Computational Statistics Analysis on MNIST Digit Classification Result - How Could Object Shape & Size Relate to Prediction Biases?[1]

1. Introduction

Recently, the advancements in machine learning, specifically in deep neural networks, have become an innovative power in diverse domains, from natural language processing to image classification. However, there is a significant challenge, the unstableness of these models can lead to biases, impacting their credibility and fairness. Biases can occur in various forms, including skewed predictions, erroneous classifications, or overgeneralizations, which can cause serious consequences such as discriminatory decisions or misleading recommendations. Therefore, gaining insights into the inner workings of these models and understanding the sources of bias has become a critical endeavor.

Within the landscape of machine learning, the domain of image classification appears to be particularly susceptible to bias scrutiny. Images, serving as input data, encapsulate rich information. The interpretation of image classification models is intricately linked to its learning process. As a common sense, the size and shape of objects within an image are potential sources of biases in image classification. Whether it be a person, an animal, or a handwritten digit, the size and shape of an object can exert influence on the model's performance and interpretability. Smaller objects can cause low-confidence predictions, introducing concerns about the reliability of machine learning models, especially in scenarios where object size plays a pivotal role. Regarding the shapes of objects, an alternation of direction can lead to ambiguity in image classification because a change in perspective might cause the object to be considered as something else.

While prior research has acknowledged the impact of object size and shape on deep neural network models, there exists divergence in understanding how precisely object size and shape impact the inference process across different scenarios. This study addresses this issue by directing attention to the MNIST dataset, chosen for its clean and isolated handwritten digits. By isolating handwritten digits, the dataset facilitates an in-depth exploration of digit size and shape as potential biases affecting the model's predictions.

The examination of bias in image classification has caught significant attention, resulting in the development of various diagnostic methods. For instance, Kim et al. from Ajou University proposed a backpropagation-based bias detection framework that mitigates prediction bias under a few-shot learning scenario (Kim 3). Meanwhile, Schaff et al. introduced a meaningful matrix

[1] Code for this project could be found from GitHub Repo: https://github.com/ziming-zh/STATS406-Final-Project

as an indicator to quantify potential bias in the decision process of convolutional neural networks (Schaff 1). However, many existing methods are inconsistent with the "black box" nature of neural network models, hindering precise identification of the mechanisms leading to biased decisions. The inherent stochastic mechanisms in neural network design further compound the complexity of bias analysis.

In response to these challenges, this work introduces a novel approach to analyze the potential influence of object size and shape in image classification, with a specific focus on different confidence levels in model inference results. Employing computational statistical techniques, this approach facilitates rigorous statistical inference, allowing a detailed dissection of the impact of object size and shape on the model's decisions. The goal is to offer a clearer understanding of how bias manifests at varying confidence levels, unraveling the intricate interplay among object size, shape, and model predictions. Through these efforts, this study contributes to the ongoing initiatives aimed at fostering transparency and equity in machine learning systems.

In the ensuing sections, we delve into the methodology employed for examining feature normality, conducting stratified sampling, and implementing bootstrapped skewness diagnostics. Subsequently, we present simulation results validating the proposed methodologies. The paper concludes with an in-depth analysis of results, shedding light on the normality of feature distributions, the outcome of the stratified sampling, and the insights gained from bootstrapped conditional mean and variance.

2. Data Collection

The MNIST Database, a large collection of handwritten digits, has become a widely used benchmark dataset in the field of machine learning and computer vision, which has even been extending to handwritten letters in recent years (Li 2). It was derived from a larger NIST Special Database 3 and Special Database 1 of monochrome images of handwritten digits in 1998. The images in the MNIST dataset have been preprocessed, including size normalization and centering, resulting in a fixed-size image of 28x28 pixels.

The MNIST dataset consists of 60,000 training examples and 10,000 test examples of handwritten digit images, with each image representing a digit from 0 to 9. These images consist of a 28x28 binary pixels point, where the black (0) indicates the background and the white (1) stands for the digits.

To establish a model that contributes to the interpretation of our backbone neural network prediction, we aim to first mine some potential influential image-based features as candidates. We focus on the factors that have actual meanings and we infer to have a possible contribution to predicting the numbers in the MNIST dataset, and we provide our analysis of how these candidate factors could influence the prediction. The overall data collection process is expressed in Figure 1.

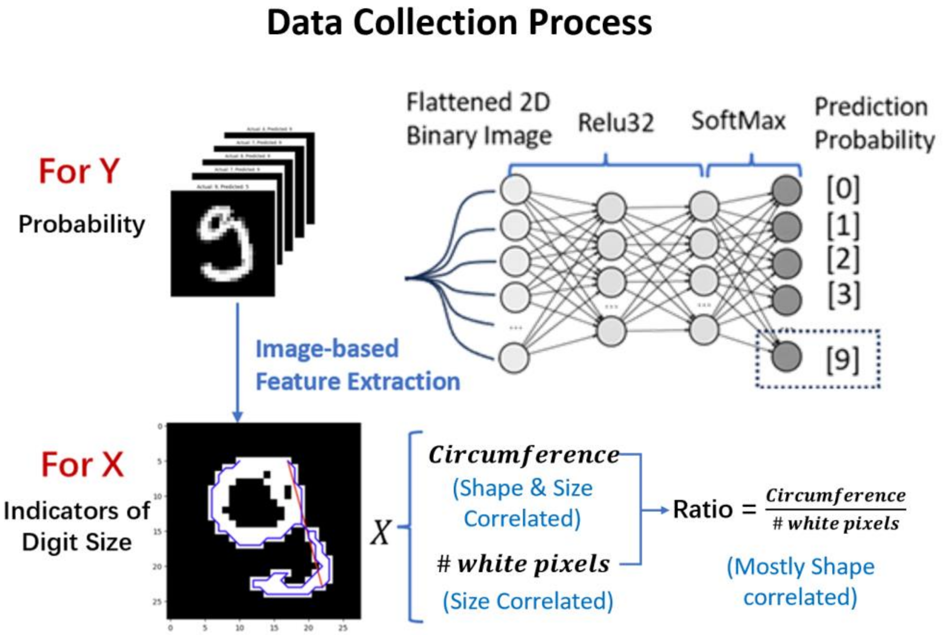

Figure 1 Schematic diagram of image-based feature extraction process

2.1 Size of the Picture ($Size$)

As the size of the handwritten digit is not always identical and different numbers have different occupations in the region, we first introduce a possible influential factor called Size, measuring the proportion of the white pixels in the whole picture. The formula is below:

$$Size \ = \frac{\#White\ Pixels}{\#Total\ Pixels}$$

This feature mainly calculated the portion of the area that the digit takes with respect to the total number of pixels of the image, which is constant 768 for all cases.

2.2 Circumference of White Pixels ($Circum$)

We consider circumference as one of the factors to be taken into account when establishing the model, as it can distinguish the writing of certain digits. In cases where the sizes are similar, digits with more complex strokes such as 4, 5, 8, 9, etc., are considered to have a larger circumference.

$$Boundary\ Points\ \epsilon = \{\sigma | Point[\sigma] = 1 \wedge \bigcup_{dir} Point[\sigma + dir] = 0\}$$

$$Circum = \sum_{(i,j)\in\epsilon} \|Point[i] - Point[j]\|_2$$

2.3 The Ratio of Circumference to Digit Size ($Ratio$)

It is well known that the circumference is dependent on the shape of the digit; however, the digit size would also interfere with the $Circum$. When the original digit is zoomed larger, the distance between two points from the original image would also be proportionally larger with respect to the size increase at the same time. In order to separate the influence of size and shape into two features, we design the $circum$: $size$ ratio as follows:

$$Ratio = \frac{Circum}{\#\ white\ pixels}$$

When the original digit is zoomed, the $Circum$ and $\#\ white\ pixels$ would change proportionally at the same time, and the ratio would remain unchanged. Therefore, in the context of following analysis, we could assume $Ratio$ as a mostly shape-correlated feature, $\#\ white\ pixels$ as a size-dominant one, and the $Circum$ is mutually influenced by shape and size of the digits.

3. Methodology

3.1 Examining Feature Normality

It is worth noting that our features are extracted from the established MNIST dataset, and we are not aware of the data-collecting process of the original samples. Therefore, the possibility of preliminary bias, for example, whether some style of handwriting is dominant in the original dataset, should be eliminated firsthand before we conduct bias analysis on the samples.

3.1.1 Visualization Method

In the process of examining the normality of features extracted from the MNIST dataset, two key visualization methods are employed: the Quantile-Quantile (Q-Q) plot and the Violin Plot.

A Q-Q plot is a statistical tool used to assess the distributional characteristics of a dataset, particularly focusing on its adherence to a theoretical distribution, such as the normal distribution. The plot compares the quantiles of the observed data to the quantiles expected under the assumption of a specified distribution. If the points in the plot align closely with a straight line, it indicates that the data is consistent with the assumed distribution. Deviations from the line suggest departures from normality, providing valuable insights into the shape and characteristics of the data distribution.

A Violin Plot combines features of a box plot and a kernel density plot to visualize the distribution of data across different categories. The plot consists of "violins," each representing a specific category or group. The width of the violin at a given point indicates the density or frequency of data points. This visualization method is particularly effective for comparing the distributional properties of multiple groups, revealing information about central tendency, spread, and the presence of outliers. Violin Plots are valuable for identifying patterns and irregularities in the data, making them a useful tool for assessing the normality and overall distributional characteristics of features extracted from the MNIST dataset.

These visualization methods play a crucial role in the initial stages of bias analysis, helping to ensure that the features under consideration exhibit a desirable level of normality and are free from potential biases introduced during the data collection process of the original MNIST samples.

3.1.2 Bootstrapped Skewness Diagnostics

To assess the skewness of the features extracted from the MNIST dataset, we employ bootstrapped skewness diagnostics. Skewness is a statistical measure that describes the asymmetry of a probability distribution about its mean. A skewness value of 0 indicates a perfectly symmetrical distribution, while positive or negative values suggest a skew to the right or left, respectively.

Bootstrapping is a resampling technique that involves repeatedly sampling with replacement from the dataset to estimate the distribution of a statistic (Hillis 22). In the context of skewness diagnostics, bootstrapping provides a robust method for estimating the skewness of the features and assessing its variability.

The following steps outline the process:

Algorithm 1: Bootstrapped Skewness Calculation:

```
a) Bootstrapped Skewness Calculation:
   For each feature in the dataset, generate multiple bootstrap samples by randomly
   drawing observations with replacement.
   Calculate the skewness for each bootstrap sample, resulting in a distribution of
   skewness values for each feature.
b) Compute Sample Mean of Bootstrapped Skewness
```

Compute the sample mean of the bootstrapped skewness values for each feature. This provides a point estimate of the skewness.

c) *Hypothesis Testing from P-value Calculation*
Formulate the null hypothesis
$H_0 \ : Skewness = 0$ *(indicating a symmetrical distribution)*
Use the bootstrapped skewness distribution to calculate the p-value associated with the null hypothesis.

Algorithm 1 Permutation Test for Asymmetricity Detection

By employing bootstrapped skewness diagnostics, we gain insights into the skewness of the feature distributions and assess whether they deviate significantly from a symmetrical shape. This analysis contributes to the overall understanding of the statistical properties of the features extracted from the MNIST dataset, complementing the information obtained from the Q-Q plot and Violin Plot visualizations in ensuring the robustness and normality of the examined features.

3.2 Stratified Sampling from Long- & Heavily-tailed Distribution

3.2.1 Detecting Asymmetricity from Permutation Test

The normality assumption from Section 3.1 with respect to the $Circ$, $Size$, and $Ratio$ is based on the common fact that the way people are taught to write digit should be largely similar for other people to recognize. However, when it comes to the prediction probability, the neural network is well-trained to binarize the classification probability. In this respect, we should expect a fairly large frequency appearing at the low end ($\rightarrow 0$) and high end ($\rightarrow 1$) of the prediction probability. Therefore, the original normality assumption does not hold anymore. A histogram visualization of the prediction probability suggests that the first and the last bin have the largest sample frequency, and the number of samples decays in an exponential shape when the prediction probability gradually approaches 0.5. Since the probability is defined on the [0,1] region, we can use truncated exponential distribution $f(x;\beta)$ to express the probability distribution at both ends.

$$f(x;\ \beta) = \begin{cases} \dfrac{\beta \cdot e^{-\beta x}}{1 - e^{-\beta}} & x \in [0,1] \\ 0 & otherwise \end{cases}$$

Another important characteristic could be discovered from the graph, although both ends of the probability seem to be exponentially distributed, the samples at the low end seem to be more aggregated at the low end than those at the high end. In other words, the exponential expression for the low-end distribution β_2 would have a larger decay rate than the high-end distribution β_1. This phenomenon is important for the stratification process, since if we can state that the distribution difference between low- and high-end is not sufficiently significant, then the distribution pattern over the whole range would be symmetric and we can classify the whole probability range into two regions. To detect whether the symmetric assumption is valid or not, we employ the permutation test as follows:

Algorithm 2: Permutation Test for Asymmetricity Detection

a) *Define the null hypothesis H_0 and the alternate hypothesis H_1:*

$$H_0: \beta_1 = \beta_2 \text{ (There is an equal decay rate between low- and high-end)}$$
$$H_1: \beta_1 < \beta_2 \text{ (There is a greater decay rate at the low end)}$$

b) *Define Test Statistic:*

$$X = \beta_2 - \beta_1$$

c) *Establish Permutation Procedure:*
Combine the decay rates from both distributions.
Randomly permute (shuffle) the combined data to create a new set of observations.
Calculate the test statistic X for the permuted data.

d) *Repeat Permutations*
Repeat the permutation process 10000 of times to create a distribution of the test statistic under the null hypothesis.

e) *Compare Observed Statistic to Permutation Distribution:*
Calculate the observed test statistic from the original data.
Compare the observed test statistic to the distribution of test statistics obtained from permutations.

f) *Calculate p-value:*
Count the number of times the permuted test statistic is as extreme as or more extreme than the observed test statistic.
The p-value is the proportion of permutations that resulted in a test statistic as extreme as the observed test statistic.

Algorithm 2 Permutation Test for Asymmetricity Detection

Based on the result of the permutation test, we could stratify the probability range into two or three regions as shown in Figure 2. The threshold values are chosen as $\{0.3, 0.7\}$ to guarantee that the length of the regions is roughly equal and we can acquire enough samples from Region II. After stratification, we can sample equal amount in each region and compute statistics to determine the difference of high-confidence samples (low-end and high-end) and low-confidence samples ($Prob \simeq 0.5$).

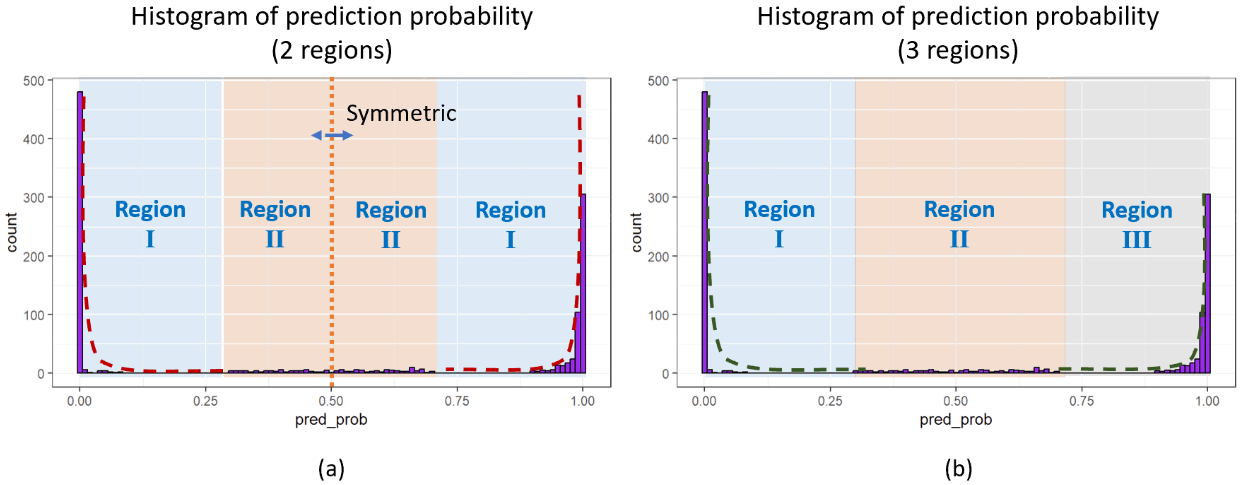

Figure 2 Possible modeling outcomes from permutation test: (a) H_1 is true, the probability range classified to two regions (b) H_1 is false, the probability range classified to three regions

3.2.2 Bootstrapping the Conditional Mean and Variance for Different Regions

After region stratification, we can successfully classify the whole probability region into high-confidence region and low-confidence region in either case (a) or (b) in Figure 2. As elaborated in the result section, we've determined that the difference between low-end and high-end is non-negligible for analysis and choose case (b) as final stratification model. Therefore, our further discussion would be based on case (b), where Region I and III are the high-confidence regions and Region II is the low confidence region. In this case, we can establish our conditional mean and variance for region m as follows:

$$\text{Conditional Mean: } Mean\left(f(\cdot)_{[i]} \,\middle|\, i \in Region[m]\right)$$

$$\text{Conditional Variance: } Var\left(f(\cdot)_{[i]} \,\middle|\, i \in Region[m]\right)$$

$$f(\cdot) = \{Circ(\cdot), Size(\cdot), \frac{Circ(\cdot)}{Size(\cdot)}\}$$

In our analysis, we further perform bootstrapping over the computed statistics to acquire the monte-carlo confidence interval. The bootstrapping process is similar to Algorithm 1 in *Section 3.1.2.*

4. Simulation

This simulation section serves to validate the two main methodologies we employ to investigate the distribution of prediction probability (exponential distribution among the high-end and low-end of the probability range) and the extracted features from image (normally-distributed and influenced from outliers). The rationale behind the selection of these specific methods is further expounded upon in the respective subsections.

4.1 Permutation Test for Exponential Distribution with Different Decay Rate

In the context of comparing decay rates (beta) between two exponential distributions, the permutation test emerges as a preferable statistical approach over traditional tests such as the t-test and Wilcoxon rank-sum test. This preference is rooted in the permutation test's non-parametric nature, rendering it robust when faced with deviations from distributional assumptions. Particularly advantageous with small sample sizes, the permutation test excels in scenarios where the assumptions of parametric tests might not be met (Elliffe 12). Moreover, exponential distributions, commonly employed to model event times, can exhibit skewed shapes, a characteristic that can challenge the Wilcoxon rank-sum test, sensitive to distributional shapes. The permutation test, being distribution-free and adaptable to specific hypotheses, offers a more flexible and reliable method for assessing differences in decay rates between the two distributions (Frattarolo 32-46). To validate this, we generate two exponentially distributed dataset with $\beta_1 = 2$ and $\beta_2 = 3$, as presented in Figure 3.

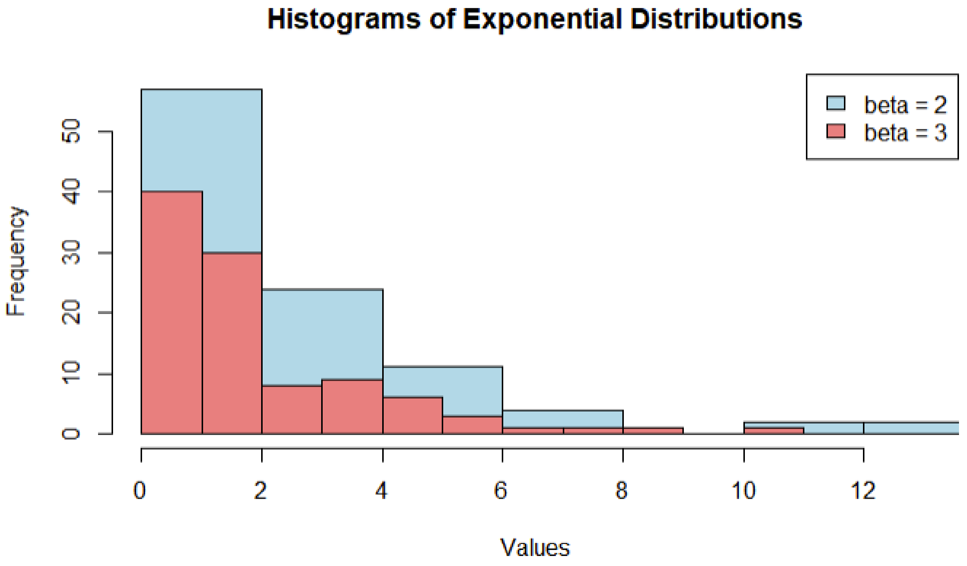

Figure 3 Histogram of exponential distribution with $\beta = 2$ (blue) and $\beta = 3$ (red)

To examine whether there is difference in decay rate between two samples, we employ hypothesis test by suggesting:

$$H_0: \beta_1 = \beta_2$$
$$H_1: |\beta_1 - \beta_2| > 0$$

Since the mean μ of an exponential distribution with decay rate β is given by the reciprocal:

$$\mu = \frac{1}{\beta}$$

By taking advantage of this characteristics, we could simplify the calculation process and establish the statistics as:

$$\delta = |\gamma_1 - \gamma_2| = \left|\frac{1}{\beta_1} - \frac{1}{\beta_2}\right| = |Mean(X_1) - Mean(X_2)|$$

where γ_1, γ_2 stands for the decay rate of data X_1 and X_2.
Our initial hypothesis could be adjusted accordingly as follow:

$$H_0: \gamma_1 = \gamma_2$$
$$H_1: |\gamma_1 - \gamma_2| > 0$$

In this case, $|\beta_1 - \beta_2| = 1$, which suggests that the alternative hypothesis H_1 is true, and we would prefer the test that could successfully reject H_0 with low p-value and high power.

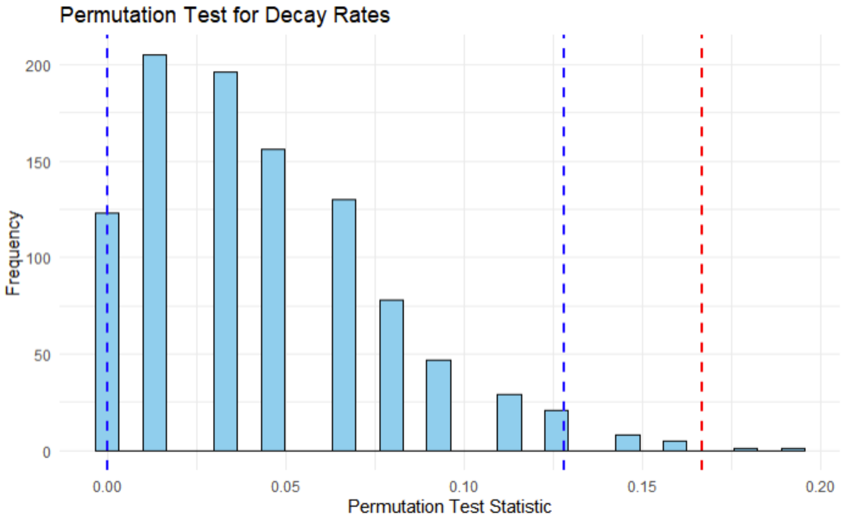

Figure 4 Result of Permutation Test for H_0 (blue line shows confidence interval at level 0.95, red line shows the observed value)

Table 1 Result Comparison of Different Statistical Test on β Examination

Test	Power
Permutation Test	0.98
Studentized-T Test	0.82
Wilcoxon Rank-sum Test	0.76

Figure 4 presents the result of the permutation test, the graph shows that the observed value is not within the confidence interval at 0.95 level, which suggests a higher than 0.95 confidence to reject the null hypothesis. Further results from studentized-t test and wilcoxon rank-sum test further show the power far less than 0.95. The specific power level can be found in Table 1. This simulation result suggests that permutation test is relatively effective to examine the exponential distribution with different decay rate.

4.2 Bootstrapping for Noised Normal Distribution with Outlier Influence

Another practice in this work we want to examine is the bootstrapping method employed to examine the mean and variance of the extracted features, which have been previously assumed as approximately normally distributed. It is not intuitively meaningful to perform a bootstrapped t-test on a well-distributed dataset, which could adequately examine using simple t-test. However, we still select bootstrapped t-test since it could provide us with confidence interval, giving a more comprehensive picture of the precision of the estimate. Moreover, we've noticed potential outliers in the distribution pattern of size-related features although they largely conform to the normal distribution shape. To better illustrate how bootstrapping methods outperform with outlier influence, the simulation is conducted between the normal distribution $N(0, \sqrt{5})$ and the noised normal distribution, whose probability density function (PDF) is defined as follow:

$$f(x) = (1 - p) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} Exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) + p \cdot g(x)$$

where:

- p is the probability of an outlier occurring,
- μ is the mean of the normal distribution (0 for simulation),
- σ is the standard deviation of the normal distribution ($\sqrt{5}$ for simulation),
- $g(x)$ is the density function of the outlier distribution.

For simplicity, we assume the outlier distribution as a standard normal distribution $g{\sim}N(0,1)$:

$$g(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} Exp\left(-\frac{x^2}{2}\right)$$

The boxplot of normal distribution $N(0, \sqrt{5})$ and a sample noised normal distribution ($p = 0.2, f = 3$) is presented in Figure 5.

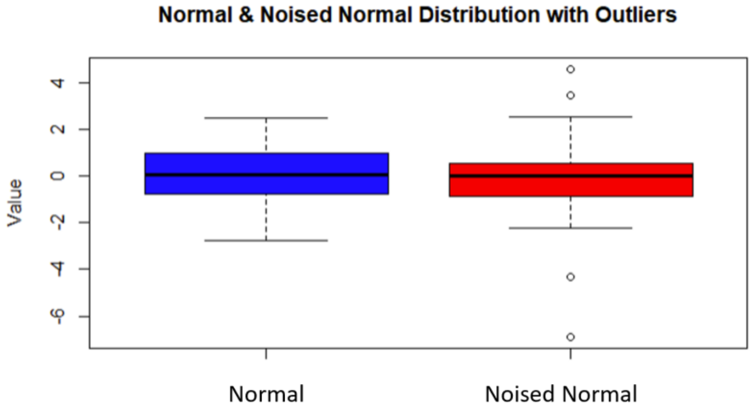

Figure 5 Normal & Noised Normal Distribution with Outliers ($p = 0.2, f = 3$)

To test whether the traditional t-test is significantly more sensitive to the outlier condition, we draw out the power curve for traditional t-test and bootstrapped t-test over a range of outlier probability $\wp = \{0,0.1,0.2,0.3,0.4\}$ and outlier factor $\mathcal{F} = \{1,2,3,4,5\}$, where higher value suggests larger outlier influence. We establish the null hypothesis H_0 and alternate hypothesis H_1 as:

$$H_0: Mean(X_{noised}) \geq Mean(X_{normal})$$

$$H_1: Mean(X_{noised}) < Mean(X_{normal})$$

As shown in Figure 6 (a) and (b), the power curves dramatically fluctuate for traditional t-test given different outlier impact, while the power values calculated from bootstrapped t-test are close to 0.5 for all cases, which is reasonable since $\mu_{noised} = \mu_{normal}$.

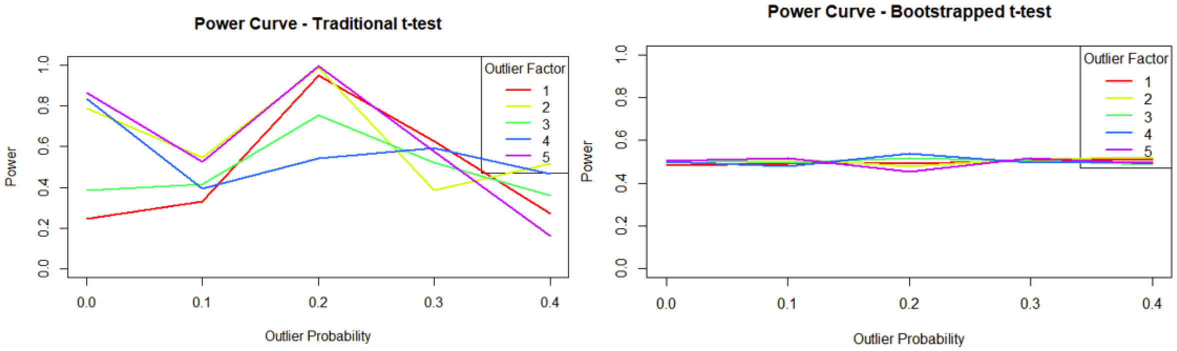

Figure 6 Power Curve of (a) Traditional T-Test (b) Bootstrapped T-Test

In conclusion, the simulation results validate our choice of permutation test to examine decay rate of exponential distribution and the efficacy of bootstrapping method to enhance the robustness of t-test analysis on noised normally-distributed datasets.

5. Result & Analysis

5.1 Normality Visualization of Feature Distribution for All Digits

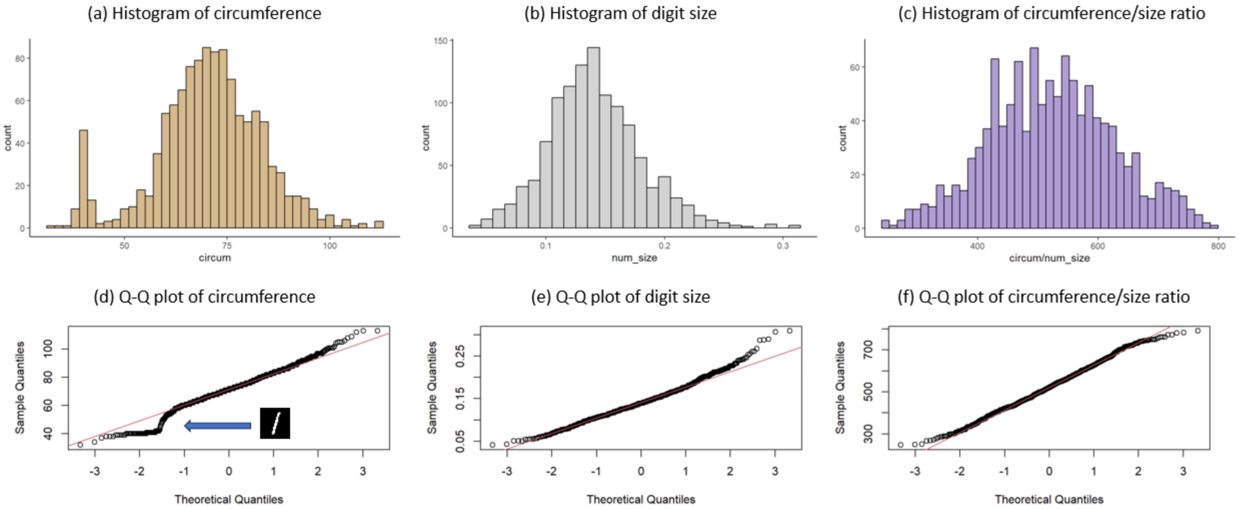

Figure 7 Histogram and Q-Q plot for normality test

Based on the distributions of the features as shown in Figure 7, an overall normal distribution is evident within our dataset. To verify the normality of the raw data, Q-Q plots were generated for each variable. Examination of these Q-Q plots reveals that all three variables exhibit a distribution close to normal, as indicated by the straight lines in the plots. Specifically, it is worth noting that from the visualization of circumference, a small peak could be discovered at the left side of the main distribution. A closer observation reveals that the peak is mainly derived from the digit “1”. Since the circumference of digit “1” is very aggregated and significantly smaller from other digits, this phenomenon is more apparent on circumference compared with other features. However, other digits form a well-established normal distribution according to the Q-Q plot. Moreover, digit “1” only accounts for approximately 10 percent of the total samples, which would not largely influence the normal distribution trend as a whole. Therefore, we still take credence of the normality of the data and does not consider any preliminary bias from unfair sample selection before our analysis process.

5.2 Stratified Sampling Result

In this section, we first conduct the permutation test of the exponential decay coefficient at high-end (β_1) and low-end (β_2). The result of the permutation result of $X = |\beta_2 - \beta_1|$ is reported as follows:

$$Mean(X) \approx 885$$
$$p - value \approx 2.03 \times 10^{-8}$$

According to the result, the p-value is considerable smaller than 0.05, which is significant enough to suggest that the null hypothesis is false. Therefore, the symmetric model in Figure 2(a) does not hold true, and we still need three regions where Region I and III are (truncated) exponentially distributed with different decaying coefficient. The result makes sense because in the actual training process, the model is given 9 times the negative samples (other than 9 digits) than the positive samples (9 digits). Based on the assumption that more sample input could lead to a better prediction accuracy, it is reasonable that the model is more confident to exclude a sample than to make sure that the sample is the specific given digit.

To further validate the permutation test result, we've fitted the distribution in Region I and III. The fitting result suggests that $\beta_1' = 500$ and $\beta_2' = 1500$. In that case, $X' = |\beta_2{}' - \beta_1{}'| = 1000$ which is close to the estimated $Mean(X)$ from permutation test. The original and fitted distributions of Region I and III are presented in Figure 8 (a) to (d). Region II is supposedly at the tail of the exponential distribution from both sides according to the established model. The samples in Region II are very few and we only collected 158 samples from 30,000 validation set samples. The histogram shows that the distribution is roughly uniform among all regions, and we can directly sample from uniform distribution. In order to perform comparisons between different stratified regions, we sampled 150 data equally from each region.

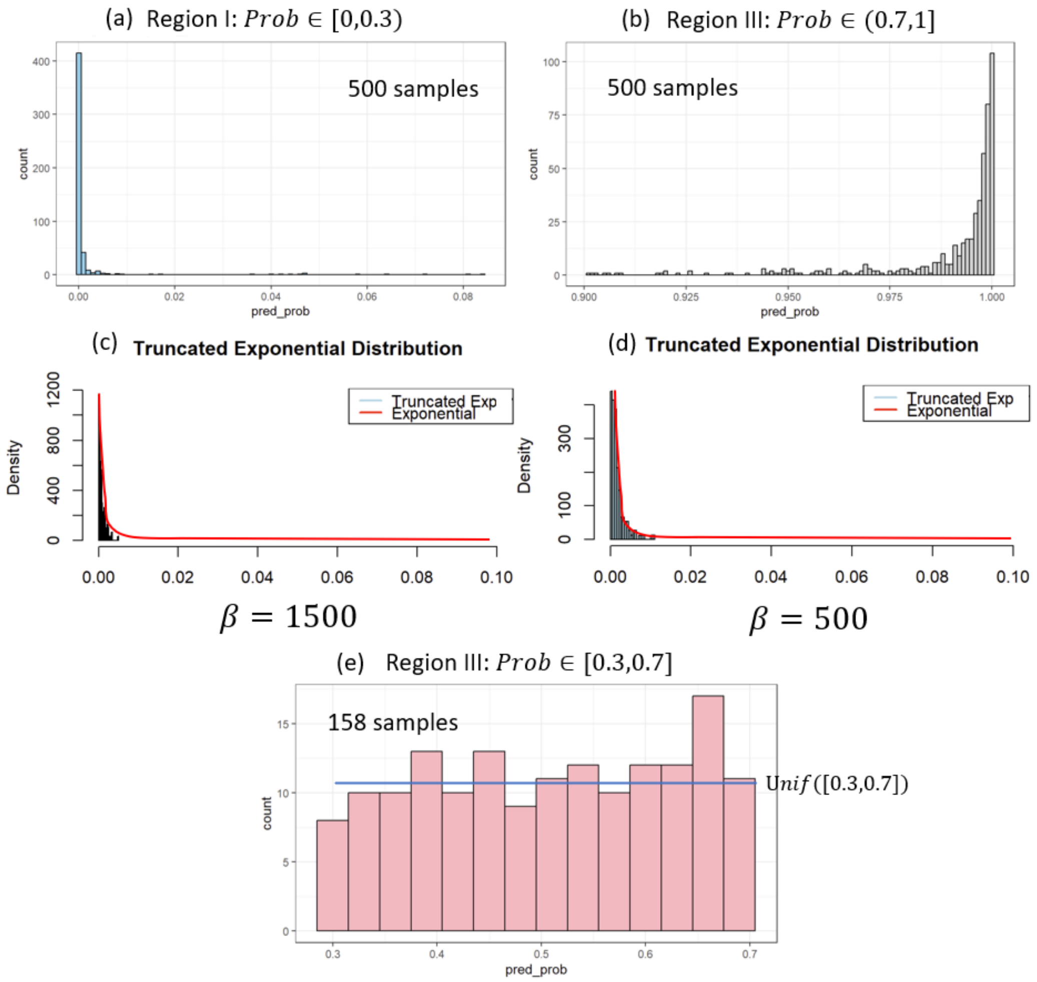

Figure 8 Stratified Sampling Result\

5.3 Bootstrapped Conditional Mean & Variance

After successfully sampling the data on each region, we perform bootstrapping to compute the conditional mean & variance for three regions together with the confidence intervals. The bootstrapping result is recorded in Table 2.

Table 2 Bootstrapped Conditional Mean & Variance for Three Regions

		Mean	Conf_low	Conf_high	Var	Conf_low	Conf_high
Circum	I	68.75	68.67	68.83	**290.13**	**288.50**	**291.76**
	III	71.21	71.17	71.24	54.46	54.07	54.84
	II	70.61	70.55	70.68	163.69	162.51	164.88

Size	I	0.1381	0.1379	0.1383	**0.0019**	**0.0019**	**0.0019**
	III	0.1490	0.1489	0.1491	0.0006	0.0006	0.0006
	II	0.1407	0.1405	0.1409	0.0018	0.0018	0.0018
Circum/ Size Ratio	I	517.78	517.29	518.26	9541.45	9475.93	9606.97
	III	488.07	487.67	488.48	6456.27	6405.28	6507.27
	II	529.52	528.94	530.11	**13736.62**	**13656.00**	**13817.25**

To better present the bootstrapping result, we have also visualized the kernel density plot of the mean statistics of the bootstrapped sample (bootstrapped sample distribution) in Figure 9. In the expected case, in order to suggest that there is no evident bias in the model prediction result, the feature statistics should satisfy:

$$Mean(\cdot \mid Region\ III) \approx Mean(\cdot \mid Region\ I) \approx Mean(\cdot \mid Region\ II)$$

$$Var(\cdot \mid Region\ III) < Var(\cdot \mid Region\ I) < \ Var(\cdot \mid Region\ II)$$

The variance of Region III is the smallest here since the "9" digits should have more aggregated feature value compared with other digits. We also assume that the variance of Region II is the largest one because those samples, when predicted with an ambiguous probability, tend to have more bizarre shapes compared with the normal situation, resulting in the largest variance in all three regions.

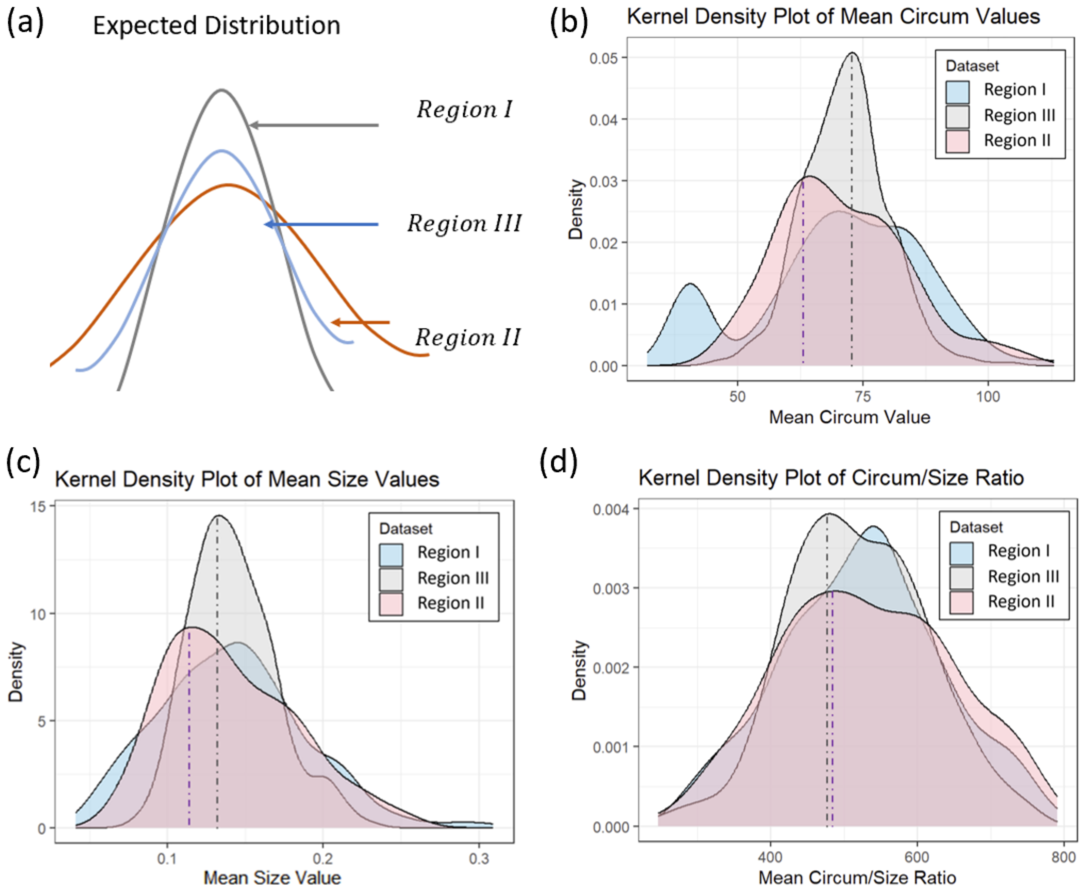

Figure 9 Kernel Density Plot of the Mean Statistics from Bootstrapping

The result shown in Table 2 and Figure 9, however, suggests that the expected condition only occurs for the feature $Ratio$, which is a shape-correlated feature. When it comes to size correlated features $Cicum$ and $Size$, although the mean value still approximates for three regions, the variance of Region I exceeds that of Region II. Moreover, the kernel density plot reveals that the peaks of the Region II has been deviated from the peak of Region I.

$$Var'(\cdot \mid Region\ III) < Var'(\cdot \mid Region\ II) < \ Var'(\cdot \mid Region\ I)$$

The different order of variance in three regions and the peak deviation both suggest possible bias for the size-correlated features. To further understand the underlying reason, we turn to a detailed observation of each digit and examine whether our previous normality assumption still holds true when it comes to separate digit.

5.4 Digit-wise visualization & Skewness Test for “9” Digit

To determine whether our previous normality assumption still holds true when it comes to separate digit, we visualize the distribution using violin plot as shown in Figure 10.

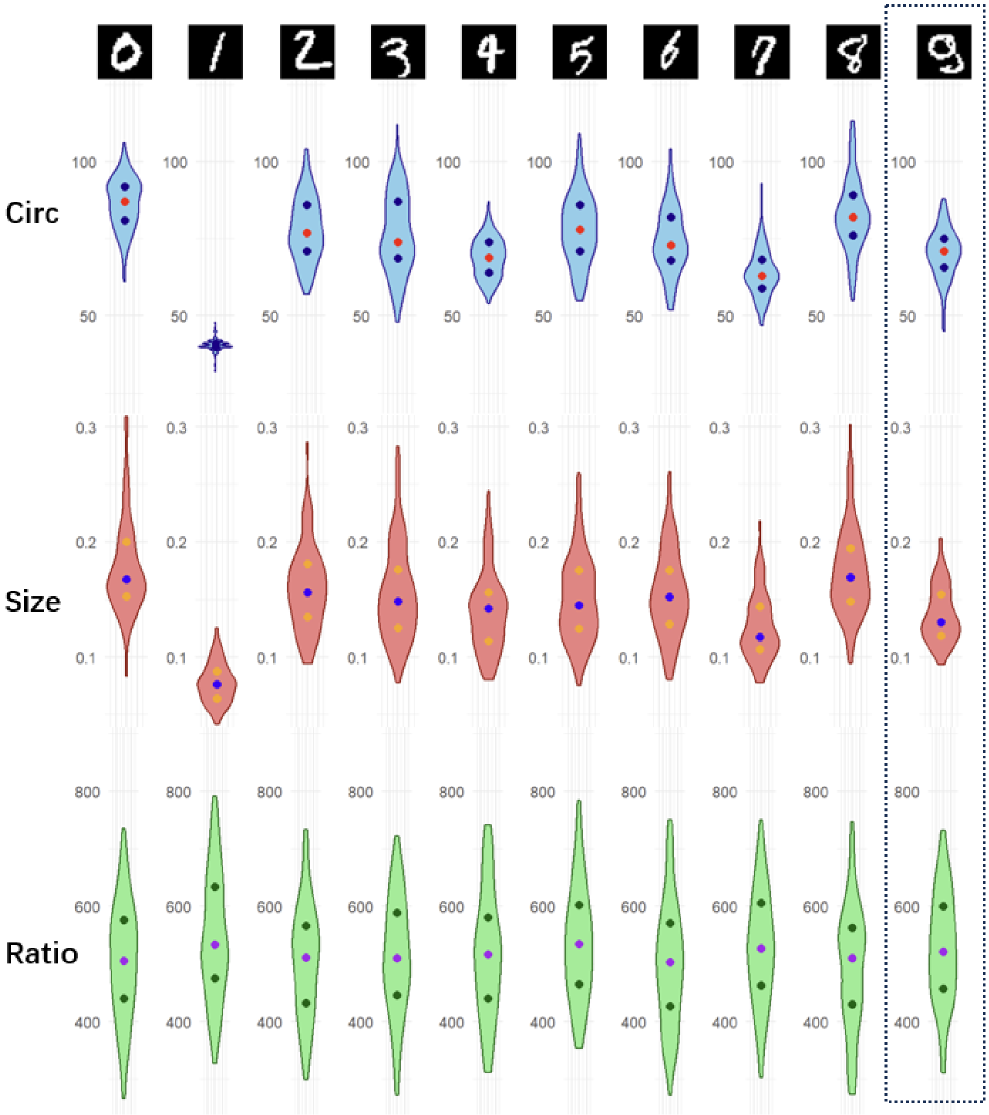

Figure 10 Violin Plot Visualization for Feature Distributions of Separate Digits

Figure 10 illustrates that the $Size$ feature exhibits pronounced skewness across most digits, notably in the case of the digit "9." Despite an overall appearance of normal distribution, primarily influenced by the digit "1" that constitutes the left tail and alleviates the skewness. As for the $Circum$ feature, its distribution is more favorable than $Size$, yet skewness persists across several digits such as "9," "0," and "7." To quantify the significance of skewness in each feature, we conducted bootstrapped skewness tests, the results of which are documented in Table 3. The table reveals that the skewness of $Circum$ is consistently smaller than 0, while that of Size is consistently larger than 0, with a confidence level exceeding 95%. The outcomes from the local normality analysis pinpoint a potential source of bias in size-related features. Unlike shape-related features, which demonstrate a normal distribution for each individual digit, the size-related digits exhibit inconsistent distributions across different digits. This inconsistency perpetuates a non-normal distribution for each digit, juxtaposed with an overall appearance of normality.

Table 3 Bootstrapped Skewness Test Result

	Statistics	0.95 Confidence Interval
Circum	-0.37	[-0.97, -0.12]
Size	0.57	[0.26, 0.99]
Ratio	-0.004	[-0.34, 0.33]

6. Discussion

Based on the MNIST dataset as our experimental playground, we have gained nuanced insights into the intricate dynamics of deep neural network models by investigating the potential biases introduced by object size and shape. Through the use of stratified sampling, bootstrapped skewness diagnostics, and a focus on feature normality, we were able to identify biases at different confidence levels in model inference.

The examination of feature normality involved rigorous statistical methods, including Quantile-Quantile plots, Violin plots, and bootstrapped skewness diagnostics. These tools allowed us to scrutinize the distributional characteristics of extracted features, ensuring their adherence to desirable levels of normality. Based on the stratified sampling from long-tailed distributions, we were able to assess bias in prediction probabilities. With our permutation test, we were able to distinguish between high-confidence samples and low-confidence samples based on their nuanced patterns.

In our simulations, we validated our chosen methodologies, with the permutation test demonstrating its effectiveness in examining decay rates in exponential distributions and bootstrapping demonstrating its robustness in analyzing noised normal distributions with outliers. These simulations bolstered the credibility of our approach and underscored its applicability to diverse scenarios.

As a result, we discovered biases in size-related features in particular for the "9" digit, despite the overall normality observed in feature distributions. The bootstrapped skewness tests accentuated the significance of this bias, highlighting inconsistencies in the distribution of size-related features across different digits. The variance analysis further discerned the disparity in variance order among the stratified regions, signaling potential biases in size-correlated features.

The implications of our research extend beyond the confines of this study. Our work contributes to the ongoing conversation on transparency and equity in artificial intelligence by providing a granular understanding of how object size and shape impact machine learning predictions. The identified biases call for a reevaluation of model architectures and training processes to mitigate the observed inconsistencies. Additionally, our findings underscore the need for continued research into more sophisticated methods for bias detection and mitigation, considering the intricate interplay of factors influencing model decisions.

In terms of practical advice, our study emphasizes the importance of scrutinizing size and shape-related features, particularly in applications where digit recognition plays a critical role. Designing models that account for the potential biases revealed in our analysis could lead to more reliable and fair machine learning systems.

Reference

Li D. "The MNIST database of handwritten digit images for machine learning research." *IEEE Signal Processing Magazine*, vol. 29, no. 6, 2012, pp. 141–142, https://doi.org/10.1109/msp.2012.2211477.

Kim J., et al. "Dataset bias prediction for few-shot image classification." *Electronics*, vol. 12, no. 11, 2023, p. 2470, https://doi.org/10.3390/electronics12112470.

Schaaf N., et al. "Towards measuring bias in image classification." *Lecture Notes in Computer Science*, 2021, pp. 433–445, https://doi.org/10.1007/978-3-030-86365-4_35.

Hillis, D., et al. "An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis." *Systematic Biology*, vol. 42, no. 2, 1 June 1993, pp. 182–192, https://doi.org/10.1093/sysbio/42.2.182.

Elliffe E., et al. "Rank-Permutation Tests for Behavior Analysis, and a Test for Trend Allowing Unequal Data Numbers for Each Subject." J*ournal of the Experimental Analysis of Behavior*, vol. 111, no. 2, 7 Feb. 2019, pp. 342–358, https://doi.org/10.1002/jeab.502.

Frattarolo L. et al. "Systemically Important Banks: A Permutation Test Approach." *SSRN Electronic Journal*, 2016, https://doi.org/10.2139/ssrn.2862546.