


# Computation of the distribution of model accuracy statistics in machine learning: Comparison between analytically derived distributions and simulation-based methods

Alexander A. Huang<sup>1</sup> | Samuel Y. Huang<sup>2</sup> 

<sup>1</sup>Northwestern University Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA

<sup>2</sup>Virginia Commonwealth School of Medicine, Virginia Commonwealth University, Richmond, Virginia, USA

## Correspondence

Samuel Y. Huang, Virginia Commonwealth School of Medicine, Virginia Commonwealth University, Richmond, VA, USA.

Email: [huangs8@vcu.edu](mailto:huangs8@vcu.edu)

## Abstract

**Background and Aims:** All fields have seen an increase in machine-learning techniques. To accurately evaluate the efficacy of novel modeling methods, it is necessary to conduct a critical evaluation of the utilized model metrics, such as sensitivity, specificity, and area under the receiver operator characteristic curve (AUROC). For commonly used model metrics, we proposed the use of analytically derived distributions (ADDs) and compared it with simulation-based approaches.

**Methods:** A retrospective cohort study was conducted using the England National Health Services Heart Disease Prediction Cohort. Four machine learning models (XGBoost, Random Forest, Artificial Neural Network, and Adaptive Boost) were used. The distribution of the model metrics and covariate gain statistics were empirically derived using boot-strap simulation ( $N = 10,000$ ). The ADDs were created from analytic formulas from the covariates to describe the distribution of the model metrics and compared with those of bootstrap simulation.

**Results:** XGBoost had the most optimal model having the highest AUROC and the highest aggregate score considering six other model metrics. Based on the Anderson–Darling test, the distribution of the model metrics created from bootstrap did not significantly deviate from a normal distribution. The variance created from the ADD led to smaller SDs than those derived from bootstrap simulation, whereas the rest of the distribution remained not statistically significantly different.

**Conclusions:** ADD allows for cross study comparison of model metrics, which is usually done with bootstrapping that rely on simulations, which cannot be replicated by the reader.

## KEYWORDS

Anderson–Darling, bootstrap, Gaussian distribution, normal distribution, simulation, sufficient statistics, variance calculations, Whitney–Mann

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Health Science Reports* published by Wiley Periodicals LLC.

## 1 | INTRODUCTION

All fields, from computer science to medicine, have significantly increased their use of machine learning algorithms.<sup>1-7</sup> These algorithms are unique in that they can predict data without explicit instructions from the modeler.<sup>1,5,8,9</sup> XGBoost and Random Forest, two well-known machine learning algorithms, have been shown to be significantly more accurate than linear and logistic regression.<sup>4,10,11</sup> However, in the case of machine learning (ML) algorithms, decreased interpretability comes at the cost of increased predictive accuracy, and ML algorithms are frequently referred to as “black boxes” because they cannot be understood.<sup>1,5,7-11</sup> When evaluating these methods, researchers heavily rely on model metrics like area under the receiver operator characteristic curve (AUROC), sensitivity, specificity, and accuracy.<sup>1,2,4</sup>

Bootstrap is commonly the method of choice to calculate the distribution of these model metrics.<sup>2,12-15</sup> As a simulation-based method, a reader cannot rerun the simulation the same way they can back-calculate from a computed *t* test.<sup>16-20</sup> It was necessary to develop analytically derived distributions (ADDs) to accurately summarize the distribution of these model metrics the same way a Gaussian distribution summarizes the mean and SD for accurate comparison of models within and between studies.

This aim of the study is to compare ADD created from formulas in the statistical literature from those derived from bootstrapping the distribution of model statistics.

## 2 | METHODS

The Heart Disease Prediction cohort from the England National Health Services database was used in this retrospective cohort study.<sup>21-23</sup> All methods in this research were carried out in accordance with guidelines detailed by the Data Alliance Partnership Board-approved national information standards and data collections for use in health and adult social care. This code was written using R version 4.2.2. We used the following packages readxl and foreign for reading in the data set, MLDataR for the data set, dplyr and ggplot2 for datavisualization, xgboost, farff, tibble, lpspline, and pROC for model creation.

### 2.1 | Model metrics of interest

The model metrics were selected due to their prevalence in the literature and included the AUROC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1, accuracy, and balanced accuracy. Metrics of feature importance assessed in this study included the gain, cover, and frequency.

### 2.2 | Independent variables

Demographic covariates included age and sex. Clinical covariates included resting blood pressure, fasting blood sugar, cholesterol,

resting electrocardiogram (ECG), presence of angina, and maximum heart rate.

### 2.3 | Dependent variable

The dependent variable of interest was a clinician's diagnosis of heart disease.

### 2.4 | Model construction and statistical analysis

Descriptive statistics for all patients and then patients stratified by heart disease were computed for all covariates and compared using  $\chi^2$  tests for categorical variables and *t* tests for continuous variables. Machine learning methods including XGBoost, Random Forest, Artificial Neural Network, and Adaptive Boosting were implemented on the data set.

### 2.5 | Bootstrap simulation and ADD compared via distribution of model metrics

#### 2.5.1 | Distribution evaluation

The distribution of each of the statistics was evaluated through comparison of summary statistics (minimum, 5th percentile, 25th percentile, 50th percentile, 75th percentile, 95th percentile, maximum, mean, SD) and the Anderson-Darling test for normality.

### 2.6 | Bootstrap simulation

A train-test set (70:30) was used within all machine-learning models in this study. Bootstrap simulation ( $N = 10,000$ ) simulations were carried out by permuting the train-test sets before training.

### 2.7 | Calculation of variance with analytical formulas to create the ADD

#### 2.7.1 | AUROC

The AUROC =  $\frac{U}{n^2}$  where  $U$  has the Mann-Whitney distribution:

$$U = \sum_{i=1}^n \sum_{j=1}^m F(X_i, Y_j), \text{ where } F(X_i, Y_j) = \begin{cases} 1 & X > Y \\ 1 & X = Y, \text{ and } X_1, \dots, X_n \\ 0 & X < Y \end{cases}$$

&  $Y_1, \dots, Y_m$  are individually identically distributed. As  $U$  has the Mann-Whitney distribution, we observe that the variance of the  $U$  distribution is  $\sigma_U^2 = \frac{n^2(2n+1)}{12}$ . Thus, as AUROC =  $\frac{U}{n^2}$ ,  $\sigma_{\text{AUROC}}^2 = \frac{\sigma_U^2}{(n^2)^2} = \frac{\sigma_U^2}{n^4} = \frac{n^2(2n+1)}{12n^4} = \frac{(2n+1)}{12n^2}$ . As the Mann-Whitney

distribution is asymptotically convergent on the Gaussian distribution at large sample sizes, the mean and SD are sufficient statistics. We further observe that for large  $n$ , the variance formula for the AUROC can be approximated as:  $\sigma_{\text{AUROC}}^2 = \frac{(2n+1)}{12n^2} \rightarrow \frac{(2n)}{12n^2} = \frac{1}{6n}$ . Furthermore, another more nuanced measurement for the variability can also be dependent on the value of AUROC itself and approximating it as a proportion yields similar approximation to the Mann-Whitney distribution for when the values of AUROC are between 0.7 and 0.9. Thus, another similarly correct analytic approximation of the AUROC is  $\sigma_{\text{AUROC}}^2 = \frac{(\text{AUROC})(1 - \text{AUROC})}{n}$ .

## 2.8 | Model metrics that can be evaluated as proportions

Multiple literature sources have treated accuracy, F1, sensitivity, specificity, PPV, and NPV as similar to proportions. Thus, we make the assumption from treating these statistics as proportions that their variance follows from the analytic formula:  $\sigma_p^2 = \frac{p(1-p)}{n}$ , where  $p$  is the proportion. Thus, the SDs for the model metrics are: accuracy:  $\sigma_{\text{Accuracy}}^2 = \frac{(\text{Accuracy})(1 - \text{Accuracy})}{n}$ , F1:  $\sigma_{\text{F1}}^2 = \frac{(\text{F1})(1 - \text{F1})}{n}$ , Sensitivity:  $\sigma_{\text{Sensitivity}}^2 = \frac{(\text{Sensitivity})(1 - \text{Sensitivity})}{n}$ .

**TABLE 1a** Summary of model metrics.

	Metrics	Minimum	5th Percentile	25th Percentile	Median	75th Percentile	95th Percentile	Maximum	Mean	SD	Range
<b>XGBoost</b>	Accuracy	0.684	0.741	0.766	0.790	0.806	0.836	0.898	0.789	0.026	0.210
	F1	0.686	0.750	0.774	0.784	0.810	0.835	0.897	0.787	0.031	0.204
	Sensitivity	0.680	0.757	0.790	0.806	0.821	0.853	0.901	0.802	0.026	0.224
	Specificity	0.592	0.708	0.749	0.786	0.815	0.852	0.947	0.789	0.037	0.348
	PPV	0.680	0.761	0.787	0.818	0.847	0.884	0.958	0.818	0.035	0.273
	NPV	0.567	0.676	0.722	0.753	0.785	0.829	0.930	0.761	0.046	0.354
	AUROC	0.772	0.831	0.856	0.867	0.884	0.903	0.948	0.868	0.025	0.171
<b>Random Forest</b>	Accuracy	0.675	0.729	0.771	0.778	0.801	0.812	0.892	0.784	0.027	0.224
	F1	0.687	0.740	0.771	0.776	0.809	0.816	0.884	0.785	0.030	0.201
	Sensitivity	0.665	0.745	0.782	0.799	0.804	0.847	0.895	0.793	0.024	0.229
	Specificity	0.584	0.709	0.748	0.784	0.803	0.845	0.927	0.771	0.041	0.340
	PPV	0.676	0.740	0.779	0.813	0.846	0.857	0.948	0.810	0.045	0.270
	NPV	0.555	0.661	0.720	0.736	0.772	0.826	0.908	0.750	0.045	0.359
	AUROC	0.757	0.824	0.842	0.860	0.887	0.900	0.928	0.857	0.022	0.175
<b>Artificial Neural Network</b>	Accuracy	0.689	0.736	0.761	0.786	0.805	0.830	0.877	0.781	0.021	0.194
	F1	0.677	0.732	0.750	0.783	0.790	0.818	0.888	0.774	0.027	0.211
	Sensitivity	0.672	0.749	0.779	0.794	0.802	0.834	0.884	0.796	0.021	0.216
	Specificity	0.591	0.707	0.749	0.768	0.799	0.835	0.928	0.768	0.035	0.327
	PPV	0.659	0.748	0.780	0.809	0.835	0.859	0.940	0.808	0.029	0.274
	NPV	0.550	0.665	0.718	0.752	0.772	0.817	0.912	0.749	0.047	0.361
	AUROC	0.751	0.821	0.839	0.866	0.882	0.891	0.949	0.847	0.027	0.192
<b>Adaptive Boosting</b>	Accuracy	0.683	0.731	0.761	0.790	0.793	0.821	0.885	0.775	0.023	0.199
	F1	0.674	0.739	0.760	0.774	0.801	0.828	0.890	0.775	0.029	0.224
	Sensitivity	0.671	0.753	0.783	0.811	0.809	0.839	0.889	0.797	0.019	0.216
	Specificity	0.585	0.694	0.746	0.777	0.803	0.853	0.941	0.772	0.045	0.354
	PPV	0.676	0.742	0.772	0.805	0.843	0.861	0.951	0.817	0.045	0.277
	NPV	0.567	0.664	0.717	0.751	0.784	0.825	0.929	0.750	0.047	0.358
	AUROC	0.755	0.815	0.840	0.860	0.865	0.894	0.929	0.862	0.025	0.175

Note: Summary of model metrics within the test set for each of the four machine learning techniques (XGBoost, Random Forest, Artificial Neural Network, and Adaptive Boosting) based upon bootstrap simulation.

Abbreviations: AUROC, area under the receiver operator characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

Specificity:  $\sigma_{\text{Specificity}}^2 = \frac{(\text{Specificity})(1 - \text{Specificity})}{n}$ , PPV:  $\sigma_{\text{PPV}}^2 = \frac{(\text{PPV})(1 - \text{PPV})}{n}$ , and NPV:  $\sigma_{\text{NPV}}^2 = \frac{(\text{NPV})(1 - \text{NPV})}{n}$ .

### 3 | RESULTS

Table 1a shows the model metrics of the four machine learning models calculated for accuracy, F1, sensitivity, specificity, PPV, NPV, and AUROC utilizing the bootstrap method. Table 1b shows the

model metrics of the four machine learning models calculated for accuracy, F1, sensitivity, specificity, PPV, NPV, and AUROC utilizing ADD. The distributions of the model metrics for the bootstrap method and ADD are approximately similar.

Table 2a shows the bootstrapped distribution for model feature importance statistics for each covariate for the selected XGBoost model. Table 2b shows the ADD for model feature importance statistics for each covariate for the selected XGBoost model. Again the values between the bootstrap simulation distribution and the ADD are very similar across the minimum, 5th percentile, 25th

**TABLE 1b** Summary of model metrics for each of the four machine learning techniques.

	Metrics	Minimum	5th Percentile	25th Percentile	Median	75th Percentile	95th Percentile	Maximum	Mean	SD	Range
<b>XGBoost</b>	Accuracy	0.684	0.751	0.773	0.789	0.805	0.828	0.898	0.789	0.024	0.215
	F1	0.686	0.748	0.771	0.787	0.803	0.826	0.897	0.787	0.024	0.210
	Sensitivity	0.680	0.764	0.787	0.802	0.818	0.840	0.901	0.802	0.023	0.222
	Specificity	0.592	0.750	0.773	0.789	0.805	0.827	0.947	0.789	0.024	0.354
	PPV	0.680	0.782	0.803	0.818	0.833	0.855	0.958	0.818	0.022	0.277
	NPV	0.567	0.720	0.744	0.761	0.777	0.801	0.930	0.761	0.025	0.363
	AUROC	0.772	0.836	0.855	0.868	0.881	0.900	0.948	0.868	0.020	0.176
<b>Random Forest</b>	Accuracy	0.675	0.744	0.768	0.784	0.800	0.823	0.892	0.784	0.024	0.216
	F1	0.687	0.746	0.769	0.785	0.801	0.824	0.884	0.785	0.024	0.198
	Sensitivity	0.665	0.755	0.777	0.793	0.809	0.832	0.895	0.793	0.023	0.229
	Specificity	0.584	0.731	0.754	0.771	0.787	0.811	0.927	0.771	0.024	0.344
	PPV	0.676	0.773	0.795	0.810	0.826	0.848	0.948	0.810	0.023	0.271
	NPV	0.555	0.709	0.733	0.750	0.767	0.791	0.908	0.750	0.025	0.354
	AUROC	0.757	0.823	0.843	0.857	0.870	0.890	0.928	0.857	0.020	0.171
<b>Artificial Neural Network</b>	Accuracy	0.689	0.742	0.765	0.781	0.797	0.820	0.877	0.781	0.024	0.188
	F1	0.677	0.735	0.758	0.774	0.791	0.814	0.888	0.774	0.024	0.211
	Sensitivity	0.672	0.757	0.780	0.796	0.811	0.834	0.884	0.796	0.023	0.212
	Specificity	0.591	0.728	0.752	0.768	0.785	0.808	0.928	0.768	0.024	0.337
	PPV	0.659	0.771	0.793	0.808	0.824	0.846	0.940	0.808	0.023	0.281
	NPV	0.550	0.708	0.732	0.749	0.766	0.790	0.912	0.749	0.025	0.361
	AUROC	0.751	0.812	0.833	0.847	0.861	0.881	0.949	0.847	0.021	0.198
<b>Adaptive Boosting</b>	Accuracy	0.683	0.736	0.759	0.775	0.791	0.815	0.885	0.775	0.024	0.202
	F1	0.674	0.735	0.758	0.775	0.791	0.814	0.890	0.775	0.024	0.216
	Sensitivity	0.671	0.759	0.781	0.797	0.813	0.835	0.889	0.797	0.023	0.217
	Specificity	0.585	0.732	0.756	0.772	0.789	0.812	0.941	0.772	0.024	0.356
	PPV	0.676	0.780	0.802	0.817	0.832	0.853	0.951	0.817	0.022	0.274
	NPV	0.567	0.709	0.733	0.750	0.767	0.791	0.929	0.750	0.025	0.362
	AUROC	0.755	0.829	0.848	0.862	0.875	0.895	0.929	0.862	0.020	0.175

Note: Summary of model metrics for each of the four machine learning techniques (XGBoost, Random Forest, Artificial Neural Network, and Adaptive Boosting) based upon the derived distribution using analytic formulas described within the study.

Abbreviations: AUROC, area under the receiver operator characteristic curve; NPV, negative predictive value; PPV, positive predictive value.

**TABLE 2a** For the XGBoost model, a summary of model gain statistics for each covariate in the model based on bootstrap simulation.

Covariates	Minimum	5th Percentile	25th Percentile	Median	75th Percentile	95th Percentile	Maximum	Mean	SD	Range
Angina	0.225	0.288	0.316	0.334	0.0353	0.383	0.456	0.335	0.029	0.231
Cholesterol	0.148	0.209	0.228	0.24	0.252	0.269	0.326	0.24	0.018	0.178
Maximum heart rate	0.081	0.114	0.129	0.139	0.15	0.165	0.201	0.139	0.015	0.12
Age	0.059	0.082	0.095	0.103	0.112	0.124	0.156	0.103	0.013	0.097
Resting blood pressure	0.027	0.051	0.061	0.069	0.076	0.087	0.109	0.069	0.011	0.082
Sex	0.026	0.038	0.044	0.049	0.054	0.062	0.082	0.049	0.007	0.056
Fasting blood sugar	0.007	0.029	0.037	0.043	0.05	0.063	0.142	0.044	0.011	0.135
RestingECG	0.003	0.012	0.017	0.02	0.024	0.029	0.043	0.02	0.005	0.04

Abbreviation: ECG, electrocardiogram.

**TABLE 2b** For the XGBoost model, a summary of model gain statistics for each covariate in the model based on analytical formulas described within this study.

Covariates gain statistic	Minimum	5th Percentile	25th Percentile	Median	75th Percentile	95th Percentile	Maximum	Mean	SD	Range
Angina	0.225	0.290	0.317	0.335	0.353	0.380	0.456	0.335	0.027	0.231
Cholesterol	0.148	0.199	0.223	0.240	0.257	0.281	0.326	0.24	0.025	0.178
Maximum heart rate	0.081	0.106	0.126	0.139	0.152	0.172	0.201	0.139	0.020	0.12
Age	0.059	0.074	0.091	0.103	0.115	0.132	0.156	0.103	0.018	0.097
Resting blood pressure	0.027	0.045	0.059	0.069	0.079	0.093	0.109	0.069	0.015	0.082
Sex	0.026	0.028	0.041	0.049	0.057	0.070	0.082	0.049	0.012	0.056
Fasting blood sugar	0.007	0.025	0.036	0.044	0.052	0.063	0.142	0.044	0.012	0.135
RestingECG	0.003	0.007	0.015	0.020	0.025	0.033	0.043	0.02	0.008	0.04

Abbreviation: ECG, electrocardiogram.

percentile, median, 75th percentile, 95th percentile, maximum, mean, and range. The SEs for the model statistics and for the gain statistics were significantly less variable from the ADD.

The Anderson–Darling test was completed to validate whether the point estimate for the mean and SD are sufficient to approximate the full model distribution are reported in Table 3. The bootstrap distribution for the model metrics and the feature gain statistics were not significantly different than a normal distribution.

Figure 1 shows the bootstrapped values for the model metrics (Balanced accuracy, Accuracy, F1, Sensitivity, Specificity, PPV, NPV, AUROC) for the XGBoost model.

Figure 2 shows the bootstrapped distribution of gain statistics calculated for covariates that included Age, Angina, Cholesterol, Fasting blood sugar, Maximum heart rate, Resting blood pressure, RestingECG, and Sex.

Figures 1 and 2 validates the observations of the Anderson–Darling test demonstrating no significant difference between the bootstrapped distribution of the model metrics and feature gain statistics from a normal distribution.

## 4 | DISCUSSION

We observed that the model metrics and model feature importance statistics for machine learning models converged on a Gaussian distribution in this retrospective, cross-sectional cohort of heart disease patients. ADDs were used to calculate sufficient Gaussian distribution statistics, including the mean and SD. It was found that there was no significant difference in the overall distribution between the Gaussian approximation of the distribution for model metrics and feature importance statistics.

Bootstrapping has previously been the primary method used to derive accuracy statistics for machine learning model distributions.<sup>18,20,24,25</sup> Bootstrapping can generate a distribution based on data without any knowledge of the distribution and without violating any assumptions that are required to utilize a distribution for inference.<sup>16,17,26–29</sup> As a result, the vast majority of packages focus upon bootstrapping and thus so too do the vast majority of studies.<sup>1,8,30–32</sup> Due to the increased computational power, these nonparametric methods can be completed efficiently and are

A		B	
Model metrics	Anderson-Darling $p$	Gain statistics	Anderson-Darling $p$
Balanced accuracy	0.53	Angina	0.23
Accuracy	0.44	Cholesterol	0.46
F1	0.46	Maximum heart rate	0.3
Sensitivity	0.18	Age	0.27
Specificity	0.36	Resting blood pressure	0.7
PPV	0.22	Sex	0.18
NPV	0.97	Fasting blood sugar	0.99
AUROC	0.64	RestingECG	0.1

TABLE 3 For the XGBoost models.

Note: (A) Summary of Anderson-Darling test for normality for model metrics. (B) Summary of Anderson-Darling test for normality for gain statistics for model covariates.

Abbreviations: AUROC, area under the receiver operator characteristic curve; NPV, negative predictive value; PPV, positive predictive value; RestingECG, resting electrocardiogram.

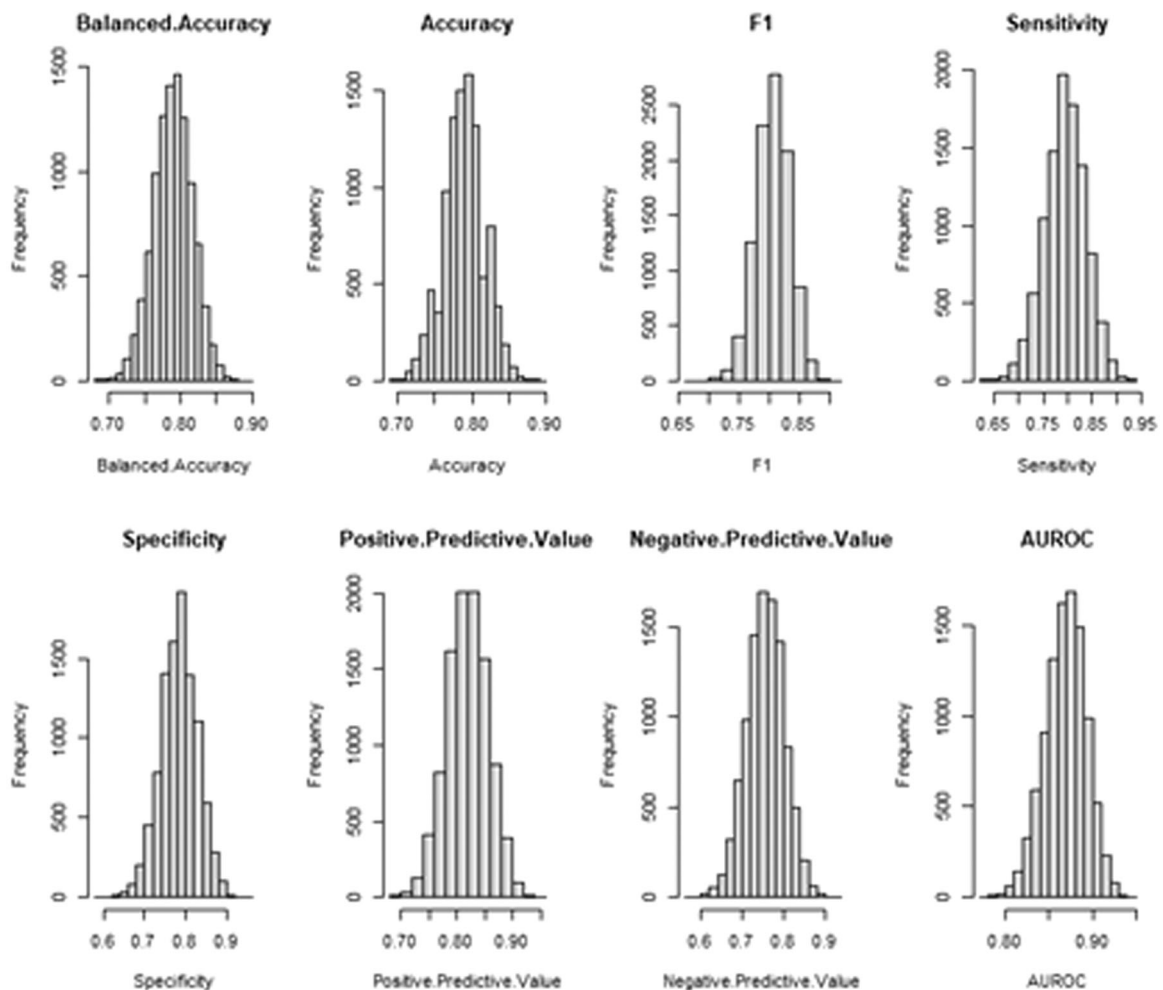
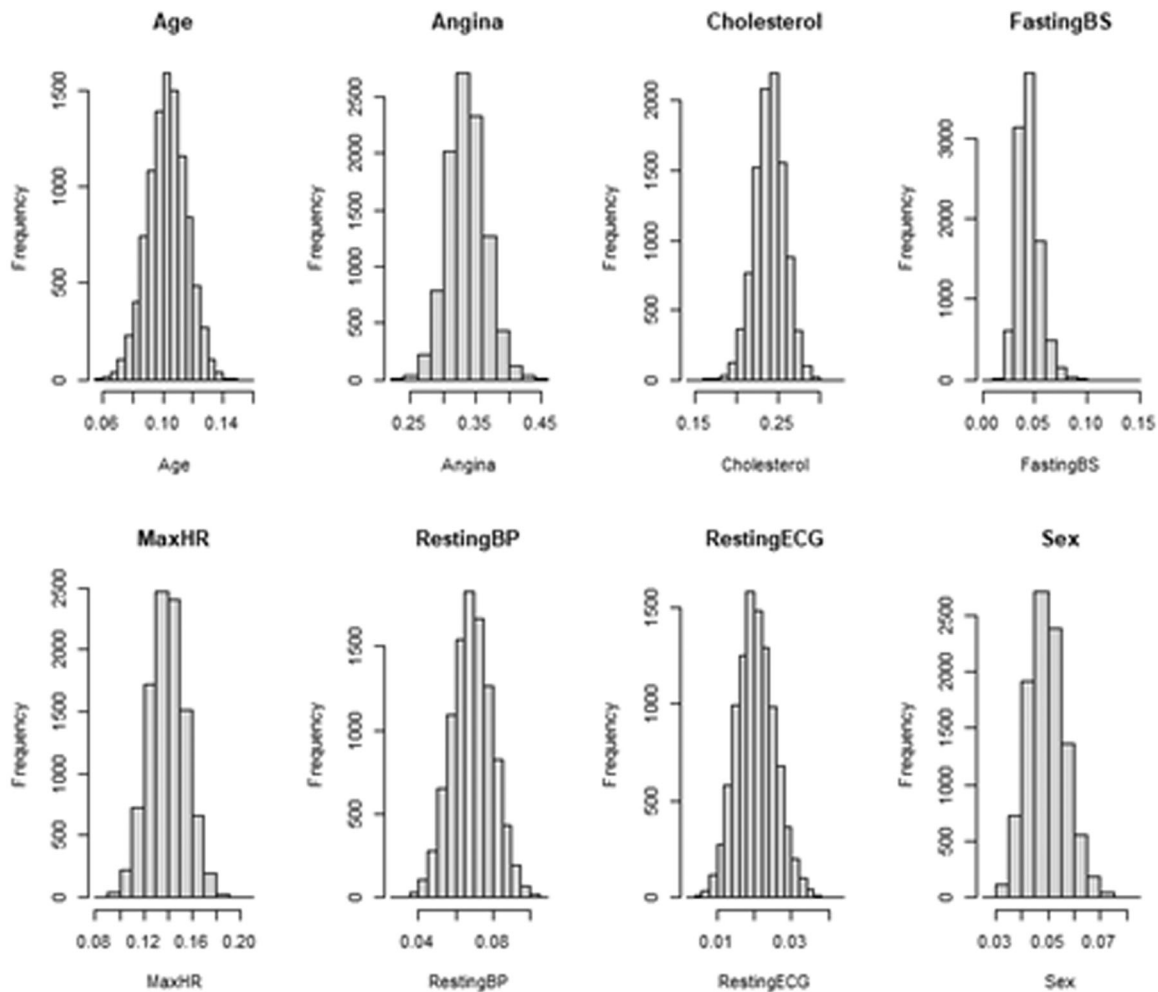


FIGURE 1 Balanced accuracy, Accuracy, F1, Sensitivity, Specificity, Positive predictive value, Negative predictive value, and area under the receiver operator characteristic curve (AUROC) for the XGboost model following bootstrap simulation.



**FIGURE 2** For the XGBoost models, the distribution of the gain statistic for all covariates: Age, Angina, Cholesterol, Fasting blood sugar (Fasting BS), Maximum heart rate (MaxHR), Resting blood pressure (RestingBP), Resting electrocardiogram (RestingECG), and Sex.

especially useful for distributions that cannot be quantified analytically.<sup>1,3,5,22,33-35</sup> The fact that bootstrapping relies on simulation is its weakness, making it difficult to replicate in other studies. Furthermore, due to differences in simulation methodology, it may be difficult to compare the results of the simulation if the results of the simulation are not summarized identically.

If a researcher wants to replicate the distribution of a study, they cannot replicate a bootstrapped distribution; however, with sufficient statistics of a random value distribution, the distribution can be exactly generated.<sup>36</sup> Thus, identification of whether the results of the simulation-based methods can be approximated with well-established random variable distributions and their sufficient statistics computed will effectively allow comparisons of models within and between studies.

What our study uniquely contributes to the medical and biostatistics literature is a rigorous comparison between bootstrap simulations and a distribution generated from a Gaussian distribution using the point-estimate for the mean and SDs (ADD). The use of Anderson–Darling methodology to evaluate bootstrap distributions to test mortality and validate visual

judgements of histograms for the normal distribution allow for strong evidence that regardless of the skewed accuracy results that come from potential imbalances of data sets and machine learning methods, the overall distribution of model metrics follow a Gaussian distribution.

The study's findings can be broadly applied to research on machine learning. To begin, they can be persuaded to employ a variety of machine-learning techniques and choose the most effective one, rather than relying solely on a single point estimate. Instead, a thorough evaluation of the estimate variances for the model metrics can be used to accurately determine which model is the most effective. As a result, we advocate that the strongest model is not only the one with the highest AUROC point estimate on a randomly selected seed but also the one with the highest distribution of multiple model accuracy statistics.<sup>16,17,24,25,37-39</sup> Furthermore, the results of this study support that the distribution of each model metric follows a normal distribution and can be modeled analytically through the Gaussian distribution and the Whitney–Mann distribution for the AUROC, which we have termed the ADD pronounced the “AD distribution.”

## 4.1 | Limitations

This study has several strengths and weaknesses. The study utilizes data from onlyappone cohort and thus may be difficult to generalize to other populations. However, as the goal was to evaluate methods to compute the variance of machine learning model statistics instead of developing models for heart disease, this is less of a concern. In addition, this study's replicability is enhanced by making use of a publicly accessible data set that is already integrated into an R package, which is in line with the paper's general recommendations. In addition, to acquire a better comprehension of the distribution of the model metrics and feature importance statistics computed from machine learning methods, subsequent studies will need to validate this approach with additional cohorts, both smaller and larger in size.

## 5 | CONCLUSION

The distribution of model metrics and feature importance measures can be summarized by making use of the Gaussian distribution and the adequate statistics of mean and SD. Based on point estimates for the model metrics, there is no significant difference between the bootstrap distribution and the Gaussian distribution. Further retrospective and prospective cohort studies utilizing the model is needed to verify the conclusion.

### AUTHOR CONTRIBUTIONS

**Alexander A. Huang:** Conceptualization; formal analysis; investigation; methodology; resources; software; supervision; visualization; writing—original draft; writing—review & editing.  
**Samuel Y. Huang:** Conceptualization; formal analysis; investigation; methodology; resources; validation; writing—original draft; writing—review & editing. All authors have read and approved the final version of the manuscript. Corresponding author had full access to all of the data in this study and takes complete responsibility for the integrity of the data and the accuracy of the data analysis.

### CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

### DATA AVAILABILITY STATEMENT

The data sets generated and analyzed within this study are available through the national health services R community at <https://nhsrcommunity.com/> and through the MLDataR package.

### TRANSPARENCY STATEMENT

The lead author Samuel Y. Huang affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

### ORCID

Samuel Y. Huang  <http://orcid.org/0000-0003-3663-004X>

### REFERENCES

- Carrington AM, Fieguth PW, Qazi H, et al. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. *BMC Med Inform Decis Mak.* 2020; 20(1):4.
- Das P, Roychowdhury A, Das S, Roychowdhury S, Tripathy S. sigFeature: novel significant feature selection method for classification of gene expression data using support vector machine and t statistic. *Front Genet.* 2020;11:247.
- Peterson LE, Coleman MA. Machine learning-based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research. *Int J Approx Reason.* 2008;47(1):17-36.
- Qiu X, Gao J, Yang J, et al. A comparison study of machine learning (Random Survival Forest) and classic statistic (Cox Proportional Hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. *Front Oncol.* 2020;10:551420.
- Zhao X, Jiang D, Hu Z, et al. Machine learning and statistic analysis to predict drug treatment outcome in pediatric epilepsy patients with tuberous sclerosis complex. *Epilepsy Res.* 2022;188:107040.
- Aguayo GA, Zhang L, Vaillant M, et al. Machine learning for predicting neurodegenerative diseases in the general older population: a cohort study. *BMC Med Res Methodol.* 2023;23(1):8.
- Manz CR, Zhang Y, Chen K, et al. Long-term effect of machine learning-triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer: a randomized clinical trial. *JAMA Oncol.* 2023;9:414-418.
- Moskowitz CS. Using free-response receiver operating characteristic curves to assess the accuracy of machine diagnosis of cancer. *JAMA.* 2017;318(22):2250-2251.
- Loftus TJ, Shickel B, Ruppert MM, et al. Uncertainty-aware deep learning in healthcare: a scoping review. *PLOS Digit Health.* 2022;1(8):e0000085.
- Hasse K, Scholey J, Ziemer BP, et al. Use of receiver operating curve analysis and machine learning with an independent dose calculation system reduces the number of physical dose measurements required for patient-specific quality assurance. *Int J Radiat Oncol Biol Phys.* 2021;109(4):1086-1095.
- Ahn S. Building and analyzing machine learning-based warfarin dose prediction models using scikit-learn. *Transl Clin Pharmacol.* 2022; 30(4):172-181.
- Guo J, Zhang X, Kong J. Prediction of bile duct injury after transarterial chemoembolization for hepatocellular carcinoma: model establishment and verification. *Front Oncol.* 2022;12:973045.
- Kubo N, Cho H, Lee D, et al. Risk prediction model of peritoneal seeding in advanced gastric cancer: a decision tool for diagnostic laparoscopy. *Eur J Surg Oncol.* 2022;49:P853-P861.
- Wojtusiak J, Bagais W, Vang J, Guralnik E, Roess A, Alemi F. The role of symptom clusters in triage of COVID-19 patients. *Qual Manag Health Care.* 2023;32(suppl 1):S21-S28.
- Nishimura K, Oga T, Nakayasu K, Ogasawara M, Hasegawa Y, Mitsuma S. How different are COPD-specific patient reported outcomes, health status, dyspnoea and respiratory symptoms? An observational study in a working population. *BMJ Open.* 2019; 9(7):e025132.
- Abolhassani A, Prates MO, Mahmoodi S. Irregular shaped small nodule detection using a robust scan statistic. *Stat Biosci.* 2022;15: 141-162.
- Borenstein M. In a Meta-Analysis, the I-squared statistic does not tell us how much the effect size varies. *J Clin Epidemiol.* 2022;152: 281-284.



18. Chen J, Wu J, Huang X, et al. Differences in structural connectivity between diabetic and psychological erectile dysfunction revealed by network-based statistic: a diffusion tensor imaging study. *Front Endocrinol (Lausanne)*. 2022;13:892563.
19. Chol-Jun K. The power-law distribution in the geometrically growing system: statistic of the COVID-19 pandemic. *Chaos*. 2022;32(1):013111.
20. Clark LV, Mays W, Lipka AE, Sacks EJ. A population-level statistic for assessing Mendelian behavior of genotyping-by-sequencing data from highly duplicated genomes. *BMC Bioinformatics*. 2022;23(1):101.
21. Quereshy HA, Quinton BA, Ruthberg JS, Maronian NC, Otteson TD. Practice consolidation in otolaryngology: the decline of the single-provider practice. *OTO Open*. 2022;6(1):2473974X221075232.
22. Guroi-Urganci I, Waite L, Webster K, et al. Obstetric interventions and pregnancy outcomes during the COVID-19 pandemic in England: a nationwide cohort study. *PLoS Med*. 2022;19(1):e1003884.
23. Datsenko A, Marriott A, Shaw J, Patel R, Foley E. Complex contraception provision during the COVID-19 pandemic, how did sexual health services fare? *Int J STD AIDS*. 2022;33(5):467-471.
24. Kondratek B. Item-fit statistic based on posterior probabilities of membership in ability groups. *Appl Psychol Meas*. 2022;46(6):462-478.
25. Foster S. A person, not a statistic. *Br J Nurs*. 2022;31(12):671.
26. Rosenbaum PR. A statistic with demonstrated insensitivity to unmeasured bias for 2 x 2 x S tables in observational studies. *Stat Med*. 2022;41(19):3758-3771.
27. Reiser M, Cagnone S, Zhu J. An extended GFFit statistic defined on orthogonal components of Pearson's chi-square. *Psychometrika*. 2022;88:208-240.
28. Park G, Jung I. A tree-based scan statistic for zero-inflated count data in post-market drug safety surveillance. *Sci Rep*. 2022;12(1):16299.
29. Han Z, Sinharay S, Johnson MS, Liu X. The standardized S-X (2) statistic for assessing item fit. *Appl Psychol Meas*. 2023;47(1):3-18.
30. Tian D, Yan HJ, Shiiya H, Sato M, Shinozaki-Ushiku A, Nakajima J. Machine learning-based radiomic computed tomography phenotyping of thymic epithelial tumors: predicting pathological and survival outcomes. *J Thorac Cardiovasc Surg*. 2023;165:502-516.
31. Lee H, Ahmad S, Frazier M, Dundar MM, Turkkahraman H. A novel machine learning model for class III surgery decision. *J Orofac Orthop*. 2022.
32. Zhu W. Making bootstrap statistical inferences: a tutorial. *Res Q Exerc Sport*. 1997;68(1):44-55.
33. Donald A, Donner A. Adjustments to the Mantel-Haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. *Stat Med*. 1987;6(4):491-499.
34. Zhi M, Zhang K, Zhang X, et al. A statistic comparison of multi-element analysis of low atmospheric fine particles (PM(2.5)) using different spectroscopy techniques. *J Environ Sci (China)*. 2022;114:194-203.
35. Chiu CY, Chang YP. Advances in CD-CAT: the general non-parametric item selection method. *Psychometrika*. 2021;86(4):1039-1057.
36. Hirji KF. A comparison of exact, mid-P, and score tests for matched case-control studies. *Biometrics*. 1991;47(2):487-496.
37. Novroski NMM, Moo-Choy A, Wendt FR. Allele frequencies and minor contributor match statistic convergence using simulated population replicates. *Int J Legal Med*. 2022;136(5):1227-1235.
38. Diao D, Diao F, Xiao B, et al. Bayes conditional probability-based causation analysis between gestational diabetes mellitus (GDM) and pregnancy-induced hypertension (PIH): a statistic case study in Harbin, China. *J Diabetes Res*. 2022;2022:2590415.
39. Aloisio KM, Swanson SA, Micali N, Field A, Horton NJ. Analysis of partially observed clustered data using generalized estimating equations and multiple imputation. *Stata J*. 2014;14(4):863-883.

**How to cite this article:** Huang AA, Huang SY. Computation of the distribution of model accuracy statistics in machine learning: comparison between analytically derived distributions and simulation-based methods. *Health Sci Rep*. 2023;6:e1214. doi:10.1002/hsr2.1214