More details for IHART

Linwei Yi, Chenhui Cui, Rubing Huang, Dave Towey, and Rongcun Wang

I. Detailed Experimental Data

This section presents a detailed overview of the experimental statistical results obtained under all experimental settings for RQ1 to RQ4. Each experiment was repeated 1000 times to ensure the reliability and accuracy of the results. All experimental outcomes are presented in the form of box plots, where the horizontal line within each box represents the median, and the "□" within the box indicate the mean.

A. RQ1. Which image-hashing strategy delivers the best IHART performance?

In the experiments for RQ1, we evaluated the fault-detection efficiency and effectiveness of the IHART method based on three image-similarity hashing strategies across four datasets and eight deep learning (DL) models. The aHash represents the average image-hashing strategy, dHash represents the difference image-hashing strategy, and pHash represents the perceptual image-hashing strategy. To comprehensively assess the performance of these three methods, we utilized six different dissimilarity (distance) measures, including Euclidean distance (ED), Pearson distance (PD), Cosine distance (CD), Chebyshev distance (CBD), Manhattan distance (MD), and Hamming distance (HD).

1) Failure-Detection Efficiency: As shown in Figure 1, we conducted a detailed statistical analysis of the F-time of the IHART based on three image-hashing strategies across different datasets and DL models. F-time measures the time required for each algorithm to detect the first fault in the testing system, a smaller F-time value indicates that the algorithm can more quickly identify the image test cases that cause the DL system to fail.

2) Failure-Detection Effectiveness: Figure 2 presents a detailed statistical analysis of the F-measure of the IHART method based on three image-hashing strategies across four datasets and eight DL models. The F-measure quantifies the number of test cases generated to identify the first fault in the DL system under test, a lower F-measure value indicates that the algorithm can more effectively detect potential faults in the DL system with fewer image test cases.

Linwei Yi and Chenhui Cui are with the School of Computer Science and Engineering, Macau University of Science and Technology, Taipa, Macau 999078, China. E-mail: 3230002105@student.must.edu.mo, and 2230004387@student.must.edu.mo.

Rubing Huang is with the School of Computer Science and Engineering, Macau University of Science and Technology, Taipa, Macau 999078, China; and also with Macau University of Science and Technology Zhuhai MUST Science and Technology Research Institute, Zhuhai, Guangdong 519099, China. E-mail: rbhuang@must.edu.mo.

Dave Towey is with the School of Computer Science, University of Nottingham Ningbo China, Ningbo, Zhejiang 315100, China. E-mail: dave.towey@nottingham.edu.cn.

Rongcun Wang is with the School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China. E-mail: rcwang@cumt.edu.cn.

B. RQ2. Does IHART significantly outperform ARTDL in terms of fault-detection efficiency?

In the experiments for RQ2, we evaluated the fault-detection efficiency of IHART and ARTDL across four datasets and eight DL models. To ensure a comprehensive assessment, we employed six dissimilarity (distance) measures to compare the performance of the two algorithms. Detailed experimental results are presented in Figure 3.

C. RQ3. Does IHART significantly outperform ARTDL in terms of fault-detection effectiveness?

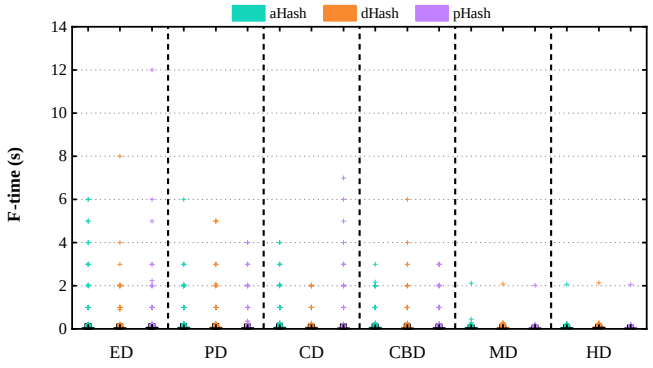
In RQ3, we continued the experimental setup from RQ2 to evaluate the fault-detection effectiveness of IHART and ARTDL. Detailed experimental results are presented in Figure 4.

D. RQ4. Can IHART maintain stable performance on single-category datasets within which the image-classification features are relatively similar?

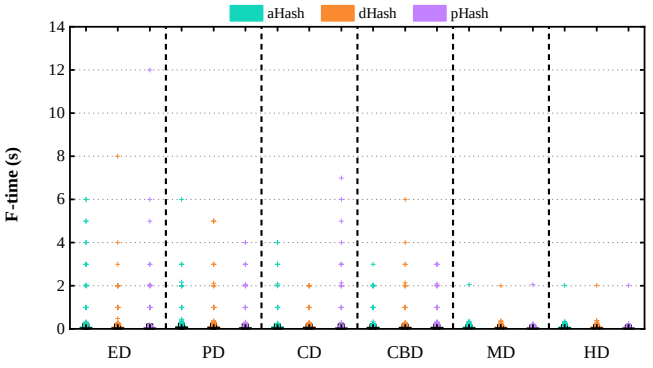
In the experiments for RQ4, we evaluated the fault-detection efficiency and effectiveness of the IHART and ARTDL algorithms on subsets of the `MNIST` and `CIFAR-10` datasets. The `MNIST` dataset comprises ten subsets of handwritten digits from "0" to "9", while `CIFAR-10` is similarly divided into ten subsets (including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck). Notably, in subset "1" of `MNIST`, both algorithms exhibited significantly higher F-time and F-measure values compared to other subsets. Consequently, the statistical results for this subset are displayed on the right side of the box plots in Figures 5(a), 5(b), 6(a), and 6(b), with reference to the y-axis on the right.

1) Failure-Detection Efficiency: Figure 5 presents a detailed statistical analysis of the F-time for IHART and ARTDL across various subsets of datasets and DL models.

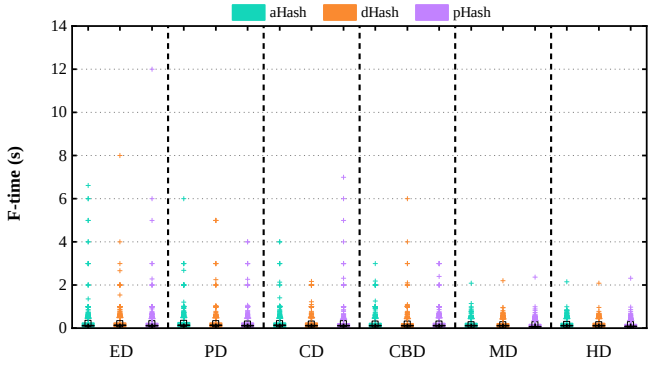
2) Failure-Detection Effectiveness: The detailed statistical analysis of the F-measure for IHART and ARTDL across different subsets of datasets and DL models is shown in Figure 6.



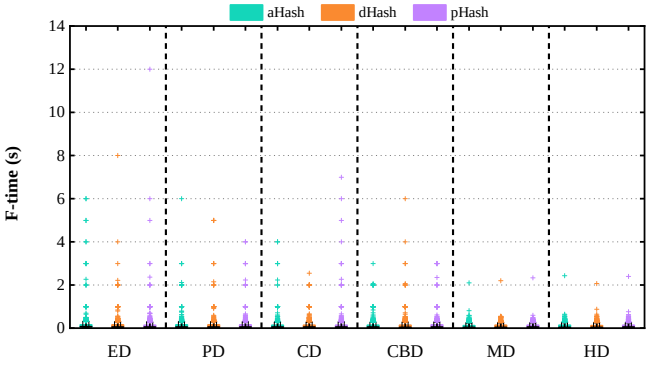
(a) MNIST dataset with LeNet-1 model



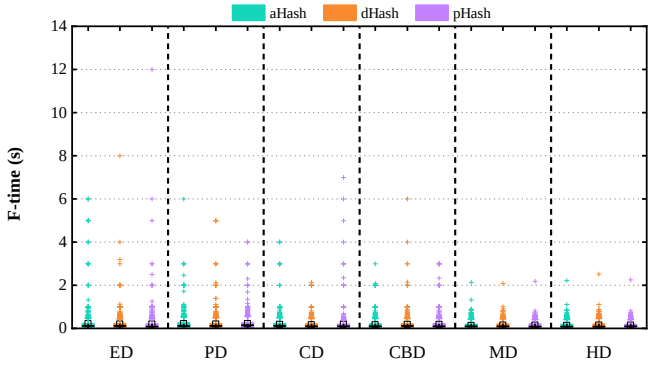
(b) MNIST dataset with LeNet-5 model



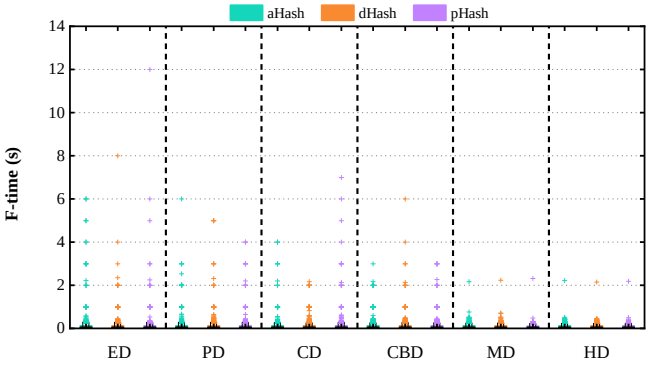
(c) CIFAR-10 dataset with ResNet-18 model



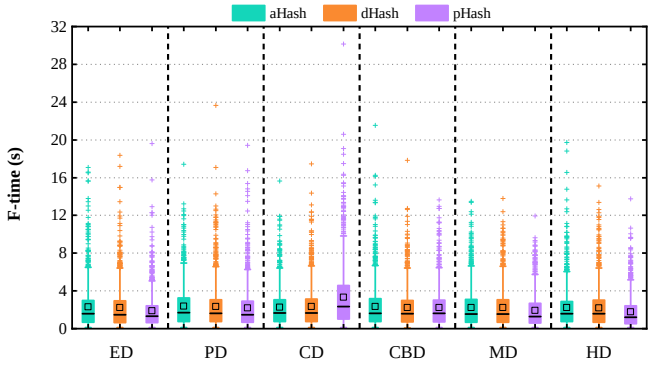
(d) CIFAR-10 dataset with ResNet-20 model



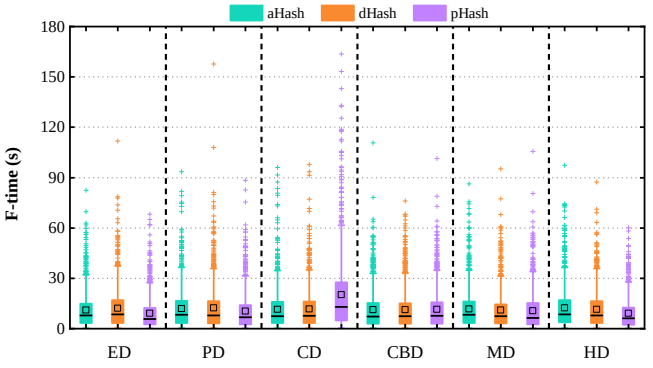
(e) SVHN dataset with ResNet-32 model



(f) SVHN dataset with ResNet-50 model

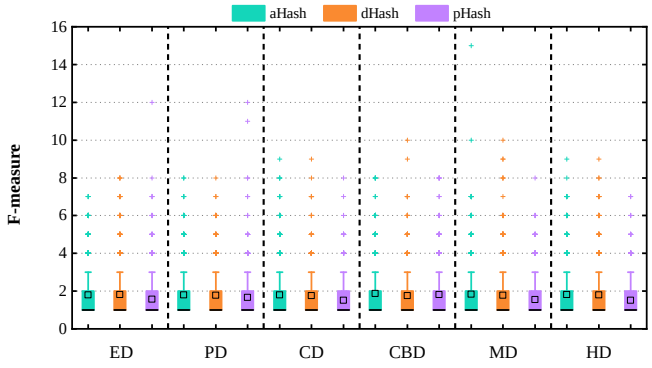


(g) DRIVING dataset with Rambo model

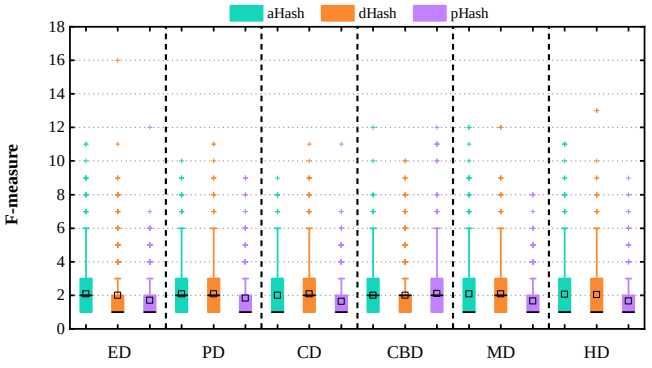


(h) DRIVING dataset with Chauffeur model

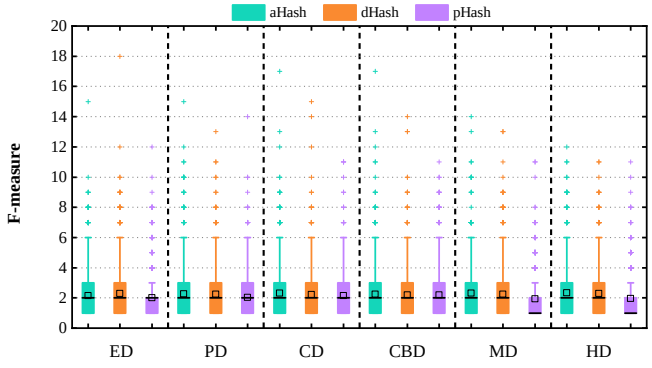
Fig. 1. F-time statistics of IHART based on three image-hashing strategies across different datasets and models.



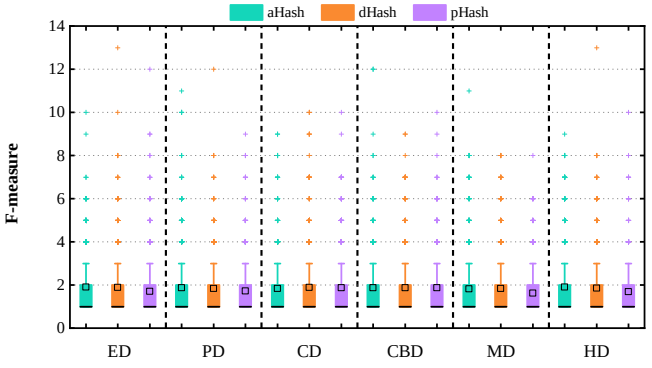
(a) MNIST dataset with LeNet-1 model



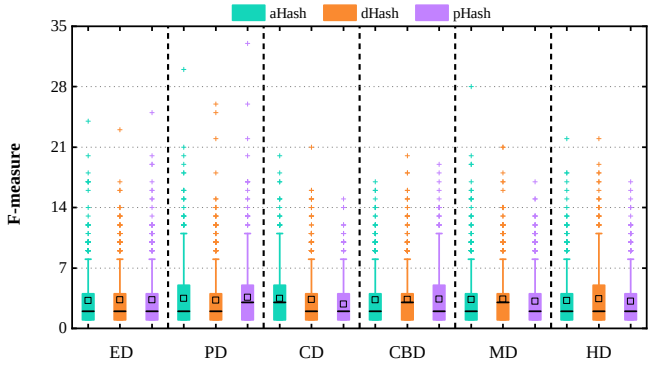
(b) MNIST dataset with LeNet-5 model



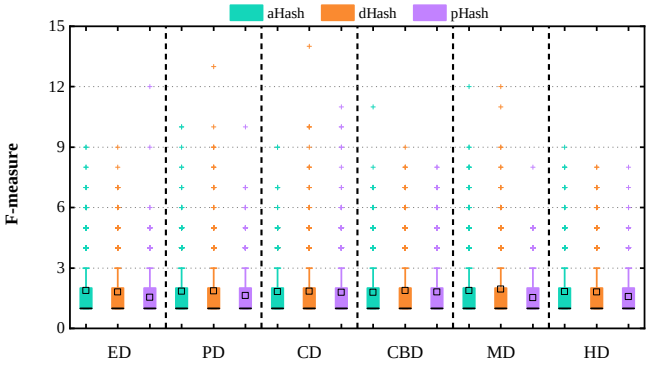
(c) CIFAR-10 dataset with ResNet-18 model



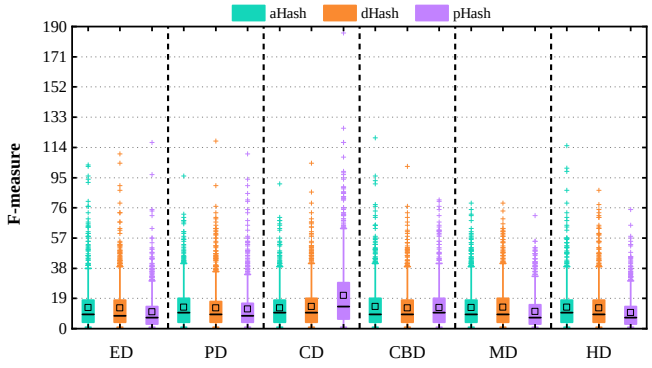
(d) CIFAR-10 dataset with ResNet-20 model



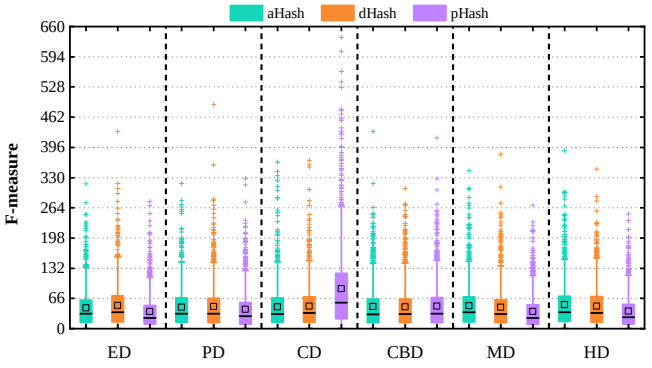
(e) SVHN dataset with ResNet-32 model



(f) SVHN dataset with ResNet-50 model

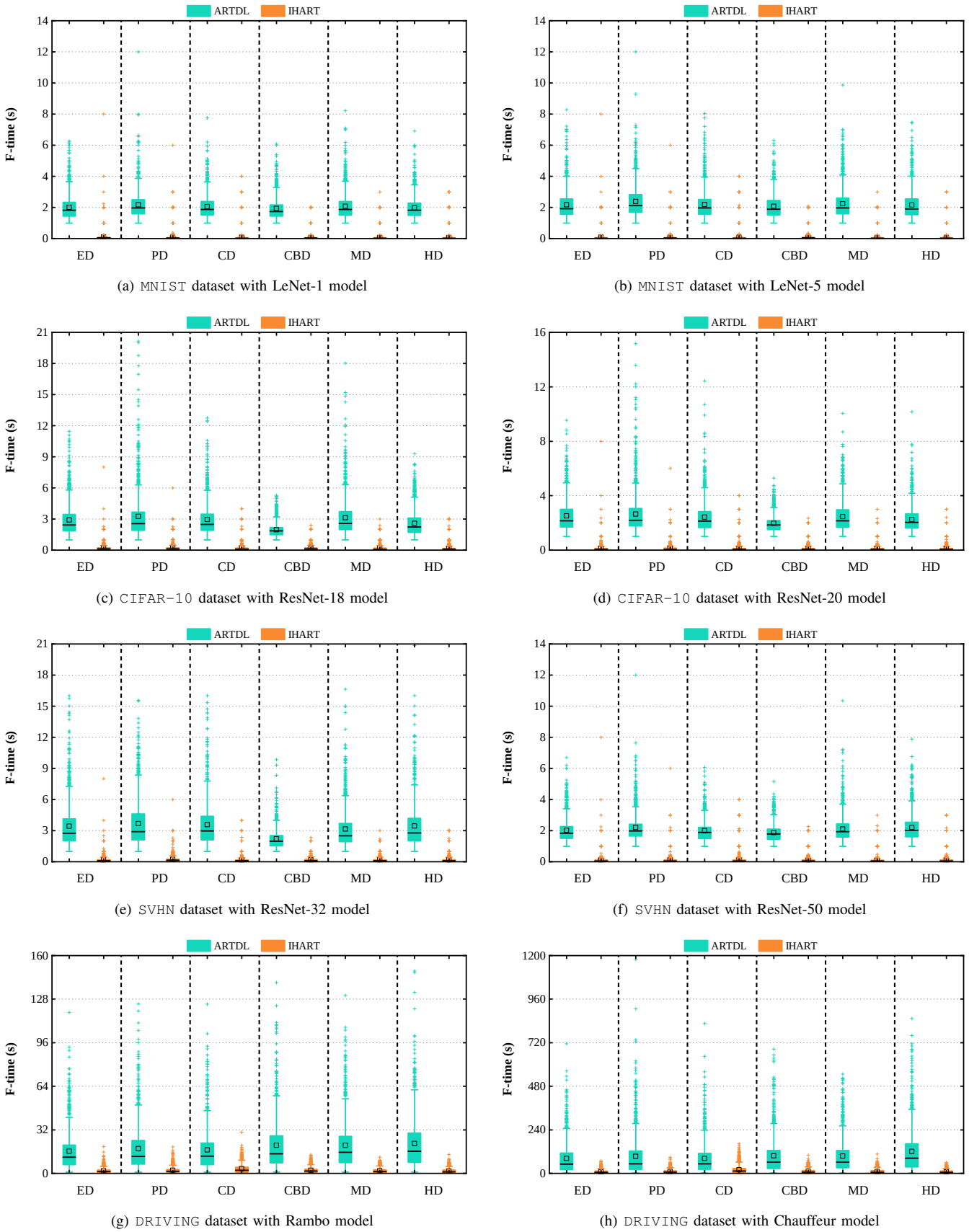


(g) DRIVING dataset with Rambo model



(h) DRIVING dataset with Chauffeur model

Fig. 2. F-measure statistics of IHART based on three image-hashing strategies across different datasets and models.



(a) MNIST dataset with LeNet-1 model

(b) MNIST dataset with LeNet-5 model

(c) CIFAR-10 dataset with ResNet-18 model

(d) CIFAR-10 dataset with ResNet-20 model

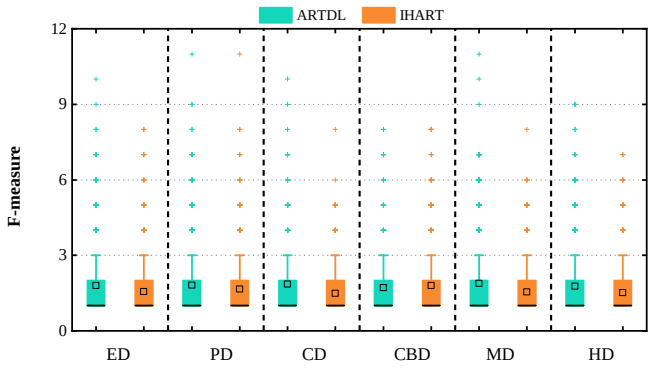
(e) SVHN dataset with ResNet-32 model

(f) SVHN dataset with ResNet-50 model

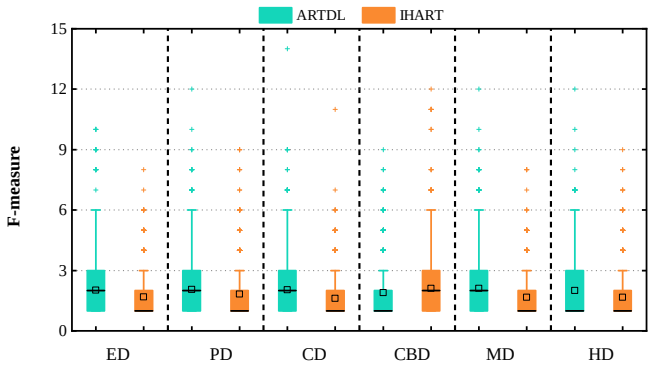
(g) DRIVING dataset with Rambo model

(h) DRIVING dataset with Chauffeur model

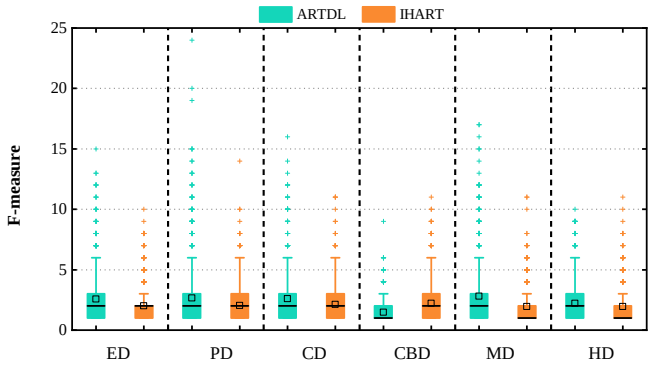
Fig. 3. F-time comparison between image hashing-based ART (IHART) and ARTDL across different datasets and models.



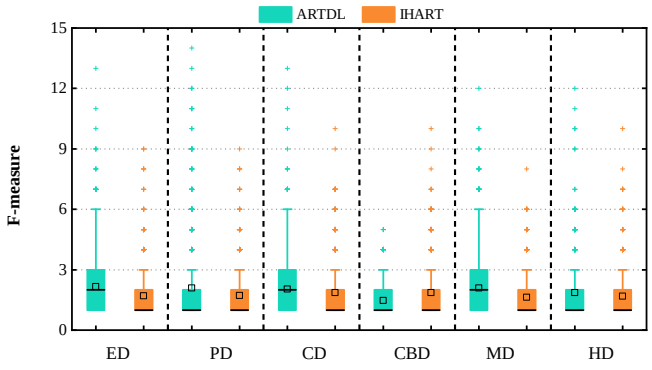
(a) MNIST dataset with LeNet-1 model



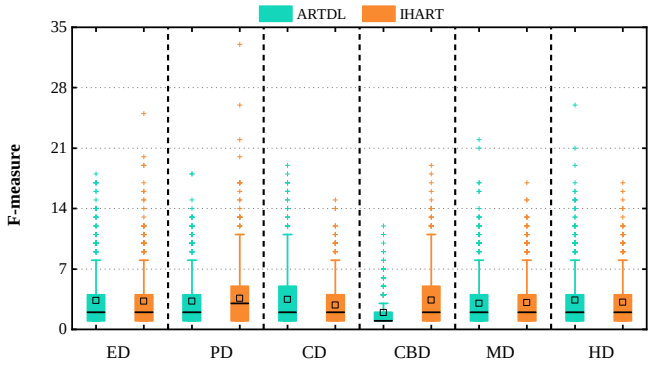
(b) MNIST dataset with LeNet-5 model



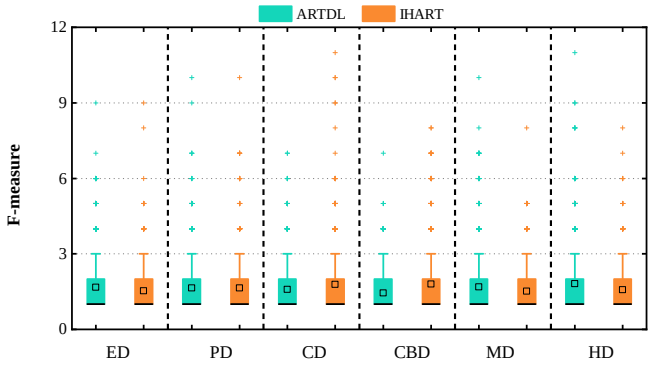
(c) CIFAR-10 dataset with ResNet-18 model



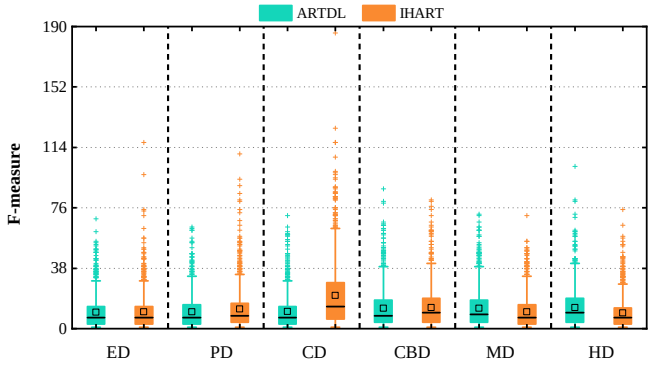
(d) CIFAR-10 dataset with ResNet-20 model



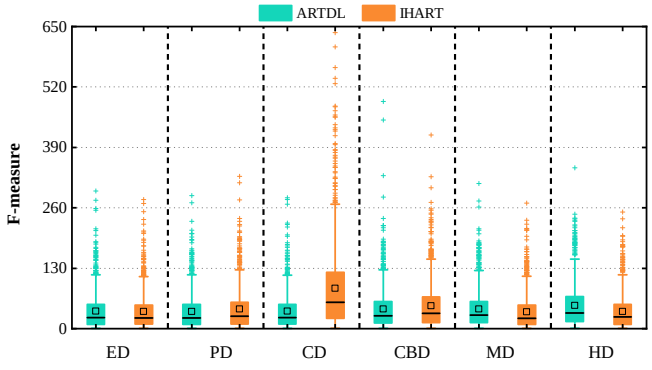
(e) SVHN dataset with ResNet-32 model



(f) SVHN dataset with ResNet-50 model

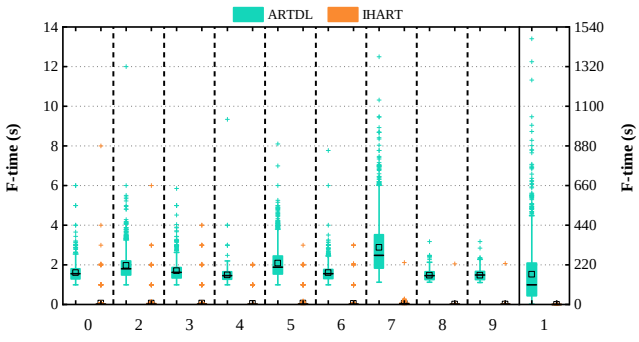


(g) DRIVING dataset with Rambo model

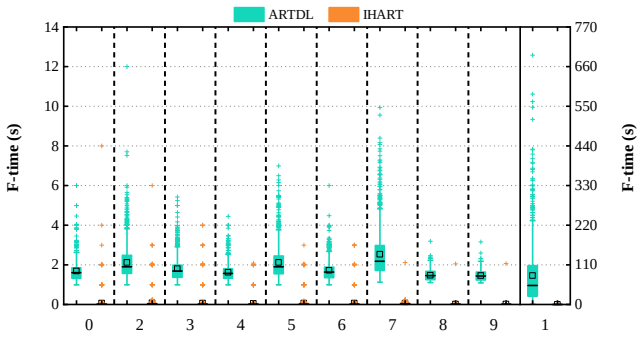


(h) DRIVING dataset with Chauffeur model

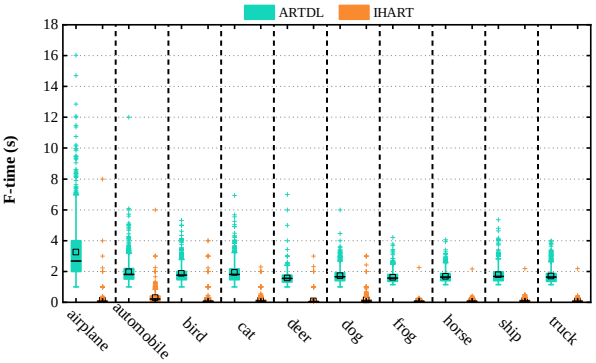
Fig. 4. F-measure comparison between image hashing-based ART (IHART) and ARTDL across different datasets and models.



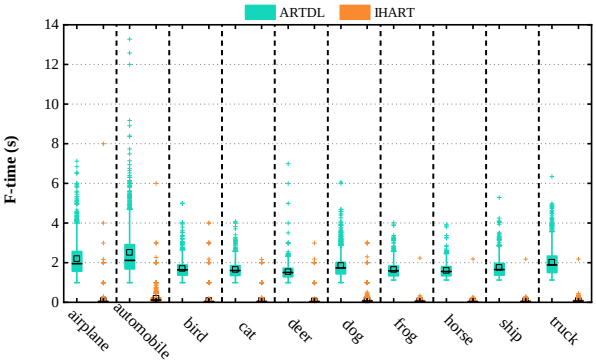
(a) 10 subsets of MNIST with LeNet-1 model



(b) 10 subsets of MNIST with LeNet-5 model

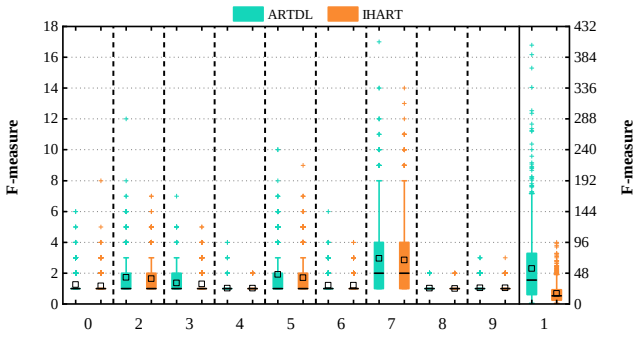


(c) 10 subsets of CIFAR-10 with ResNet-18 model

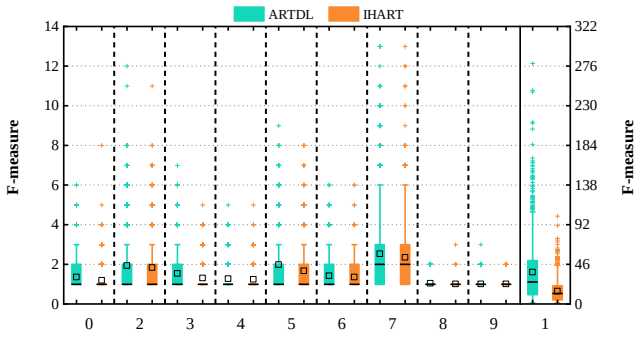


(d) 10 subsets of CIFAR-10 with ResNet-20 model

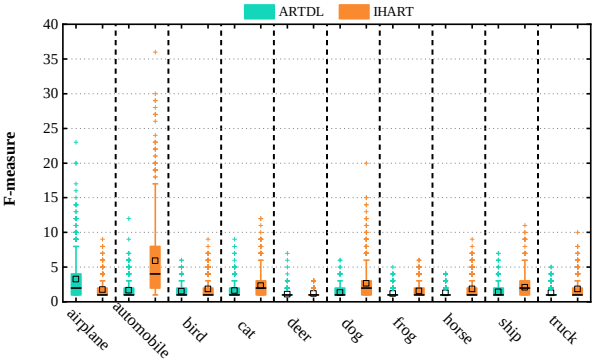
Fig. 5. F-time comparison between image hashing-based ART (IHART) and ARTDL across different datasets and models.



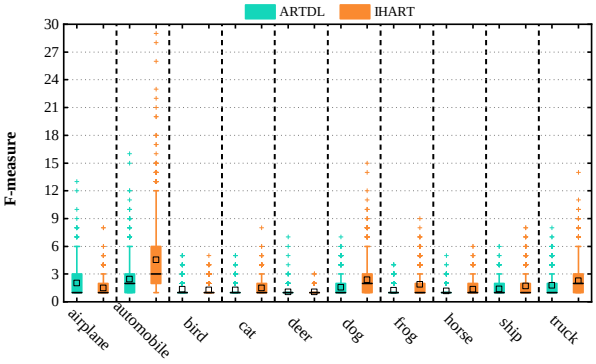
(a) 10 subsets of MNIST with LeNet-1 model



(b) 10 subsets of MNIST with LeNet-5 model



(c) 10 subsets of CIFAR-10 with ResNet-18 model



(d) 10 subsets of CIFAR-10 with ResNet-20 model

Fig. 6. F-measure comparison between image hashing-based ART (IHART) and ARTDL across different datasets and models.