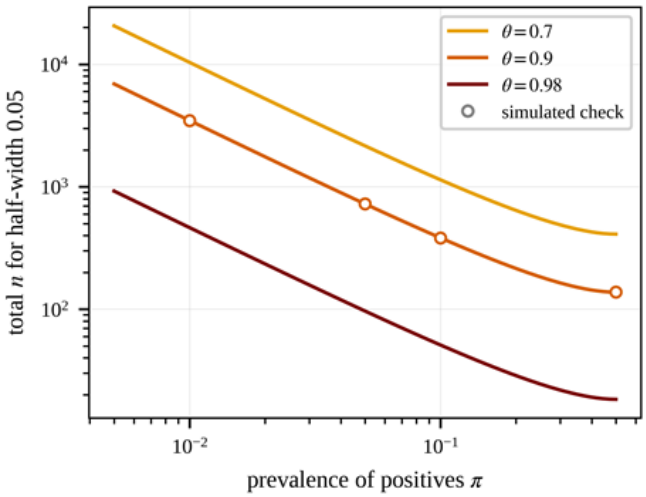

Figure 4: Total sample size needed for a 95% interval of half-width 0.05, Eq. (4), against prevalence. Circles: simulated check at θ = 0.9.

6.4. Coverage of five intervals

Table 4 and Figure 5 give coverage with balanced classes. At AUC 0.7 all five intervals are close to 95% from 20 positives on. At AUC 0.9 with 20 positives and 20 negatives, the logit interval covers 95.0%, the percentile bootstrap 91.0%, and the DeLong and binormal Wald intervals 88.2% and 87.6%. The Hanley–McNeil Wald interval is closer to 95% (90.2%) only because its variance is too large, which is an accident that fails in the rare-positive case: with 5% positives it reaches 98.7% at 50 positives, while the DeLong interval is 92.3%.

Table 4: Coverage (%) of nominal 95% intervals with balanced classes ($n_0 = n_1$).

θ	n_1	Wald DeLong	Wald H–M	Wald binorm.	Logit DeLong	Boot-strap
0.7	20	93.7	93.2	92.8	96.0	97.3
0.7	50	93.7	93.7	93.6	95.0	93.3
0.9	20	88.2	90.2	87.6	95.0	91.0
0.9	50	92.8	94.6	92.6	95.7	93.0
0.98	20	73.7	81.2	73.7	84.7	81.7
0.98	50	86.0	93.3	87.9	93.4	86.7

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

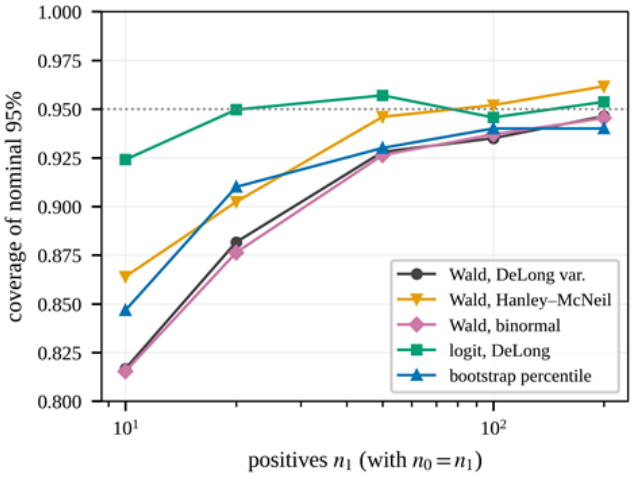

Figure 5: Coverage of nominal 95% intervals at θ = 0.9 with balanced classes. Wald intervals use three variance estimates; the logit interval and the percentile bootstrap are described in Section 3.2.

At AUC 0.98 the picture changes (Figure 6). With balanced classes and 50 positives, coverage is 86.0% for DeLong, 93.4% for the logit interval and 86.7% for the bootstrap. With 5% positives and 20 positives (380 negatives) the five intervals cover 79.2% (DeLong), 86.9% (logit), 78.7% (bootstrap), 87.6% (binormal) and 97.9% (H–M). Even 100 positives with 1,900 negatives leave the DeLong interval at 89.8%. The best performer in these cells is the Hanley–McNeil interval, not because it is good but because it is wide.

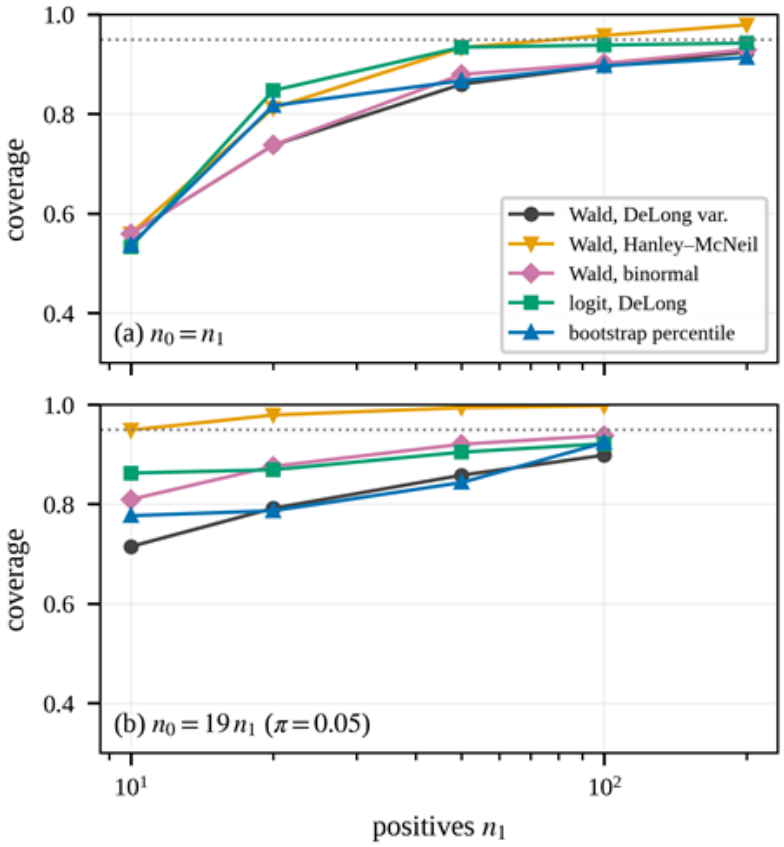

Figure 6: Coverage at θ = 0.98. (a) Balanced classes. (b) 5% positives. Nominal level 0.95 dotted.

6.5. The boundary

Proposition 2 explains the failures. Table 5 compares the closed-form probability p_1 of Eq. (5) with simulation; they agree within 0.006. At AUC 0.98 with 10 positives and 10 negatives, 44% of samples separate the classes perfectly, so no interval of the kinds above can cover more than 56% of the time, and the observed coverage for the best of them is 56%. At AUC 0.995 the bound is 53% with 20 positives and

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

Table 5: Probability p_1 that every positive outscores every negative ($n_0 = n_1 = n$): Eq. (5) and simulation, and the coverage bound $1 - p_1$.

θ	n	p_1 (Eq. 5)	p_1 (sim.)	Bound $1 - p_1$
0.9	10	0.054	0.052	0.946
0.9	20	0.001	0.001	0.999
0.98	10	0.435	0.430	0.565
0.98	20	0.126	0.123	0.874
0.98	50	0.002	0.002	0.998
0.98	100	0.000	0.000	1.000
0.995	10	0.761	0.762	0.239
0.995	20	0.474	0.468	0.526
0.995	50	0.084	0.082	0.916
0.995	100	0.004	0.004	0.996

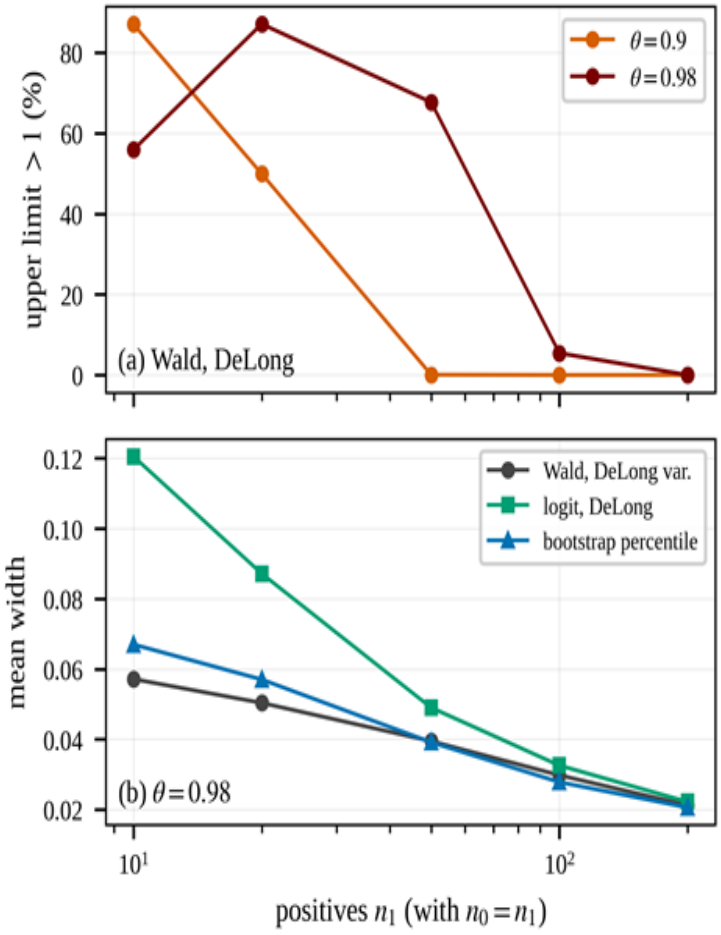

Figure 7: (a) Share of DeLong Wald intervals whose upper limit exceeds 1. (b) Mean interval width at θ = 0.98.

6.6. Real Data

Table 6 and Figure 8 give coverage of the population AUC. For the weak feature (AUC 0.72) all methods are close to 95% from 20 positives on, as at θ = 0.7 in simulation (DeLong 93.9%, logit 94.9%, bootstrap 93.3% at 20 positives and 180 negatives). For the two strong scores (AUC 0.9953 and 0.9955) they are not. With 20 positives and 180 negatives, the DeLong interval covers 45.1% on breast cancer and 62.5% on digits, and the bootstrap 52.0% and 66.7%. The H–M interval is highest, at 67.7% and 84.1%, for the same reason as before. With 50 positives and 450 negatives the best method reaches 85% on breast cancer and 95% on digits. The boundary probability explains this: the simulated p_1 with 20 positives and 180 negatives is 31% on breast cancer and 16% on digits.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

Table 6: Real data, coverage (%) of nominal 95% intervals with nine negatives per positive, and the probability p_1 (%) of perfect separation (2,000 repetitions; 300 for the bootstrap).

Score	n_1	Wald DL	Wald H–M	Wald bin.	Logit DL	Boot.	p_1
b., all features	10	32.1	38.9	38.9	34.2	36.3	60
b., all features	20	45.1	67.7	52.4	57.6	52.0	31
b., all features	50	61.8	85.5	64.3	73.8	63.3	3
b., one weak feature	10	90.5	96.7	93.8	93.7	89.7	–
b., one weak feature	20	93.9	98.0	95.9	94.9	93.3	–
b., one weak feature	50	93.6	98.4	95.9	94.5	92.3	–
digits	10	43.3	53.5	53.5	50.0	44.7	46
digits	20	62.5	84.1	70.7	76.1	66.7	16
digits	50	74.9	95.3	79.8	83.9	78.0	1

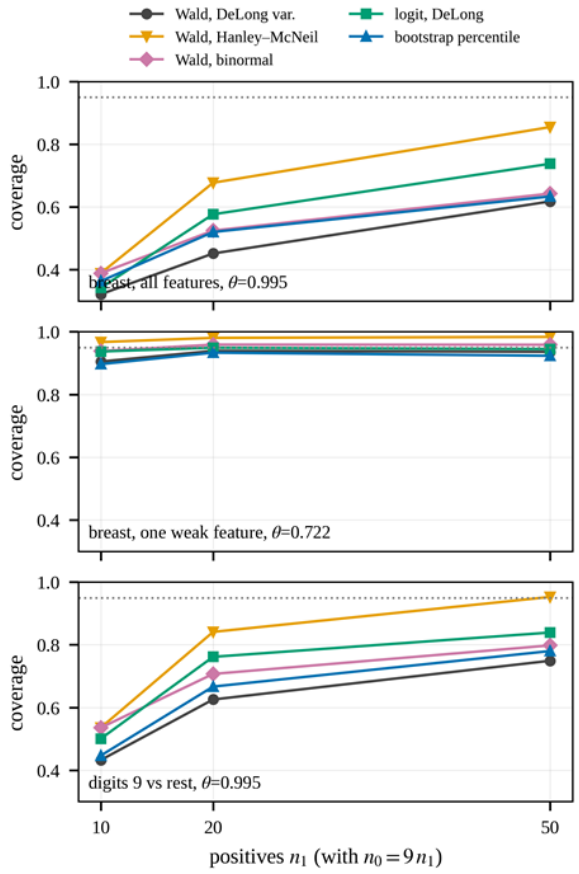

Figure 8: Real data: coverage of the population AUC with nine negatives per positive (2,000 repetitions; 300 for the bootstrap). Nominal level dotted.

6.7. Planning Table

Table 7 turns Eq. (4) into a planning aid. For a half-width of 0.05 the number of positives needed falls from 205 at AUC 0.7 to 9 at AUC 0.98 with balanced classes, because a more separable classifier is estimated more precisely; the smallest values are not credible, since at such sizes perfect separation is likely (Eq. 5) and the large-sample formula does not apply, and halving the half-width to 0.03 multiplies it by about 2.8. With 5% positives, the number of positives is smaller than in the balanced design, since extra negatives help, but the total number of cases is about 5 times larger. Figure 9 shows that the half-width depends mainly on n_1: for the same number of positives, adding 18 negatives per positive shrinks it by at most the factor in Table 2. These sample sizes are for the width; Table 3 and Section 6.4 show that the coverage needs more positives still, and at AUC above 0.95 the boundary probability of Eq. (5) should be checked first.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

Table 7: Sample size for a 95% interval of half-width h (Eq. 4): total cases and positives, balanced and with 5% positives. Entries with fewer than about 30 positives are unreliable (Section 6.5).

θ	h	n (π=0.5)	n_1 (π=0.5)	n (π=0.05)	n_1 (π=0.05)
0.7	0.05	410	205	2,160	108
0.7	0.03	1140	570	6,000	300
0.8	0.05	290	145	1,525	76
0.8	0.03	805	403	4,237	212
0.9	0.05	138	69	725	36
0.9	0.03	382	191	2,013	101
0.95	0.05	60	30	313	16
0.95	0.03	165	83	871	44
0.98	0.05	18	9	97	5
0.98	0.03	51	26	268	13

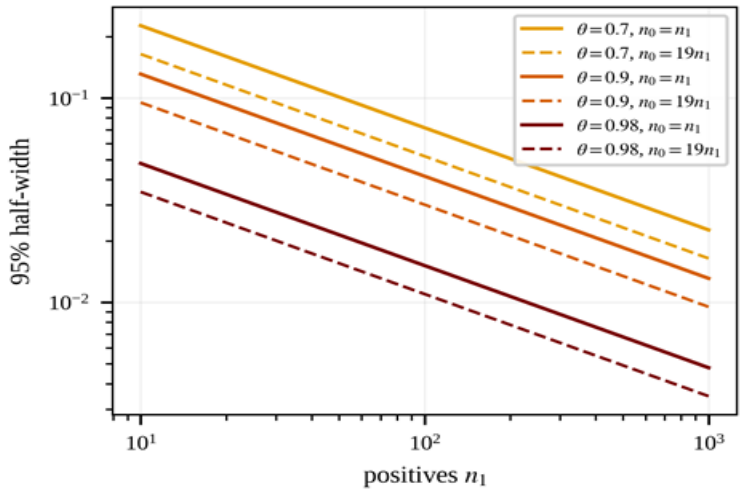

Figure. 9 Half-width of a 95% interval from Eq. (3) as a function of the number of positives, for balanced classes (solid) and 19 negatives per positive (dashed).

6.8. A Practical Recipe

The results suggest a short procedure. First, count positives, not cases. For a half-width of 0.05 at AUC 0.9 about 35 to 70 positives are needed, whatever the prevalence (Table 3), and more negatives than ten per positive add almost nothing (Eq. 3, Corollary 2). Second, use Eq. (4) and Table 7 to plan a study, and treat it as a lower bound on width, not a guarantee of coverage. Third, estimate the probability of perfect separation (Eq. 5) before trusting any interval: if it is above a few percent, the interval is unreliable by Proposition 2. Fourth, prefer the DeLong-based logit interval or a bootstrap to the Wald interval, and do not rely on the Hanley–McNeil variance when the AUC is high. Fifth, when positives are few and the AUC is near one, report the point estimate with the number of positives and say that an interval is not reliable, or use a model-based or exact method that does not collapse at the boundary.

Discussions and Limitations

What we think carries over. Three points seem general. The variance of the AUC estimate has a floor set by the number of positives, so sample size should be planned in positives. Perfect separation by chance is a real event at high AUC, and it limits every interval that is built from the data alone. And an approximation that over-estimates the variance can look like a better interval simply because it over-covers.

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

Synthetic setting. The binormal model has equal variances and Gaussian scores. Real scores are skewed and bounded, which changes Q_1 and Q_2 but not the form of Eq. (2). We did not study unequal variances, ties, or clustered data. The coverage figures have simulation error of 0.4 points (analytic) and 1.3 points (bootstrap).

Real data. We treated each dataset as the population and sampled from it with replacement, which makes the true AUC known but gives the sampling distribution a discrete support, and the AUC values are close to 1. We used logistic regression scores and one weak feature, and did not study other models or other datasets. We did not study score-based, exact or Bayesian intervals for the AUC, or bias-corrected bootstrap variants, which might behave better at the boundary.

7. Conclusion

The uncertainty of an AUC estimate is set by the number of positives. In the binormal model the exact variance has a closed form, extra negatives beyond about ten per positive add little, and a 95% interval of half-width 0.05 at AUC 0.9 needs 138 cases with balanced classes but 3,478 with 1% positives. Near AUC one, a finite sample can separate the classes perfectly, which caps the coverage of the usual intervals by $1 - p_1$, and in our simulations and real data the coverage fell to between 63% and 84% at 20 positives. Next steps are to study intervals that do not collapse at the boundary, to relax the binormal assumption, and to extend the sample-size rule to comparisons between two classifiers.

References

[1]. J. A. Hanley, B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

[2]. E. R. DeLong, D. M. DeLong, D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[3]. D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychology*, 12(4):387–415, 1975.

[4]. H. B. Mann, D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18(1):50–60, 1947.

[5]. W. Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.*, 19(3):293–325, 1948.

[6]. B. Efron, R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[7]. C. Cortes, M. Mohri. Confidence intervals for the area under the ROC curve. In *NIPS*, 2004.

[8]. T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[9]. X. Robin et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77, 2011.

[10]. L. D. Brown, T. T. Cai, A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001.

[11]. D. M. Green, J. A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, 1966.

[12]. F. Pedregosa et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[13]. W. N. Street, W. H. Wolberg, O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *SPIE*, vol. 1905, 1993.

[14]. E. Alpaydin, C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998.

[15]. A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

A. Derivations

Eq. (2). Write $\hat{\theta}$ as the average of the indicators $1[X_i > Y_j]$. Two indicators that share a positive have covariance $Q_2 - \theta^2$, those that share a negative have covariance $Q_1 - \theta^2$, those that share both have variance $\theta(1 - \theta)$, and others are independent. Counting the pairs of indicators of each type gives Eq. (2).

Eq. (3). Divide the terms of Eq. (2) by $n_1 n_0$ and drop the terms that are of lower order, namely $\theta(1 - \theta)/(n_1 n_0)$ and the corrections of -1.

Eq. (4). Set $n_1 = \pi n$ and $n_0 = (1 - \pi)n$ in Eq. (3) and solve $1.96\sqrt{\mathrm{Var}(\hat{\theta})} = h$ for n.

Eq. (5). See the proof of Proposition 2. The integral is evaluated numerically on $[-10, 12]$.

B. Parameters

Synthetic: $\theta \in \{0.7, 0.9, 0.98\}$; variance study 20,000 repetitions with $(n_1, n_0) \in \{(20, 20), (20, 380), (100, 100)\}$; negatives study $n_1 = 50$ with n_0/n_1 from 1 to 100; sample-size check 3,000 repetitions; coverage study 3,000 repetitions per cell, balanced $n_1 \in \{10, 20, 50, 100, 200\}$ and $n_0 = 19\, n_1$ with $n_1 \in \{10, 20, 50, 100\}$; bootstrap 200 stratified resamples on the first 300 repetitions; p_1 with 20,000 repetitions. The binormal interval uses a 400-point lookup of the two integrals on θ from 0.5001 to 0.99999. Real data: logistic regression ($C = 1$) on standardised features with five-fold out-of-fold decision values; population AUC computed on all cases; $n_1 \in \{10, 20, 50\}$; $n_0 \in \{n_1, 9\, n_1\}$; 2,000 repetitions, 300 for the bootstrap. All figures were produced by code at 600 dpi and are printed at their stored width. Code, results and figures are in the supplementary archive.

How To Cite:

Sruthi Pushadapu , " Computational and Statistical Reliability of Machine Learning Performance Metrics: AUC Variance and Confidence Interval Failure
", International Journal of Research and Development in Engineering Sciences(IJRDES), Volume 8 , Issue 1 , pp 20 – 30 , January 2026.
CrossRef DOI: https://doi.org/10.63328/ijrdes-v8ri1p3

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 2582 - 4201(Online) , http://www.ijrdes.com/ , IJRDES, Vol. 8, Issue 6 , 2026 , https://doi.org/10.63328/IJRDES-V8RI6P3

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJRDES , All Rights Reserved
www.jacksparrowpublishers.com