Journal of Information Systems Engineering and Management
2025, 10(16s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Performance Evaluation of Enhanced Deduplication Model with Image Augmentation using Deep Learning (IDME-IR)

Priya Vij [1], Dr. Dalip [2]
[1,2] M.M. Institute of Computer Technology & Business Management
Maharishi Markandeshwar (Deemed to be University), Mullana Ambala, Haryana, India
Vijpriya1@gmail.com

ARTICLE INFO	ABSTRACT

Received: 01 Dec 2024
Revised: 25 Jan 2025
Accepted: 08 Feb 2025

This study aims to evaluate the performance of the Image Duplicate Matching and Elimination - Image Retrieval (IDME-IR) Deduplication Model and an image augmentation method for image retrieval across key parameters such as accuracy, precision, recall, and F1-score. The IDME-IR Deduplication Model focuses on eliminating redundant or near-duplicate images from large datasets, ensuring a cleaner and more efficient retrieval process. Meanwhile, image augmentation techniques are employed to enhance dataset diversity, improving the robustness of retrieval systems by simulating real-world variations in lighting, orientation, and noise. Both methods are evaluated within the context of image retrieval tasks, with the performance metrics being computed across various datasets. After Evaluating the Performance of IDME-IR with Image Augmentation we get accuracy 93.55%, Precision 92.1%, F1-Score 92.7% and Recall 93%. Similarly, without applying image augmentation techniques the result has been observed as accuracy 84.2%, Precision 85.3%, F1-Score 84.7% and Recall 85.4%.

Keywords: IDME-IR; Image Augmentation; Accuracy; Precision; Recall; F1-score.

1. INTRODUCTION

In modern deep learning applications, image augmentation plays a significant role in tasks like object detection, image classification, and semantic segmentation. It introduces variability into training data without needing additional manual annotations, allowing models to better recognize features in varied environments and lighting conditions [1][2]. With advancements in techniques like random transformations, mixup, and GAN-based augmentations, this practice continues to evolve, enabling more robust and accurate machine learning models.

In the realm of image retrieval, the quality and size of datasets play a pivotal role in the accuracy and efficiency of retrieval systems. Two critical techniques—deduplication and image augmentation—have been developed to improve dataset integrity and performance. The IDME-IR Deduplication Model is designed to address the issue of redundant or near-duplicate entries in large datasets, which can lead to inefficiencies and skewed retrieval results. By identifying and removing these duplicates, the model ensures that the dataset remains clean and concise, allowing retrieval algorithms to focus on unique entries.

On the other hand, image augmentation involves applying a range of transformations to images, such as rotation, scaling, and noise injection, to artificially increase the diversity of a dataset [3]. Augmentation helps machine learning models generalize better to new, unseen data by simulating real-world variations in lighting, orientation, and background noise. This technique is particularly valuable in image retrieval systems, where robustness to varying conditions is essential for accurate performance. Figure 1 represents the classical image data augmentation taxonomy.

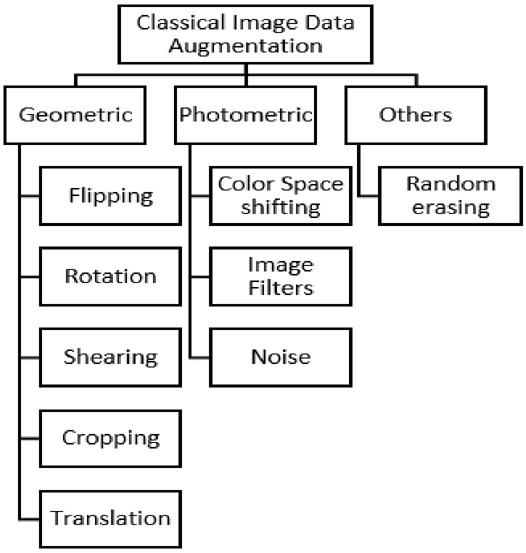

Figure 1: Classical image data augmentation taxonomy [4][5]

Figure 1 represents a systematic categorization of traditional methods used to augment image datasets. These methods include geometric transformations like flipping, rotation, scaling, and cropping, as well as pixel-level adjustments such as brightness, contrast, and color changes. Such techniques are employed to enhance dataset variability, improving the robustness and generalization of machine learning models. Table 1 shows the basic image manipulations and concise description

Table 1: Basic image manipulations and concise description [6][7].

Methods for Image Augmentation	Description
Flipping Flip	The image horizontally, vertically, or both.
Flip Rotation	Rotate the image at an angle.
Scaling Ratio	Increase or reduce the image size.
Noise injection	Add noise into the image.
Color space	Change the image color channels.
Contrast Change	The image contrast.
Sharpening	Modify the image sharpness.
Translation	Move the image horizontally, vertically, or both.
Cropping	Crop a sub-region of the image.

This study aims to evaluate the combined effects of the IDME-IR Deduplication Model and image augmentation on image retrieval performance. The evaluation is based on key metrics including accuracy, precision, recall, and F1-score. By analyzing these metrics, the study seeks to determine how deduplication and augmentation impact retrieval system effectiveness, both independently and together. Ultimately, the goal is to provide insights into how these techniques can optimize image retrieval processes, improving both dataset quality and retrieval accuracy for real-world applications. The author's contribution is given below:

- To develop an enhanced deduplication model.
- To evaluate performance of proposed model with existing models.

The rest of the paper is organized as follows. Section 2 presents the various existing deduplication techniques. Section 3 presents our methodology and proposed work. Section 4 explain the results and discussion. Section 5 discuss the conclusions and provides directions for future work.

2. LITERATURE REVIEW

Han et al. [8] introduced a multi-scale feature fusion approach that integrates features extracted from different scales to enhance retrieval performance under diverse image conditions. However, as datasets grow in size and complexity, models often suffer from overfitting and struggle with generalization, particularly in cross-domain retrieval scenarios. This technique has shown significant benefits, especially in situations where obtaining labeled data is challenging or expensive. Data augmentation techniques, including geometric transformations, color modifications, and synthetic image generation using Generative Adversarial Networks (GANs), have demonstrated their effectiveness in improving model robustness by simulating real-world variations during training.

Xu M et al. [9] conducted a comprehensive review of image augmentation algorithms, classifying them into three categories: model-free, model-based, and policy-optimization-based approaches. To assess the objectives of image augmentation, they examined the challenges involved in deploying deep learning models for computer vision tasks and explored the concept of vicinity distribution. Their findings indicated that image augmentation significantly enhances model performance, with various algorithms tailored to address specific challenges. For example, intensity transformations are useful for handling occlusion, while model-based techniques effectively mitigate class imbalance and domain shift. Their analysis suggests that innovative methods can be inspired by emerging challenges, and selecting the appropriate augmentation strategy depends on the unique constraints of a given dataset.

Wang et al. [10] employed different deep network architectures and applied augmentation techniques such as 3D rotation, flipping, scaling, and the addition of random noise. Their experimental results indicated that test-time augmentation improves segmentation accuracy while also providing uncertainty estimation for segmentation outcomes. While numerous data augmentation strategies exist within the deep learning community, optimizing their application remains a critical area of research.

Lewy D et al. [11] focused on two key aspects of data augmentation: mixing images and selecting augmentation policies. In image classification tasks, data augmentation primarily serves to expand training datasets and enhance model robustness by generating diverse variations of images that resemble real-world test cases. It is widely recognized as an effective regularization technique that helps mitigate overfitting during training. Furthermore, the limited availability of annotated training data poses a significant challenge in deep learning applications, particularly for niche domains requiring domain expertise. In such cases, data augmentation plays a crucial role in expanding training datasets without incurring additional annotation costs.

Hrga I et al. [12] analyzed the impact of augmentation techniques on face image classification across datasets of different sizes using two transfer learning approaches. Their study considered both objective and subjective target attributes, revealing that simple affine transformations with minimal intensity did not provide significant benefits. However, image-mixing techniques became increasingly effective as dataset size grew, particularly when fine-tuning models. The results also demonstrated that image mixing had the strongest regularization effect. Nevertheless, due to the uniformity of scenes and the absence of typical challenges like occlusion, techniques involving information dropout did not emerge as a dominant factor, suggesting the need for further research on more complex image datasets.

Jinwoo K [13] highlighted the significance of data augmentation and adversarial learning in improving image retrieval models by enhancing generalization and resilience to perturbations. By generating diverse training samples and incorporating adversarial examples, these methods address key challenges associated with large-scale datasets and real-world variability [14]. However, further advancements are required to develop more adaptive and efficient augmentation strategies.

Ismael et al. [15] applied data augmentation techniques to address data scarcity and class imbalance in MRI image classification for brain cancer. They experimented with various augmentation methods, including flipping (horizontal and vertical), rotation, shifting, zooming, shearing, and brightness adjustments. Their findings revealed that different augmentation techniques influenced class performance differently. For example, brightness adjustment achieved 96% accuracy for one class, while rotation improved accuracy to 98% for the same class. Similarly, brightness and rotation achieved 99% and 98% accuracy, respectively, for another class. By combining all augmentation strategies, they attained a 99% overall accuracy—an improvement of 4% over models trained without augmentation.

Additionally, Gour et al. [16] introduced ResHist, a 152-layer convolutional neural network (CNN) based on residual learning, for breast cancer histopathological image classification. Their data augmentation strategy incorporated stain normalization, image patch generation, and affine transformations to enhance model performance. Experimental results demonstrated that the proposed augmentation techniques significantly improved classification accuracy compared to pre-trained architectures such as AlexNet, VGG16, VGG19, GoogleNet, Inception-v3, ResNet50, and ResNet152.

3. EXISTING DEDUPLICATION TECHNIQUES

In image processing, particularly for tasks like image deduplication, various hashing techniques are commonly employed to efficiently identify duplicate images. Below is an overview of Convolutional Neural Networks (CNNs) and several key image hashing methods, including Perceptual Hashing, Difference Hashing, Wavelet Hashing, and Average Hashing.

- **Convolutional Neural Networks (CNNs)**

Convolutional Neural Networks (CNNs) are deep learning models that specialize in processing image data by learning layered feature representations. In deduplication tasks, CNNs can extract detailed feature embeddings from images, allowing for comparison based on learned patterns rather than exact pixel similarity[18]. By leveraging these embeddings, CNNs can detect duplicates even if images are transformed, such as through resizing or minor rotation. This capability makes CNNs highly effective for large datasets and high-resolution images, where detecting content similarities extends beyond simple visual matching [19].

- **Perceptual Hashing (P Hash)**

Perceptual Hashing (P Hash) generates a distinct fingerprint for each image based on its overall visual characteristics rather than pixel-by-pixel information. By transforming the image into the frequency domain (often through Discrete Cosine Transform), P Hash captures the image's core structure and pattern. This makes it resilient to small transformations like resizing or slight color adjustments, making it effective in deduplication tasks for visually similar, but not identical, images [20][21].

- **Difference Hashing (D Hash)**

Difference Hashing (D Hash) identifies structural elements within an image by comparing pixel intensity differences in a low-resolution grayscale version of the image. This binary hash represents variations in brightness and patterns, resulting in a simplified fingerprint [22]. D Hash is efficient and well-suited for images with minor brightness or contrast shifts, making it useful for detecting duplicates that have undergone small lighting adjustments.

- **Wavelet Hashing (W Hash)**

Wavelet Hashing (W Hash) uses a wavelet transform to evaluate the image across multiple resolution scales, capturing essential structural details within the low-frequency components [23]. This multi-resolution approach makes W Hash robust against compression, noise, and minor occlusions, making it a reliable method for deduplicating images with subtle modifications. It is particularly useful for identifying duplicates in cases where images are compressed or resized [24].

- **Average Hashing (A Hash)**

Average Hashing (A Hash) creates a binary hash by comparing each pixel's intensity to the overall average in a low-resolution grayscale version of the image [25]. This method is computationally efficient and works well for detecting exact duplicates or images with minimal differences. However, A Hash is less robust to transformations like rotation or compression, so it is best suited for identifying straightforward duplicates with only slight changes [26].

4. PROPOSED WORK

The flowchart in Figure 2 illustrates the Image Duplicate Matching and Elimination (IDME-IR) Deduplication Model, which is designed to improve image retrieval performance. Here's a breakdown of the process:

1. **Start with an Image Dataset**: The process begins with inputting an image dataset as the source for deduplication and retrieval improvement.
2. **Feature Extraction (e.g., CNN)**: Features are extracted from images using a Convolutional Neural Network (CNN) or a similar technique, allowing the system to compare image similarities.

3. **Calculate Feature Similarity**: After extracting features, the system calculates the similarity between feature vectors to assess how closely related or duplicated the images are.
4. **Duplicate Detection**: Images with high similarity scores are identified as duplicates.
5. **Deduplicate the Dataset**: The duplicates are removed, leaving a refined image dataset, which boosts storage efficiency and improves retrieval accuracy.
6. **Apply Image Augmentation**: Techniques such as rotation, scaling, and cropping are applied to the images, enhancing the dataset's diversity and making the model more robust for retrieval tasks.
7. **Enhanced Image Retrieval**: With a deduplicated and augmented dataset, the image retrieval system achieves better accuracy and efficiency.

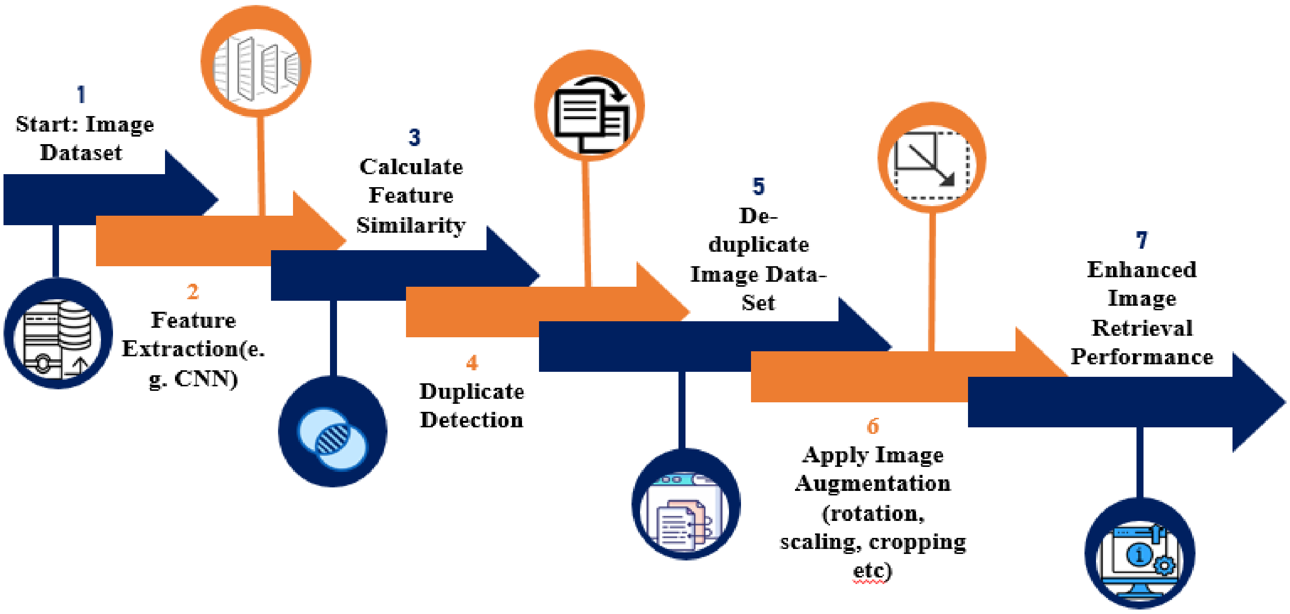

Figure 2: Basic flow chart for IDME-IR Deduplication Model and image augmentation for image retrieval performance

5. TOOLS AND IMPLEMENTATION

- **Framework and Libraries**
  - Deep Learning Framework: Use a framework such as PyTorch or TensorFlow to develop the IDME-IR system, as they provide support for data augmentation, model training, and evaluation.
  - Augmentation Library: Utilize libraries like Albumentations or Torchvision's transforms for a wide range of augmentation techniques.
- **Model Architecture**
  - IDME-IR Implementation: Develop the IDME-IR model with a suitable backbone architecture for feature extraction, such as ResNet or EfficientNet, followed by a layer for similarity measurement (e.g., cosine similarity or a learned metric).
  - Augmentation Pipeline: Configure an augmentation pipeline to apply specified augmentations only to the training images, leaving the validation and test datasets unchanged.
- **Experimental Setup**
  - Hardware: Leverage a GPU (such as NVIDIA V100 or A100) to accelerate the training and evaluation processes.
  - DataLoader: Utilize a DataLoader that incorporates augmentation techniques, batch processing, and parallel data loading.

- Logging and Tracking: Employ experiment tracking tools such as Weights & Biases, TensorBoard, or MLflow to record metrics, visualize results, and compare the performance of augmented and non-augmented models.
- Random Seed Control: Ensure reproducibility by setting random seeds across relevant libraries (e.g., random, numpy, and torch).

Data set- The dataset which we used is part of the ISIC (International Skin Imaging Collaboration) Archive, which hosts publicly accessible dermatology image datasets used for the purpose of developing and evaluating machine learning models for skin lesion analysis, especially in the context of detecting melanoma and other skin conditions. Link- https://challenge.isic-archive.com/data/#2016.

- **Training Data (900 Images)**
  - The training dataset contains 900 dermoscopic images. These images are typically used to train machine learning models in tasks such as classification, segmentation, or detection of skin lesions.
  - The images in this set likely represent a variety of skin lesion types, including benign lesions (non-cancerous) and malignant lesions (such as melanoma).
  - Image augmentation is a common technique applied to training datasets to artificially expand the diversity of the training set. This might include transformations like rotation, scaling, cropping, and color adjustments, which help the model generalize better by simulating different imaging conditions or slight variations in lesion appearances.
- **Test Data (379 Images)**
  - A separate test dataset of 379 images was provided. This dataset is used to measure the performance of systems (machine learning models) that were trained using the 900-image training dataset as shown in figure 3.
  - Unlike the training set, the test set is kept separate during the training process and is used only for final evaluation. This ensures that the model's performance is tested on unseen data, which helps assess its generalization ability to new, real-world examples.

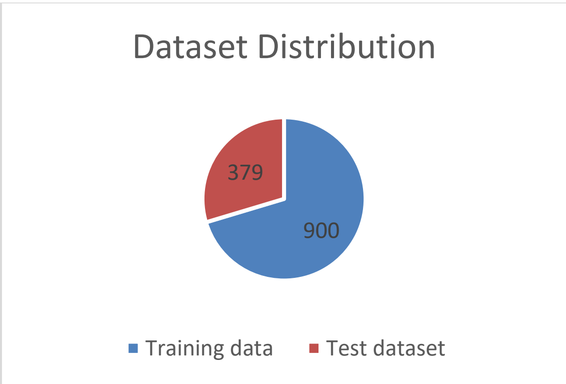

Figure 3: Dataset Distribution

```
// Pseudocode for Image Deduplication and Retrieval Algorithm

// Step 1: Preprocessing Stage
        Input:
          I = {I_1, I_2, ..., I_n}  // Collection of images
             A = {a_1, a_2, ..., a_m}  // Augmentation configurations
// Step 2: Feature Extraction
        For each image I_i in I:
            f_i = F(I_i)          // Extract feature vector using feature extraction model F
// Step 3: Augmentation
        For each image I_i in I:
```

```
                   Apply augmentations A to create variations of I_i
// Step 4: Image Similarity Computation
                 For each pair of images (I_i, I_j) in I:
                         sim(f_i, f_j) = cosine_similarity(f_i, f_j)   // Compute similarity using cosine similarity
// Step 5: Duplicate Detection
                 For each pair (I_i, I_j) in I:
                          If sim(f_i, f_j) > T:      // Threshold T for similarity
                                      Mark I_i and I_j as duplicates
// Step 6: Clustering
                 Use clustering algorithm (e.g., DBSCAN) to group similar images into clusters
// Step 7: Deduplication
                 For each cluster of duplicate images:
                         Select one representative image
                                      Remove the rest from the image set

Output:
   I_dedup = I - Duplicates   // Deduplicated image set

// Step 8: Image Retrieval
               Input: I_q   // Query image

               // Query Processing
                         f_q = F(I_q)        // Extract feature vector for query image
                      f_q_aug = F(A(I_q)) // Extract feature vectors for augmented query images
               // Retrieval
               For each image f_i in Features_dedup:
                  Compare sim(f_q, f_i) and sort based on similarity scores
Output:
   Top-K most similar images

// Step 9: Performance Enhancement via Image Augmentation
                 Match multiple variations of the query image I_q for robustness in retrieval
```

Performance Metrics for IDME-IR

To evaluate the model performance, the following metrics are commonly used [17] see table 2.

Table 2: Performance Metrics used for IDME-IR

Parameter	Definition	Formula
Accuracy	The proportion of correctly identified images, including duplicates and non-duplicates.	$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{All Data}}$
Precision	Measures the fraction of correctly identified duplicates out of all images labeled as duplicates.	$\text{Precision} = \frac{\text{TruePositive}}{\text{All Actual Positives}}$
F1-Score	The harmonic mean of precision and recall.	$\text{F1 Score} = \frac{\text{TruePositive}}{\text{TruePositive} + \frac{1}{2(\text{FalsePositive} + \text{FalseNegetive})}}$
Recall	Measures the fraction of true duplicates that were successfully identified.	$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$

6. RESULTS AND DISCUSSION

The results of the IDME-IR (Image Decomposition and Matching Engine - Image Retrieval) model with image augmentation and without image augmentation evaluation is done on the basis of Accuracy, precision, F1 score and recall. Refer Table 3.

Table 3: IDME-IR with or without Image Augmentation

S.No	Techniques Used	Accuracy	Precision	F1-Score	Recall
1	Image Augmentation with Deduplication Model(IDME-IR)	93.5	92.1	92.7	93
2	With-out Image Augmentation and Deduplication Model(IDME-IR)	84.2	85.3	84.7	85.4

The Figure 4 compared various performance metrics for two different models: the "Image Augmentation with Deduplication Model (IDME-IR)" and the "without Image Augmentation and Deduplication Model (IDME-IR)." The metrics displayed include Accuracy, Precision, F1-Score, and Recall, with specific values represented on the y-axis ranging from 78 to 96. The chart visually contrasts these metrics across both models, highlighting the differences in performance between them.

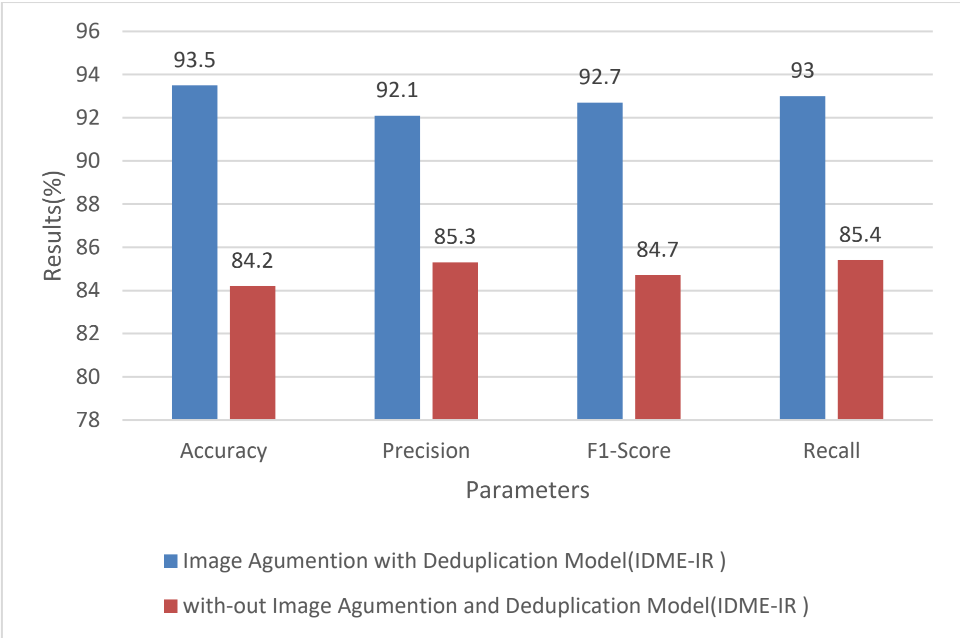

Figure 4: IDME-IR model with or without Image Augmentation

Table 4 provides a comparison of various deduplication techniques such as CNN, P Hash, D Hash, W Hash, A Hash, IDME-IR with image Augmentation and IDME-IR without image Augmentation based on various parameters (Accuracy, Precision, F1-Score and Recall).

Table 4: Comparison of Various Deduplication Techniques

Hashing Algorithm	**Accuracy (in %)**	**Precision (in %)**	**F1-Score (in %)**	**Recall (in %)**
CNN	65	65.80	64.10	63.40
P Hash	61	61.00	62.32	63.00
D Hash	79	78.30	77.25	74.68
W Hash	73	74.20	73.85	71.95
A Hash	68	67.30	67.74	69.21
IDME-IR without image Augmentation	84.20	85.30	84.70	85.40
IDME-IR with image Augmentation	**93.50**	**92.10**	**92.70**	**93.00**

The Figure 5 displays accuracy percentages for various hashing techniques. The x-axis lists the techniques, and the y-axis represents accuracy in percentage from 0% to 100%.The percentage values range from 65% to 100%, indicating the performance of different methods such as CNN, P Hash, D Hash, W Hash, A Hash, and IDME-IR with and without image augmentation. This technical representation allows for a comparative analysis of how each method performs under different conditions. The chart likely serves to visualize the effectiveness of these hashing algorithms in relation to image processing tasks.

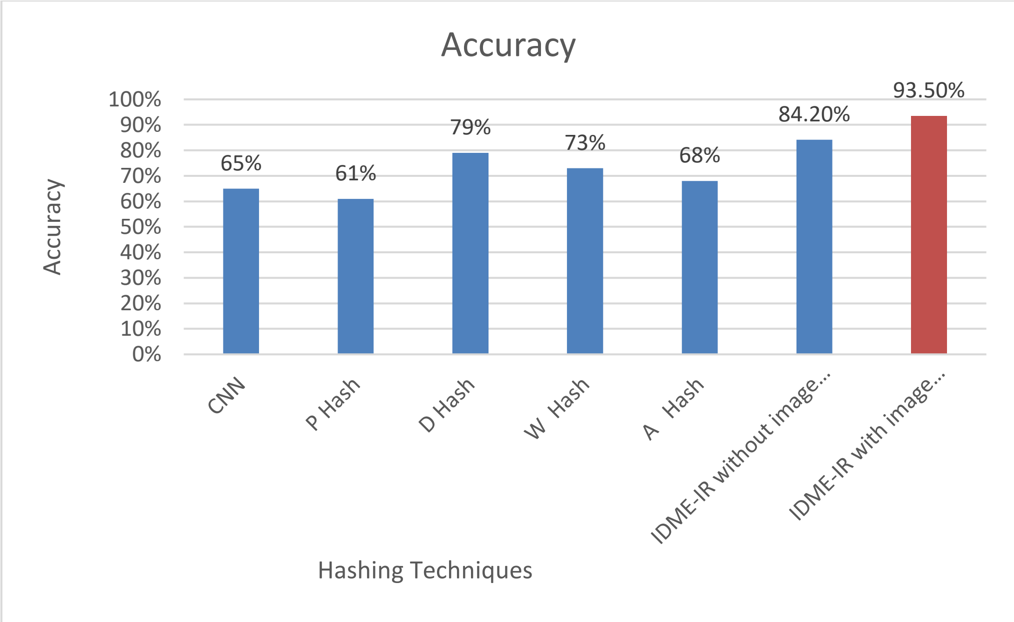

Figure 5: Performance comparison of Hashing Techniques in terms of accuracy

The Figure 6 shows the Precision (in %) for various image deduplication methods. The IDME-IR model with image augmentation has the highest precision at 92.10%, followed by the IDME-IR model without augmentation at 85.30%. Among the hashing methods, D Hash performs the best with 78.30%, followed by W Hash at 74.20% and A Hash at 67.30%. CNN achieves 65.80%, while P Hash has the lowest precision at 61.00%. This illustrates the strong performance of the IDME-IR model, especially with augmentation, in improving precision.

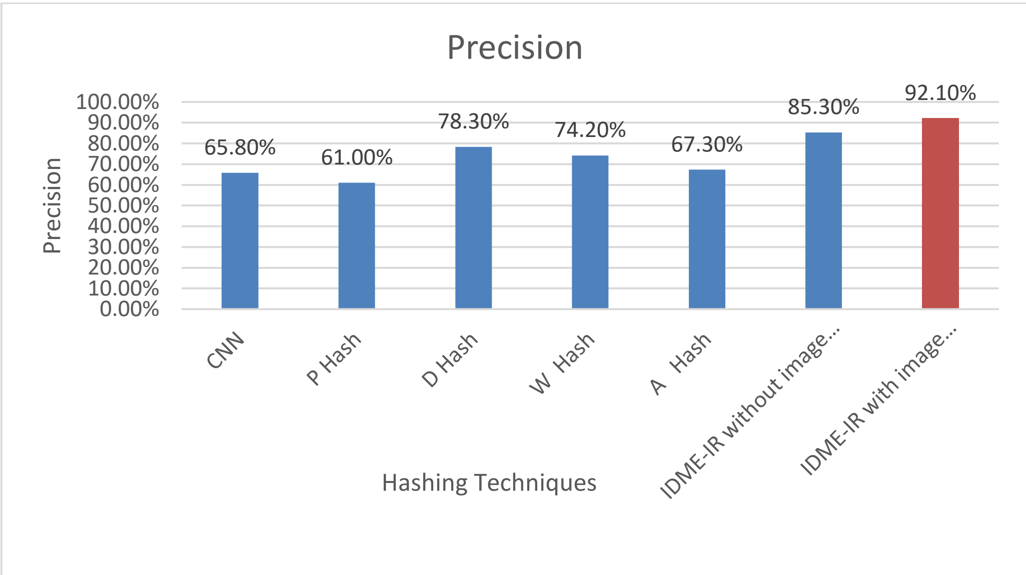

Figure 6: Performance comparison of hashing techniques in terms of Precision

The Figure 7 displays the F1-Score (in %) comparison for different image deduplication methods. The IDME-IR model with image augmentation achieves the highest F1-Score of 92.70%, followed by the IDME-IR model without augmentation at 84.70%. Among hashing methods, D Hash performs best with a score of 77.25%, followed by W Hash at 73.85% and A Hash at 67.74%. CNN and P Hash have the lowest F1-Scores, with 64.10% and 62.32%, respectively. This highlights the superior performance of the IDME-IR model, especially when using image augmentation.

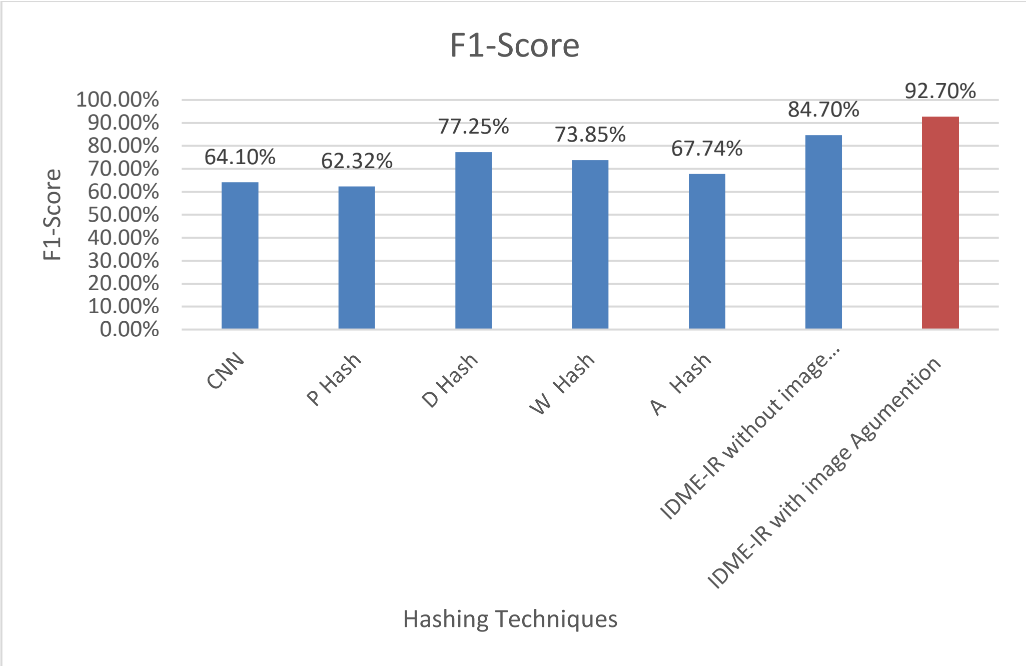

Figure 7: Performance comparison of hashing techniques in terms of F1-Score

The Figure 8 presents a comparison of recall percentages across various methods of image deduplication and retrieval, highlighting the impact of the IDME-IR Deduplication Model and image augmentation techniques. The methods are evaluated based on their performance in terms of recall, which measures the ability to retrieve relevant images from a dataset.

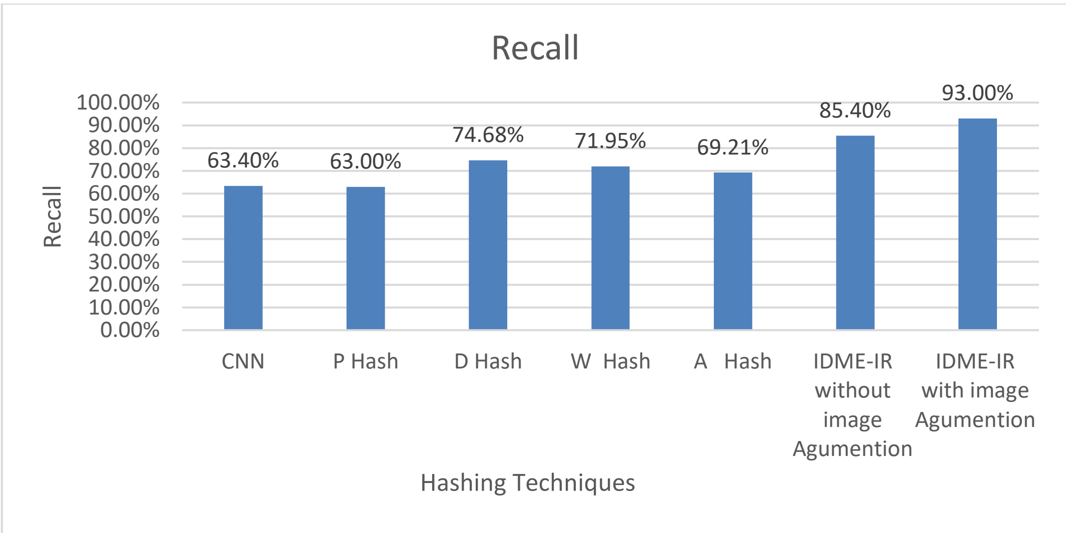

Figure 8: Performance comparison of hashing techniques in terms of Recall

- CNN (Convolutional Neural Network) and P Hash (Perceptual Hashing) both demonstrate recall values of approximately 63.40% and 63.00%, respectively, indicating their similar effectiveness in identifying relevant images.
- D Hash (Difference Hashing) shows improved performance with a recall of 74.68%, while W Hash (Wavelet Hashing) achieves 71.95% recall, slightly lower but still above CNN and P Hash.
- A Hash (Average Hashing) lags behind with a recall of 69.21%.
- The IDME-IR model without image augmentation achieves a significant improvement in recall, reaching 85.40%. This demonstrates the effectiveness of the deduplication model alone.

- The highest performance is observed in the IDME-IR model with image augmentation, where recall reaches 93.00%. This shows the substantial benefit of applying augmentation techniques, such as rotation and scaling, in further enhancing the model's ability to retrieve relevant images.

7. CONCLUSION AND FUTURE WORK

The evaluation of the **IDME-IR Deduplication Model** and image augmentation methods for image retrieval has demonstrated that both techniques contribute significantly to enhancing retrieval performance across various parameters, including **accuracy**, **precision**, **recall**, and **F1-score**. The deduplication model effectively eliminates redundant data, streamlining the retrieval process and improving precision by reducing unnecessary matching of duplicate images. On the other hand, image augmentation enhances the dataset's diversity, which boosts recall and ensures that the system can generalize better to different variations in images.

The combined use of deduplication and augmentation strikes a balance between dataset cleanliness and diversity, leading to improved performance metrics overall. However, careful fine-tuning of augmentation parameters is necessary to ensure that increased dataset size through augmentation does not introduce unnecessary noise or overly similar images that could impact precision negatively.

Ultimately, this study highlights the importance of using both deduplication and augmentation in tandem to optimize image retrieval systems. The improvements in F1-score, which balances precision and recall, demonstrate the potential of these techniques to enhance the overall reliability and robustness of image retrieval models in real-world applications. Future work could focus on refining these techniques further to maximize their impact in specific domains of image retrieval.

REFERENCES

[1] Xu M, Yoon S, Fuentes A, Park DS. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognition. 2023 May 1;137:109347.

[2] Islam T, Hafiz MS, Jim JR, Kabir MM, Mridha MF. A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions. Healthcare Analytics. 2024 May 8:100340.

[3] Alimisis P, Mademlis I, Radoglou-Grammatikis P, Sarigiannidis P, Papadopoulos GT. Advances in Diffusion Models for Image Data Augmentation: A Review of Methods, Models, Evaluation Metrics and Future Research Directions. arXiv preprint arXiv:2407.04103. 2024 Jul 4.

[4] Suto J. Using Data Augmentation to Improve the Generalization Capability of an Object Detector on Remote-Sensed Insect Trap Images. Sensors. 2024 Jul 11;24(14):4502.

[5] Khalifa NE, Loey M, Mirjalili S. A comprehensive survey of recent trends in deep learning for digital images augmentation. Artificial Intelligence Review. 2022 Mar;55(3):2351-77.

[6] Yang S, Xiao W, Zhang M, Guo S, Zhao J, Shen F. Image data augmentation for deep learning: A survey. arXiv preprint arXiv:2204.08610. 2022 Apr 19.

[7] Yang Z, Sinnott RO, Bailey J, Ke Q. A survey of automated data augmentation algorithms for deep learning-based image classification tasks. Knowledge and Information Systems. 2023 Jul;65(7):2805-61.

[8] Yunpeng Gong, Yongjie Hou, Chuangliang Zhang, and Min Jiang. Beyond augmentation: Empowering model robustness under extreme capture environments. arXiv preprint arXiv:2407.13640, 2024.

[9] Xu M, Yoon S, Fuentes A, Park DS. A comprehensive survey of image augmentation techniques for deep learning. Pattern Recognition. 2023 May 1;137:109347.

[10] Wang, G., Li, W., Ourselin, S. and Vercauteren, T., 2019. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4 (pp. 61-72). Springer International Publishing.

[11] Lewy D, Mańdziuk J. An overview of mixing augmentation methods and augmentation strategies. Artificial Intelligence Review. 2023 Mar;56(3):2111-69.

[12] Hrga I, Ivasic-Kos M. Effect of Data Augmentation Methods on Face Image Classification Results. InICPRAM 2022 (pp. 660-667).

[13] Jinwoo K. A Review of Image Retrieval Techniques: Data Augmentation and Adversarial Learning Approaches. arXiv preprint arXiv:2409.01219. 2024 Sep 2.
[14] Chen Y, Yang XH, Wei Z, Heidari AA, Zheng N, Li Z, Chen H, Hu H, Zhou Q, Guan Q. Generative adversarial networks in medical image augmentation: a review. Computers in Biology and Medicine. 2022 May 1;144:105382.
[15] Ismael SA, Mohammed A, Hefny H. An enhanced deep learning approach for brain cancer MRI images classification using residual networks. Artificial intelligence in medicine. 2020 Jan 1;102:101779.
[16] Gour M, Jain S, Sunil Kumar T. Residual learning based CNN for breast cancer histopathological image classification. International Journal of Imaging Systems and Technology. 2020 Sep;30(3):621-35.
[17] Makhlouf A, Maayah M, Abughanam N, Catal C. The use of generative adversarial networks in medical image augmentation. Neural Computing and Applications. 2023 Dec;35(34):24055-68.
[18] Vij, P., Dalip (2023). Comparative Analysis of Image Augmentation and Data Deduplication Techniques. In: Senjyu, T., So–In, C., Joshi, A. (eds) Smart Trends in Computing and Communications. SMART 2023. Lecture Notes in Networks and Systems,vol645. Springer, Singapore.https://doi.org/10.1007/978- 981-99-0769-4_26.
[19] Kaur R, Bhattacharya J, Chana I. Deep CNN based online image deduplication technique for cloud storage system. Multimedia Tools and Applications. 2022 Nov;81(28):40793-826.
[20] Vij P, Dalip D. Performance Analysis of Dimensionality Reduction Techniques for Datasets using Deep Learning. In2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) 2023 Jul 6 (pp. 1-6). IEEE.
[21] Chakiat A, Oli N, Modi VK. Deduplication of advertisement assets using deep learning ensembles. In2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA) 2020 Dec 14 (pp. 889-894). IEEE.
[22] Akbar M, Ahmad I, Mirza M, Ali M, Barmavatu P. Enhanced authentication for de-duplication of big data on cloud storage system using machine learning approach. Cluster Computing. 2024 Jun;27(3):3683-702.
[23] Usharani S, Dhanalakshmi K. An image storage duplication detection method using recurrent learning for smart application services. The Journal of Supercomputing. 2023 Jul;79(10):11328-54.
[24] Hema C, Marquez FP. Storage Enhancement in the Cloud Using Machine Learning Technique and Novel Hash Algorithm for Cloud Data Security. InInternational Conference on Management Science and Engineering Management 2022 Jul 14 (pp. 516-526). Cham: Springer International Publishing.
[25] Chen L, Xiang F, Sun Z. Image deduplication based on hashing and clustering in cloud storage. KSII Transactions on Internet and Information Systems (TIIS). 2021;15(4):1448-63.
[26] Kumar RN, Sathiya T, Mohanta HC, Ahmad SS, Kumar A, Alshammri GH, Atiglah HK. Research Article Secure Data Deduplication System with Cyber Security Multikey Management in Cloud Storage.

BIOGRAPHIES OF AUTHORS

Priya Vij holds a Bachelor of Computer Applications (BCA) degree and a Master of Computer Applications (MCA) degree, both from Kurukshetra University. She is currently pursuing a Ph.D. at the M.M. Institute of Computer Technology & Business Management, Maharishi Markandeshwar (Deemed to be University), Mullana. With a strong academic foundation, Priya Vij has contributed to the field of computer science through publications in various conferences, showcasing her dedication to advancing knowledge in her area of expertise. Her primary research interests lie in the domain of Deep Learning, a field at the forefront of artificial intelligence. She can be reached at vijpriya1@gmail.com for academic or professional inquiries.

	**Dr. Dalip** is working as an Associate Professor in M.M. Institute of Computer Technology & Business Management, M.M (Deemed to be University). Prior to this appointment he was a researcher at the same University. He is medalist in Master of Technology. He has awarded 08 copyrights by Govt. of India and published Dataset Mango Leaf Species (Indian) on Kaggle. He received his Master of Technology in Information Technology from M.M University, Master of Computer Application from Kurukshetra University and Ph.D in Information Technology from the M.M (Deemed to be University).