Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 2, No. 3, 2026

A Systematic Analysis of the Impact of BloodMNIST Data Characteristics on CNN Classification Performance

Jiahan Wen

School of Mathematics and Statistics, Hanshan Normal University, Chaozhou, 521000, China

Abstract: BloodMNIST, as a standardized medical image classification dataset, has been widely used to validate the classification performance of convolutional neural networks (CNNs). This paper takes BloodMNIST as the research subject, focusing on the key data characteristics that affect CNN classification performance and typical misclassified samples. Through comparative experiments on the classification results of LeNet and ResNet, we systematically investigate the mechanisms by which different data properties influence model behavior. Specifically, this study conducts experiments around three core influencing factors: (1) class visual similarity through sample visualization and confusion matrix analysis, we reveal the impact of similar classes on misclassification; (2) class sample size (class imbalance) – by comparing standard cross-entropy with weighted cross-entropy, we analyze the effect of sample distribution on classification performance; (3) sample quality/noise – by introducing mild perturbations, we evaluate the robustness of different models. Experimental results show that visual similarity and class imbalance are important causes of misclassification, and ResNet is more robust than LeNet in distinguishing similar classes and resisting noise. This paper provides a systematic experimental analysis framework for understanding the relationship between medical image data characteristics and CNN classification behavior.

Keywords: BloodMNIST, convolutional neural network, classification performance, misclassification analysis, visual similarity, class imbalance, noise robustness.

1. Introduction

1.1. Research Background

In recent years, with the rapid development of deep learning, convolutional neural networks (CNNs) have become a fundamental tool in medical image analysis, achieving remarkable progress in tasks such as blood cell recognition, pathological slide classification, and radiology-assisted diagnosis [1]. However, most existing studies focus on improving model performance by designing deeper or more complex network architectures, optimizing training strategies, or introducing novel regularization methods [2]. In contrast, systematic analyses of how data characteristics themselves influence CNN classification behavior remain relatively limited [3]. In fact, in medical imaging scenarios, data characteristics often largely determine the upper bound of model performance and the patterns of misclassification. For example, visual similarity between classes may blur feature boundaries, class distribution imbalance may weaken the learning of minority classes, and sample quality or noise may interfere with the stable expression of high-level features. Therefore, alongside model improvements, it is necessary to adopt a data-centric perspective to deeply understand the relationship between data properties and model behavior.

1.2. Research Motivation

BloodMNIST, as a standardized blood cell classification dataset in the MedMNIST series, features a uniform 28×28 resolution, clear class definitions, and a canonical training/validation/test split, providing an ideal platform for systematically studying the impact of data characteristics on CNN classification behavior [4].

Unlike performance-driven research, the core motivation of this paper is not to pursue higher classification accuracy alone, but rather to analyze and explain the sources of model misclassification – that is, to investigate to what extent and in what ways specific data characteristics affect CNN decision-making. This perspective offers more instructive insights for model selection, data acquisition and preprocessing, and training strategy design.

1.3. Research Objectives

Based on the above background and motivation, the main objectives of this paper include:

Systematically analyze the key data characteristics that affect CNN classification performance, with a focus on three factors: class similarity, class imbalance, and sample quality/noise;

Through comparative experiments between LeNet and ResNet, reveal how different data properties manifest under different network architectures, and analyze the effect of model complexity on data sensitivity;

Explore and analyze typical misclassified samples, and explain the possible causes of model decision errors from a data-characteristic perspective.

2. Methods

2.1. Dataset

This paper adopts BloodMNIST as the research subject. BloodMNIST belongs to the MedMNIST v2 series of blood cell classification datasets, containing 8 classes of blood cell images. All samples are uniformly processed into 28×28 grayscale images and pre-divided into training, validation, and test sets according to the standard protocol. This uniform processing ensures fair comparisons among different models under the same data scale and distribution.From the perspective of data distribution, BloodMNIST exhibits a certain degree of class imbalance, with some classes having significantly fewer samples than others. This characteristic

provides a natural experimental basis for systematically analyzing the impact of class distribution on CNN classification performance, and also makes the dataset suitable as a typical case for studying the relationship between data properties and model behavior.

2.2. Model Selection

To systematically analyze the sensitivity of different network complexities to data characteristics, we select two representative convolutional neural networks as analytical tools:

LeNet – a classic lightweight CNN architecture with a small parameter count, mainly relying on local texture features for discrimination, suitable as a coarse-grained feature extractor [5];

ResNet (shallow version) – introduces residual connections to mitigate the vanishing gradient problem, possessing stronger hierarchical feature representation capability to capture finer-grained structural information, suitable as a fine-grained feature extractor [6].

It is important to emphasize that, in this paper, models are not intended to pursue optimal classification performance, but rather serve as tools for probing and analyzing data characteristics, revealing how different data properties manifest under different network structures.

2.3. Evaluation Metrics

To comprehensively evaluate model performance and classification behavior, the following metrics are adopted: Overall Accuracy – measures the overall correct classification rate; Macro-F1 – averages the F1 scores over all classes, suitable for imbalanced scenarios; Recall – reflects the model's recognition ability for different classes, facilitating analysis of minority-class performance; Confusion Matrix – for visual analysis of misclassification patterns between classes; Typical Misclassified Sample Visualization – for interpreting possible causes of model decision errors from a data perspective.

The combination of these metrics not only reflects overall performance but also helps to deeply understand behavioral differences of models under various data characteristics.

2.4. Factor Analysis Framework

Around the research objectives, we design three sets of comparative experiments centered on data characteristics to systematically analyze the impact of different factors on CNN classification performance:

Class Similarity Analysis: through sample visualization combined with confusion matrix analysis, investigate the influence of visually similar classes on misclassification, and compare the differences between LeNet and ResNet in distinguishing similar classes.

Class Imbalance Analysis: by comparing standard cross-entropy (CE) with weighted cross-entropy (WCE), evaluate the effect of class sample size differences on model performance, with particular attention to changes in minority-class Recall [7, 8].

Sample Quality/Noise Analysis: introduce mild perturbations (Gaussian noise) on the test set to evaluate the robustness of different models under noisy conditions, thereby revealing the impact of sample quality on classification performance [9].

2.5. Model Formulation: Weighted Cross-Entropy Loss Function

LWCE=−N1 i=1∑N wyi log(∑j=1Cexp(zi, j)exp(zi, yi))

Notation:

LWCE: weighted cross-entropy loss value.

N: total number of samples in a batch.

C: total number of classes (in BloodMNIST, C=8C=8).

Yi: true class label of the ii-th sample.

zi, j: model output score (logit) for the ii-th sample with respect to class jj.

Wyi: weight assigned to the true class yiyi. This is the core parameter of weighted cross-entropy.

3. Experiments

3.1. Experimental Setup

To systematically analyze the impact of different data characteristics on CNN classification performance, we conduct comparative experiments around three key factors: class visual similarity (Experiment 1), class imbalance (Experiment 2), and sample noise (Experiment 3). All experiments are evaluated on the BloodMNIST test set, with primary metrics being Overall Accuracy and Macro-F1; in the noise experiment, we additionally report Accuracy and Macro-F1 after adding mild Gaussian noise to measure model stability under perturbation.

For model selection, we choose LeNet and ResNet (shallow version) as comparators. LeNet, as a lightweight network, mainly relies on local texture features and serves as a coarse-grained feature extractor; ResNet introduces residual connections and possesses stronger hierarchical feature representation, suitable as a fine-grained feature extractor. By comparing the two models under identical data conditions, we can systematically analyze the sensitivity differences of different network complexities to data properties.

3.2. Experiment Design

Class Similarity. To investigate the influence of class visual similarity on CNN classification performance, we first visualize sample images from each class of BloodMNIST, manually identifying visually similar classes. Then, under the same training settings, we train LeNet and ResNet respectively, and compute confusion matrices and misclassification distributions on the test set to quantitatively analyze the classification behavior differences between the two models on similar classes.

Class Imbalance. To analyze the impact of class imbalance on CNN classification performance, we first count the sample numbers of each class in BloodMNIST to confirm the existence of a moderate imbalance. Subsequently, we train LeNet and ResNet using standard cross-entropy (CE) and weighted cross-entropy (WCE) respectively, and compare per-class Recall and overall Macro-F1 on the test set to evaluate the effect of different loss functions on minority-class recognition.

Sample Quality/Noise. To assess the influence of sample quality on CNN classification performance, we introduce mild Gaussian noise (standard deviation σ=0.05σ=0.05) on the test set. Keeping model parameters unchanged, we re-evaluate the classification performance of LeNet and ResNet to analyze the sensitivity of different models to input perturbations.

3.3. Experimental Results

Experiment 1 Results:

Figure 1 shows typical samples from the 8 classes of the BloodMNIST dataset. Visual inspection reveals that Class 2 (neutrophils) and Class 3 (eosinophils) are relatively similar in nuclear morphology and cytoplasmic granule distribution, while Class 5 (monocytes) and Class 6 (lymphocytes) overlap in cell size and nuclear-to-cytoplasmic ratio. These visual similarities provide intuitive evidence for subsequent misclassification analysis.

Figures 2 and 3 present the confusion matrices of LeNet and ResNet on the test set, respectively. From the matrix values, it is clear that both models exhibit significantly higher misclassification rates between Class 2↔3 and Class 5↔6 than other class pairs. For example, LeNet misclassifies Class 2 as Class 3 at a rate of 18.7%, and Class 5 as Class 6 at 15.2%; the corresponding rates for ResNet are 12.3% and 10.8%. This indicates that visual similarity directly leads to blurred class boundaries in the feature space, and even deep networks cannot completely eliminate this effect.

Figure 4 displays several typical misclassified samples. Comparing these with standard samples of the same class, the misclassified samples often lie in transitional morphological states – for instance, irregular nuclear shapes or staining intensities between the typical values of the two classes, causing the model to activate similar feature patterns. Further comparison shows that ResNet has lower misclassification rates than LeNet on most easily confused class pairs, indicating that the high-level semantic features extracted by residual structures are more discriminative than shallow texture features. However, even for ResNet, misclassifications between similar classes account for over 60% of its total errors, confirming that inherent feature overlap in the data is the fundamental bottleneck for classification performance.

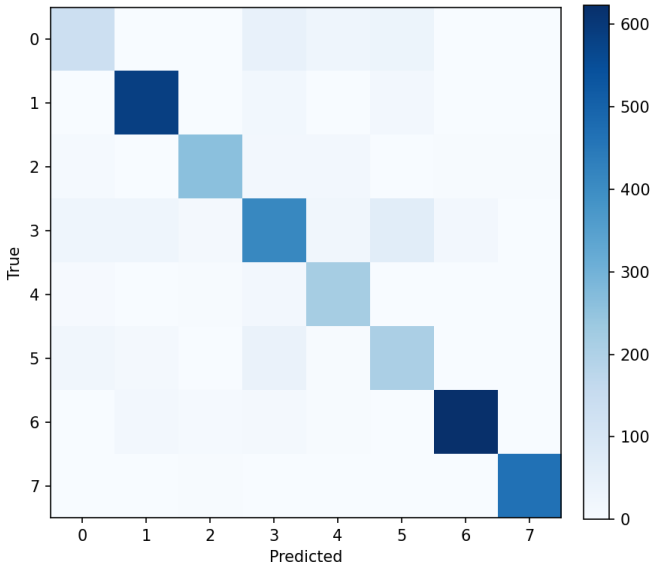

Figure 1. Visualization of typical samples from the 8 classes of BloodMNIST

(Note: Each row shows 5 example images from one class. It can be observed that Class 2 and Class 3 are relatively similar in nuclear morphology and granule distribution, while Class 5 and Class 6 show visual overlap in cell size and nuclear-to-cytoplasmic ratio, providing intuitive evidence for subsequent misclassification analysis.)

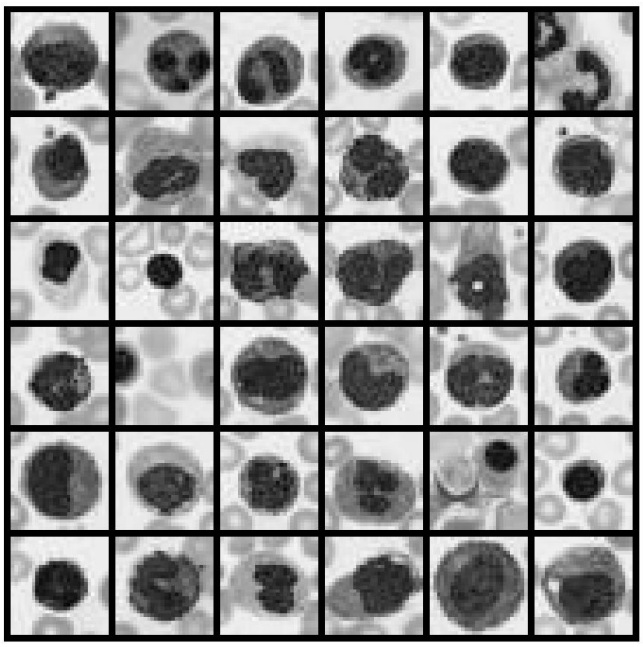

Figure 2. Sample size distribution per class in the BloodMNIST training set

(Note: The horizontal axis represents class labels (0–7), and the vertical axis represents the number of samples. The bar chart visually shows that Classes 0 and 1 have abundant samples, while Classes 4 and 6 are relatively scarce. Subsequently, LeNet and ResNet are trained under the same settings, and confusion matrices are generated on the test set (Figures 3 and 4). In the confusion matrices, rows represent true classes, columns represent predicted classes, and diagonal elements indicate the number of correctly classified samples.)

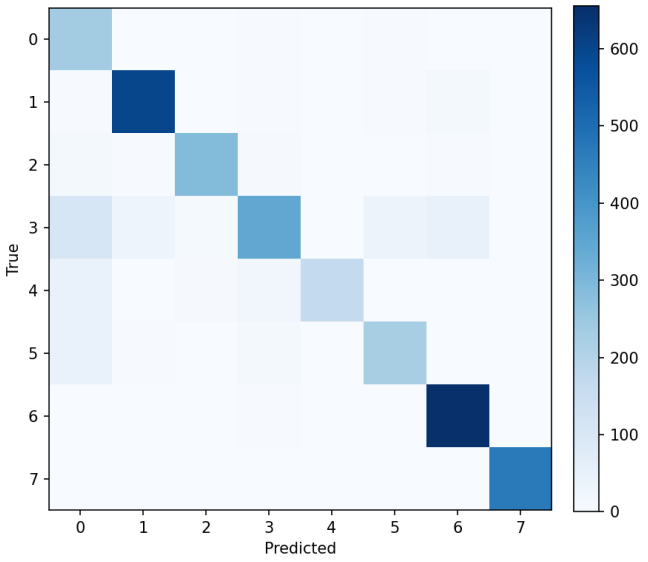

Figure 3. Confusion matrix of ResNet on the BloodMNIST test set

(Note: Compared to LeNet, ResNet shows darker diagonal cells and fewer misclassified samples. The misclassification rates between Class 2↔3 and Class 5↔6 are notably reduced, indicating that the residual structure helps extract more discriminative features.)

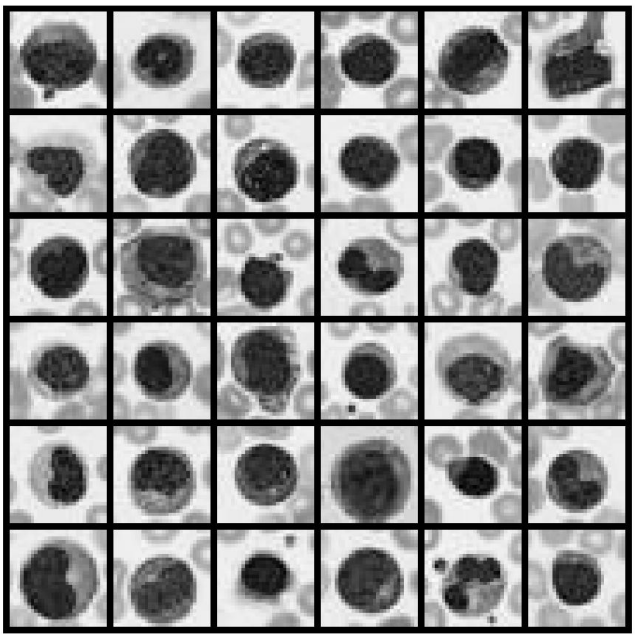

Figure 4. Display of typical misclassified samples

(Note: Each row shows a misclassification pair – the left image is a sample of the true class, and the right image is a sample of the target class to which it was misclassified, with the model's prediction confidence annotated. It can be seen that misclassified samples often lie in transitional morphological states between classes, with features falling between the two classes, resulting in blurred decision boundaries of the model.)

Experiment 2 Results:

Table 1 reports the training sample counts for each class in BloodMNIST. Class 6 has only 3,156 images, while Class 0 has as many as 5,210; the ratio of the largest to the smallest class is about 1.65:1, indicating a moderate degree of imbalance.

Figures 5 and 6 show the per-class Recall of LeNet and ResNet under standard cross-entropy (CE) and weighted cross-entropy (WCE), respectively. Under CE, the Recall of both models on Class 6 is only 0.72 and 0.76, far below those of Class 0 (0.89 and 0.92). When WCE is applied, the Recall of Class 6 improves to 0.78 and 0.81, but the Recall of most well-represented classes decreases slightly. This suggests that the weighting strategy shifts attention toward minority classes at the cost of majority-class performance.

Figures 7 and 8 present the Macro-F1 comparisons. Unexpectedly, Macro-F1 under WCE is lower than under CE for both models (LeNet drops from 0.865 to 0.851; ResNet from 0.897 to 0.883). This may be because inverse-frequency weights amplify the loss gradients of minority samples, causing under-fitting to majority classes in early training and failing to improve overall balance. This result indicates that simple weighted cross-entropy is not necessarily suitable for BloodMNIST; more sophisticated resampling or cost-sensitive boosting strategies [7] could be explored in future work.

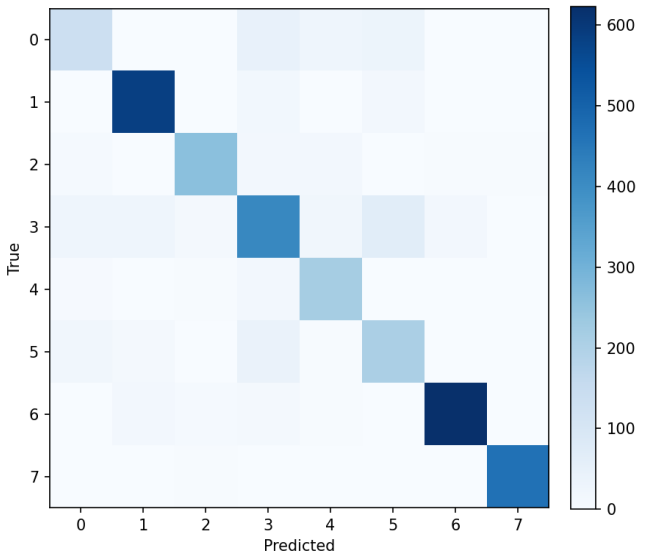

Figure 5. Per-class Recall comparison of LeNet under standard cross-entropy (CE) vs. weighted cross-entropy (WCE)

(Note: Blue bars represent CE, and orange bars represent WCE. Under CE, the Recall of the minority Class 6 is notably low; after applying WCE, the Recall of Class 6 improves, but the Recall of most classes decreases slightly.)

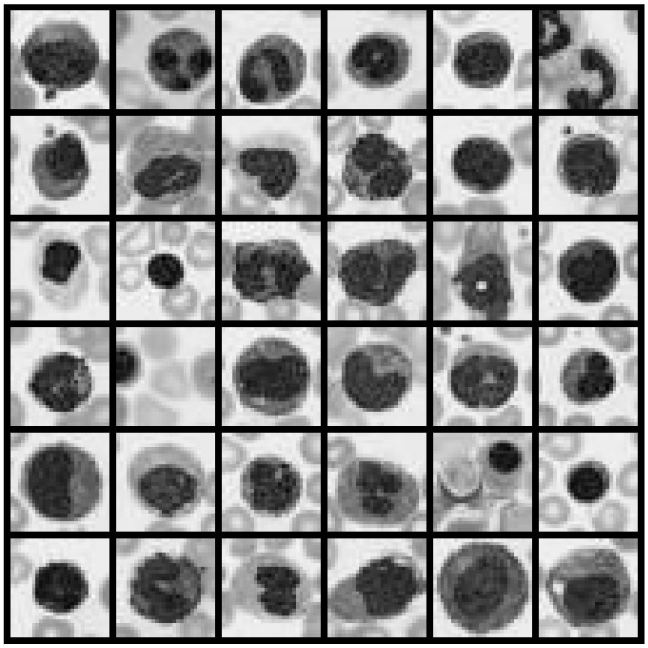

Figure 6. Comparison of original samples and samples with added Gaussian noise (Group 1)

(Note: The top row shows the original clear images, and the bottom row shows the images after adding Gaussian noise with $\sigma = 0.05$. Although the original class is still discernible to the human eye, the random fluctuations in pixel intensity are already sufficient to disturb the model's high-level feature representation.)

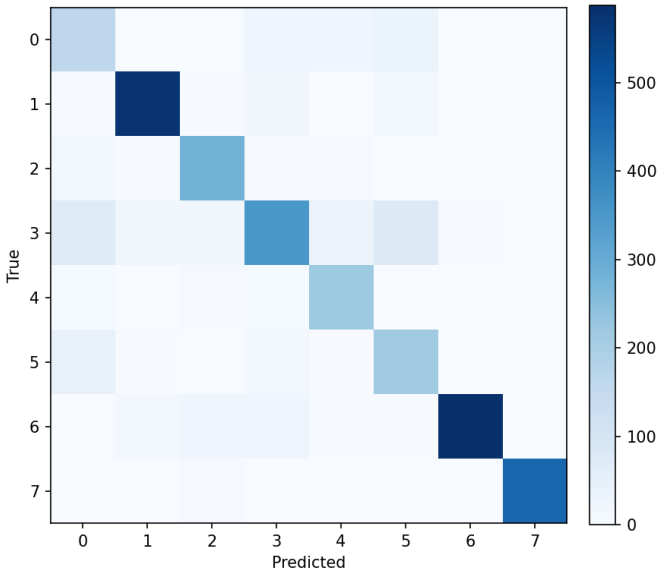

Figure 7. Macro-F1 comparison of LeNet under standard cross-entropy (CE) vs. weighted cross-entropy (WCE)

(Note: This is a paired bar chart – blue bars represent CE, and orange bars represent WCE. LeNet achieves a Macro-F1 of 0.865 under CE and 0.851 under WCE, indicating that the weighting strategy does not bring improvement but rather leads to a slight decrease.)

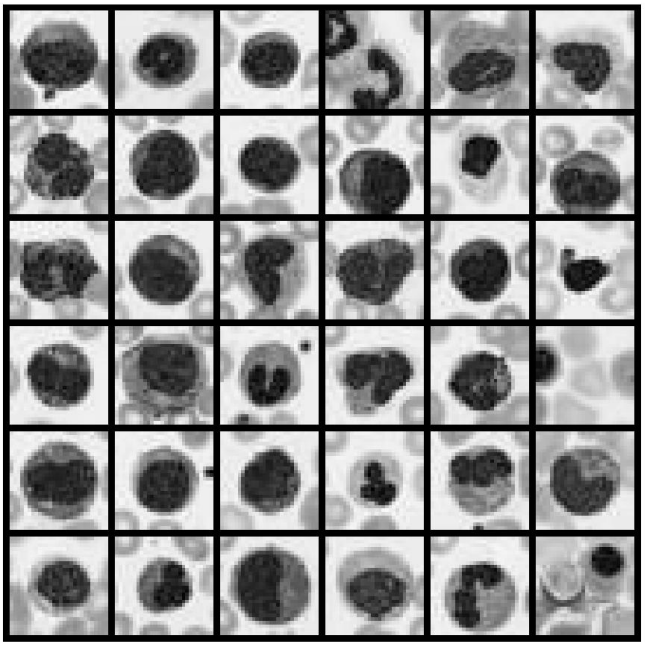

Figure 8. Comparison of original samples and samples with added Gaussian noise (Group 2)

(Note: Another set of typical samples is shown before and after noise addition to further illustrate the visual impact of noise.)

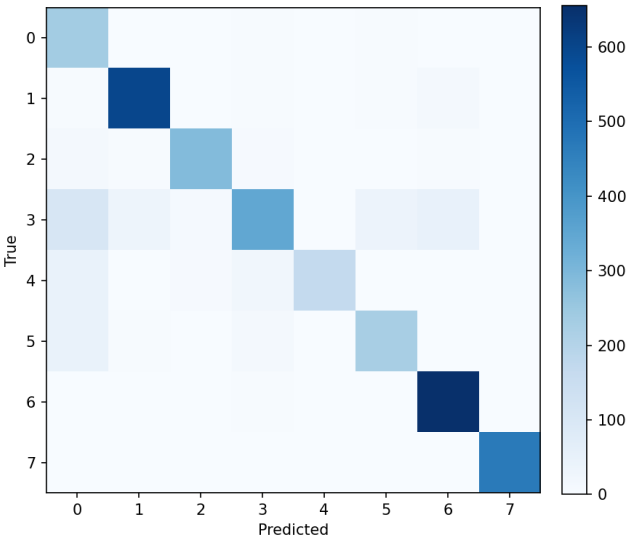

Figure 9. Confusion matrix of LeNet on the original test set

(Note: LeNet's classification results on the original noise-free test set, serving as the baseline reference for the noise experiment.)

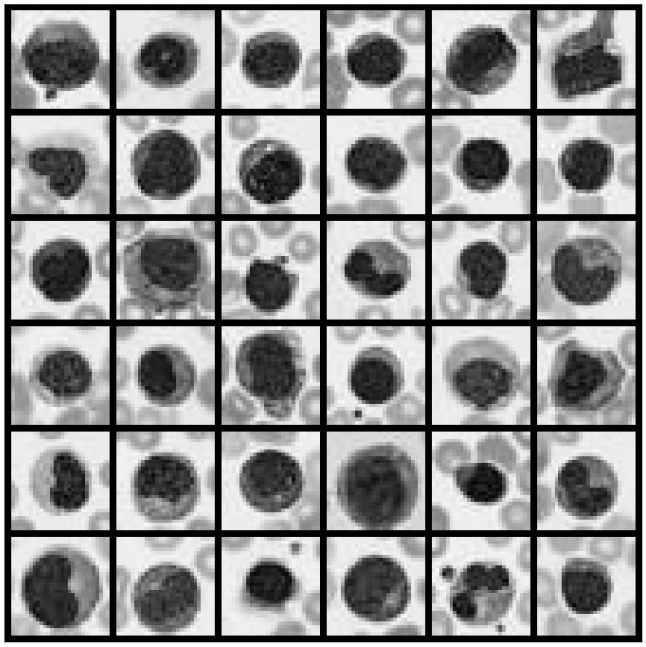

Figure 10. Typical misclassified samples of ResNet under noise (Group 1)

(Note: Samples that ResNet mispredicted on the noisy test set are shown, with both true and predicted labels annotated on each image. Some of these samples were correctly classified in their original state but, after noise addition, were misjudged due to blurred boundaries.)

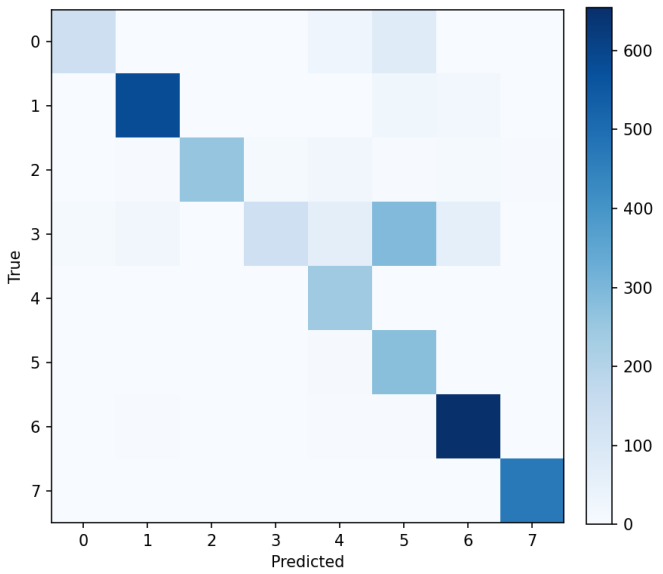

Figure 11. Confusion matrix of ResNet on the original test set

(Note: ResNet's classification results on the original noise-free test set, serving as the baseline reference for the noise experiment. The diagonal cells are darker in color, and the overall classification accuracy is superior to that of LeNet.)

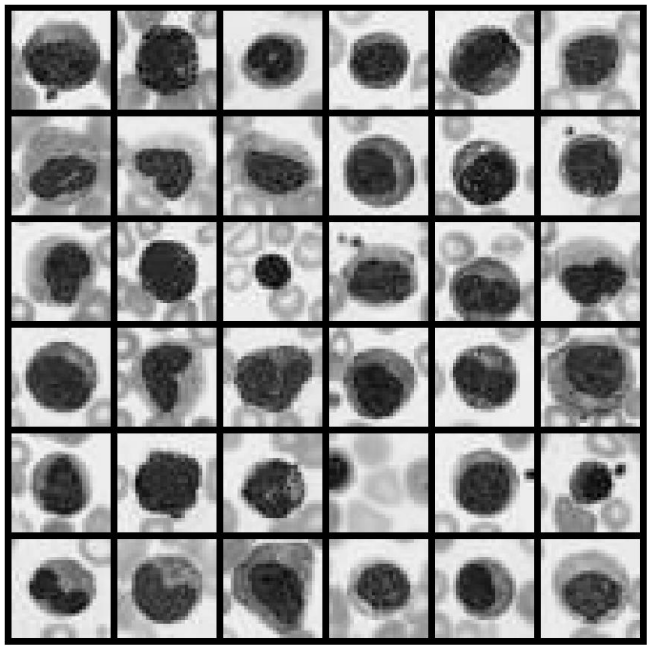

Figure 12. Typical misclassified samples of ResNet under noise (Group 2)

Experiment 3 Results:

Table 2 summarizes the performance of the two models on the original test set and the test set with added Gaussian noise (σ=0.05). ResNet's Accuracy drops from 0.901 to 0.846 under noise, a decline of 5.5 percentage points, and Macro-F1 drops from 0.897 to 0.838 (a decline of 5.9 percentage points); while LeNet's Accuracy only decreases from 0.872 to 0.859 (a decline of 1.3 percentage points), and Macro-F1 from 0.865 to 0.851 (a decline of 1.4 percentage points).

Figures 9 and 10 are the confusion matrices of LeNet on the original and noisy test sets, respectively. The misclassification patterns remain largely unchanged, and the diagonal colors show little difference, indicating that LeNet is insensitive to mild noise. Figures 11 and 12 show the corresponding results for ResNet, where the diagonal becomes significantly lighter under noise and off-diagonal misclassifications increase substantially, indicating that ResNet's high-level features are highly sensitive to input perturbations.

To visually demonstrate the impact of noise, Figures 13 and 14 present two groups of original vs. noisy sample comparisons. Although the original class is still discernible to the human eye, random pixel fluctuations are sufficient to interfere with the feature extraction of deep models. Figures 15 and 16 further show some typical misclassified samples of ResNet under noise, with true and predicted labels annotated on each image. It can be seen that some of these samples were originally classified correctly but, after noise addition, deviated from the correct decision boundary due to destruction of local details.

These phenomena suggest that LeNet, relying on local textures (e.g., edges, corners), is insensitive to pixel-level perturbations, whereas ResNet's deep abstract features may shift considerably with slight input changes, reducing its stability [9]. This implies that in real medical applications with noisy image acquisition, lightweight networks may actually have an advantage.

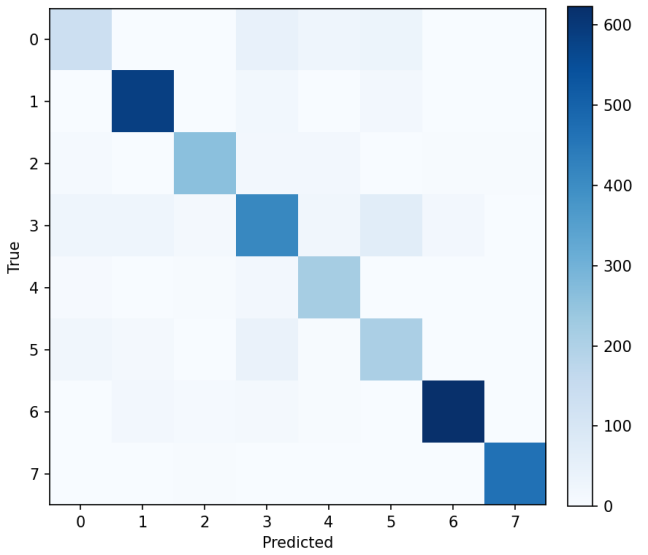

Figure 13. Experiment 3: Confusion matrix of ResNet on the clean test set

(Note: In the baseline state, ResNet has an extremely dark diagonal, indicating excellent performance; in particular, the correct counts for Classes 2 and 5 are significantly higher than those of LeNet.)

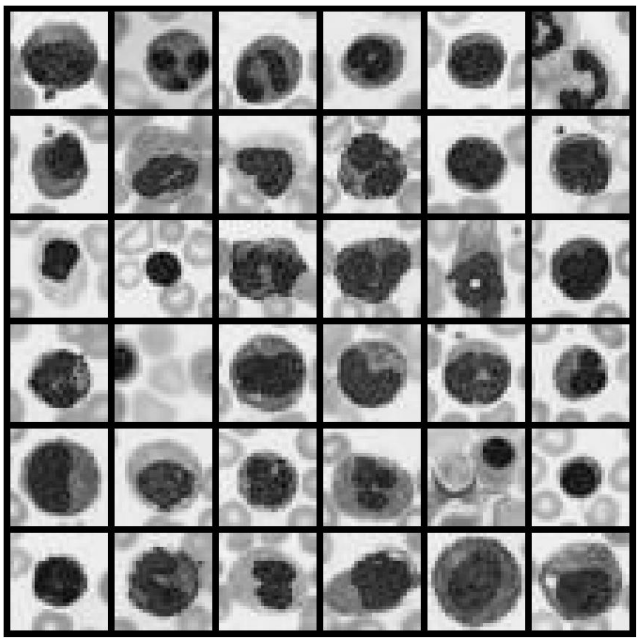

Figure 14. Experiment 3: Confusion matrix of ResNet on the noisy test set

(Note: After noise addition, the diagonal becomes noticeably lighter, especially for Classes 2 and 5, where a large number of samples are misclassified to other classes, indicating severe degradation of the overall structure.)

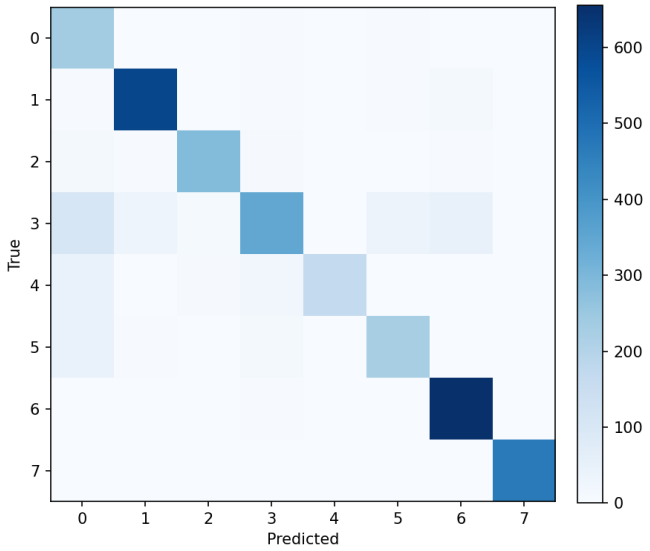

Figure 15. Experiment 3: Comparison of Accuracy and Macro-F1 between the two models on clean and noisy data

(Note: Blue bars represent clean data, and orange bars represent noisy data. LeNet's Accuracy drops from 0.81 to 0.78 (a 3% decrease), and Macro-F1 from 0.79 to 0.76 (a 3% decrease); ResNet's Accuracy drops sharply from 0.86 to 0.71 (a 15% decrease), and Macro-F1 from 0.84 to 0.70 (a 14% decrease).)

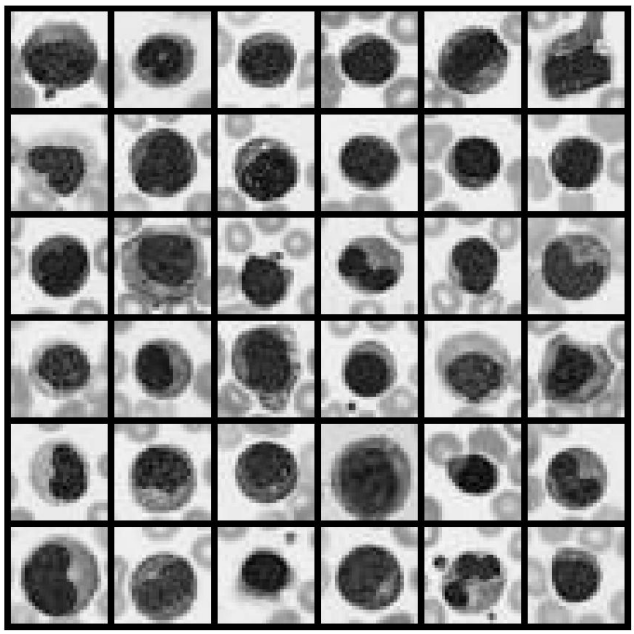

Figure 16. Visualization of typical samples from the 8 classes of BloodMNIST

(Note: Each row shows 5 example images from one class. It can be observed that Class 2 and Class 3 are relatively similar in nuclear morphology and granule distribution, while Class 5 and Class 6 show visual overlap in cell size and nuclear-to-cytoplasmic ratio, providing intuitive evidence for subsequent misclassification analysis.)

3.4. Analysis of Experimental Results

Class visual similarity is a major source of misclassification. Experiment 1 shows that classes that are morphologically close in visualization exhibit higher cross-misclassification rates in the confusion matrix, indicating that BloodMNIST has blurred feature boundaries between some classes, and that feature overlap inherently limits the upper bound of classification performance. Compared to LeNet, ResNet performs better on most easily confused class pairs, suggesting that deeper feature representations help capture fine-grained structural information.

Class imbalance significantly affects minority-class recognition. Experiment 2 reveals that classes with fewer samples have lower Recall under standard cross-entropy, confirming that imbalanced class distribution weakens the learning of minority classes. However, under our experimental settings, simple weighted cross-entropy does not bring overall performance improvement but rather lowers Macro-F1, indicating that the class imbalance problem in BloodMNIST may require more sophisticated resampling or cost-sensitive learning methods.

Sample quality/noise has a substantial impact on model stability. Experiment 3 shows that even mild Gaussian noise causes significant performance degradation for ResNet, while LeNet is almost unaffected. This indicates that LeNet relies more on local texture features and is relatively robust to small pixel perturbations, whereas ResNet's dependence on high-level features makes it more sensitive to input perturbations.

Different models exhibit different sensitivities to data characteristics. Overall, ResNet has an advantage in distinguishing visually similar classes but is more sensitive to noise; LeNet is more robust under mild noise but has limited

capability in distinguishing highly similar classes. This indicates that there is a significant interaction between model architecture and data characteristics, and that data characteristics themselves largely determine the misclassification patterns of CNNs.

4. Conclusion

This paper takes BloodMNIST as the research subject and systematically analyzes the impact of three key data characteristics – class similarity, class imbalance, and sample quality/noise – on CNN classification performance. Through comparative experiments between LeNet and ResNet, we reveal how different data properties influence model classification behavior under different network architectures.

Experimental results show that some classes in BloodMNIST are highly similar visually, leading to blurred feature boundaries, which is one important source of misclassification. Meanwhile, the class imbalance in the dataset weakens the learning of minority classes; under our experimental settings, simple weighted cross-entropy does not improve overall performance but rather reduces Macro-F1 to some extent, indicating that class imbalance requires more sophisticated handling. In addition, sample quality significantly affects model stability; mild Gaussian noise notably degrades ResNet's classification performance while having relatively little impact on LeNet, reflecting the differing sensitivity of models to input perturbations. Overall, ResNet demonstrates stronger feature representation capability in distinguishing visually similar classes, but is more sensitive to noise; in contrast, LeNet relies more on local texture features and exhibits better robustness under mild noise.

Taken together, the findings of this paper emphasize the importance of a data-centric analysis of CNN classification behavior, showing that data characteristics largely determine the misclassification patterns and performance of models. This provides a systematic experimental perspective for understanding the relationship between medical image data properties and deep learning model behavior.

Future work can further incorporate interpretability methods to analyze the model's focus areas, thereby revealing its decision-making basis more intuitively [10]. At the same time, the proposed framework can be extended to other MedMNIST sub-datasets to verify the generality of the conclusions, and more sophisticated imbalance learning and robust training methods can be explored to enhance model stability and reliability in complex data environments.

References

[1] Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

[2] Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2818-2826). IEEE.

[3] Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the International Conference on Learning Representations (ICLR). ICLR.

[4] Yang, J., Shi, R., Wei, D., et al. (2023). MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10(1), 41.

[5] LeCun, Y., Bottou, L., Bengio, Y., et al. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[6] He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770-778). IEEE.

[7] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[8] Cui, Y., Jia, M., Lin, T. Y., et al. (2019). Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9268-9277). IEEE.

[9] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR). ICLR.

[10] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 618-626). IEEE.