Duplication Detection in Medical Imaging Datasets: A Comparative Study of ISIC Data on Kaggle

VERON HOXHA
veho@itu.dk
STUDENT NUMBER: 21851

IT UNIVERSITY OF COPENHAGEN

RESEARCH PROJECT
KIREPRO1PE
3RD SEMESTER, AUTUMN 2024
DATA SCIENCE MSC

Duplication Detection in Medical Imaging Datasets: A Comparative Study of ISIC Data on Kaggle

Veron Hoxha
IT Univeristy of Copenhagen
veho@itu.dk

***Abstract*—The International Skin Imaging Collaboration (ISIC) datasets are noteworthy and the leading datasets for researchers in machine learning for medical image analysis due to their large content size and the wide use across many papers due to the ability to be used for most of the tasks focused on classification and segmentation, especially in the field of skin cancer detection and malignancy assessment. ISIC datasets images spread on releases throughout 2016 - 2020 with a fresh new release in 2024, where each one of the releases has a specific task to be solved which contributes towards the medical field. The ISIC datasets notably contain numerous duplicates, underscoring the importance of further investigation into this problem. The presence of duplicate datasets is problematic because these often lack images and are prepared for specific machine learning tasks that do not require the entire dataset. Furthermore, there are cases where images in the duplicate datasets have been mixed across sets which lead to data leakage. This is a critical problem in the medical field, where the integrity of data is important to ensuring accurate diagnoses and research outcomes. In this study, we adopt a systematic approach that progresses from simple to more sophisticated image comparison methods to identify how similar ISIC datasets on Kaggle are compared to the official ISIC Challenge datasets and how effective each of these applied methods is. We find that duplicate datasets result in missing images, images being preprocessed, and even images being misplaced across training, testing, and validation sets. The findings from our research set a foundation for future improvements in the processes used to detect and understand ISIC duplicate datasets not only on Kaggle but also on other platforms, aiming to support the trustworthiness of medical imaging resources for research.**

The code necessary to replicate this report is available at Duplicate Datasets - GitHub Repo.

I. Introduction

Medical imaging datasets are crucial in advancing medical research by providing essential insights that improve diagnostic accuracy through machine learning. It is known that even medical experts with several years of experience can occasionally miss specific findings or wrongly discard them for this reason the usage of medical imaging data is crucial in the world of health with a potential chance that they may see a tremendous positive impact in the near future as a result of new research and approaches enabled by AI [1, 2]. Popular platforms like Kaggle and Hugging Face provide a wide range of medical imaging datasets for research and development. However, a common issue is the frequent occurrence of duplicate datasets. These duplicates often come with incomplete titles, descriptions, licenses, and missing, preprocessed, or misplaced images, which can lead even to data leakage, thus compromising research validity and reproducibility.

Before using medical images for research data preparation is the most critical stage [3]. Alabduljabbar et al. [4] highlights the necessary steps when working with medical images, it mentions a few steps such as ethical process, quality control, and data structure playing a crucial role before jumping into potential investigation. Despite these efforts to inform the audience to take those actions, the presence and usage of duplicate datasets continues to rise and pose challenges, indicating the need for more investigation.

In this context, we explore the world of dataset duplicates for a very specific dataset: The International Skin Imaging Collaboration (ISIC) [5, 6]. A widely known dataset with multiple competitions over the years where initial research was conducted on the same topic and dataset by Jiménez-Sánchez et al. [7]. More concretely, we want to know **"How similar are ISIC datasets available on Kaggle compared to the official ISIC Challange datasets?" In addition, we want to investigate how effective are various image comparison techniques in identifying exact and near-exact duplicates within the image datasets.**

In Section II and III we show the process we followed to collect and pre-process the data together with the preliminary analysis/findings. In Section IV we address the main methods used to detect the duplicates. In Section V we showcase the results of our methods across the duplicate ISIC datasets picked for deeper analysis. Finally, in Section VI, we delve into the potential biases present in our dataset and disclose the primary limitations associated with its use which are followed by Section VII which talks about the future work and in the end conclusion in Section VIII.

II. Data

A. Data Collection

Data for this research was primarily sourced from the official ISIC challenge website [6], and from Kaggle, a well-known platform for data science competitions and community interaction. The Kaggle API [8] along with the Croissant [9] schema, a sophisticated dataset format was used to ensure that metadata collection was both methodical and reliable. This approach not only made it easier to gather the metadata but also set a strong foundation for potential future work by establishing a repeatable method for metadata extraction. The extracted information was organized according to the fields outlined in Appendix A, Table IV.

1) ISIC Challenge: Data from the ISIC challenge website [6] provided official benchmarks for comparison, which are crucial for analyzing the ISIC duplicate datasets from Kaggle. ISIC dataset challenges started in 2016 and continued every year until 2020 [10, 11, 12, 13, 14]. A new challenge was introduced in June 2024 [15] after 4 years of break which for the moment does not have all the data available for the public and thus was excluded from this research. It's worth mentioning that the ISIC dataset challenges share images that are used between one year and another year. Research by Cassidy et al. [16] found out that for the challenges between the years of 2016-2020, the total number of binary identical image files across all training and test sets is 13,976 which includes duplicates found both within individual training and test sets and across training and test sets.

2) Kaggle: Kaggle provides diverse options of datasets through its API, particularly those related to the ISIC datasets, which are central to this study. The platform's large dataset repository was crucial, offering a rich source of medical imaging data.

The Kaggle datasets downloaded kept their original names provided by the creators, except that any spaces in the names were replaced with dashes ("-"). In contrast, the datasets from the official ISIC website had to be downloaded separately for training, testing, and validation sets. Consequently, these were added in one directory named based on the challenge year in the format of `"ISIC-(YEAR_OF_THE_CHALLENGE)-CHALLENGE"`, the year of the challenge of ISIC starts from 2016 to 2020. Within these directories, as well as their Kaggle counterparts, the data was categorized into three subfolers (sets): `"train"`, `"test"`, and `"val"`. This organization gave the opportunity for a more detailed analysis, allowing for specific investigations into the presence of missing or duplicate images in these sets, rather than just in the overall dataset.

III. Exploratory Data Analysis

A. Data Processing

We established an initial criterion to refine the selection process for our study, which involves identifying datasets from Kaggle that contain the term "ISIC" in their titles. Further, to narrow down the dataset metadata sample, matches were identified between duplicate datasets on Kaggle and the official ISIC challenge datasets based on the matched corresponding year tags within their titles, resulting in 138 matched datasets.

For each matched pair, the content size difference of the datasets was calculated in megabytes (MB). This difference, termed and added as *"Size Difference (%)"* in the retrieved metadata, shows the difference between the content size of the Kaggle dataset and that of the original ISIC challenge dataset. The *"Size Difference (%)"* named as *percentage_difference* in the code when calculated, is found using the formula:

$$percentage\_difference = \frac{|kaggle\_content\_size - original\_size|}{original\_size \times 100}$$

Here, *kaggle_content_size* denotes the content size of the dataset on Kaggle in MB, and *original_size* refers to the content size of the ISIC challenge dataset in MB. This metric results stored in the *"Size Difference (%)"* column, fundamentally represent the relative content size variation between the matched duplicate and the original ISIC dataset for a given year.

Figure 1 shows the distribution of the 138 examined potential duplicate datasets in terms of *"Size Difference (%)"*. This histogram provides a visual representation of how these duplicates diverge in data volume relative to their original counterparts. Notably, the majority of duplicates reveal a *"Size Difference (%)"* greater than 50%, with the 60-100% range being particularly predominant. This significant difference raises concerns about the completeness of the duplicate datasets, as it implies substantial content is missing compared to the originals. Such insights are crucial for evaluating the integrity and variability of the datasets being analyzed.

Due to limitations as mentioned in Subsection VI-C three matched datasets listed in Table I were selected by random for an in-depth analysis of duplications, more details about those datasets can also be seen in Table II. These datasets represent three consecutive years of ISIC challenges and are categorized into three distinct size difference brackets (1-30%, 30-60%, and 60-100%). This categorization facilitates a significant comparison across different editions of the ISIC challenges and allows for a detailed examination of duplicates that vary significantly in content size. Each dataset picked, including the official ones and the selected Kaggle duplicates, was manually downloaded to ensure accuracy in our comparative analysis, however, it is worth mentioning that the Kaggle dataset in *"Data 2019"* was missing the *"test"* set, also the *"val"* set does not exist in the original and duplicate dataset at all as can be seen in Table II.

B. Preliminary Analysis/Findings

The initial phase of our research involved exploring Kaggle datasets that included the term *"ISIC"* in their titles, as discussed in Subsection III-A. As of October 2024, this investigation identified 860 datasets that collectively consume a significant amount of storage. Notably, these datasets occupy approximately 2,642.14 GB, which clearly contrasts with the 75.15 GB used by the official ISIC datasets hosted on the ISIC challenge website.

A preliminary analysis was conducted to explore the relationship between content size *(contentSize)*, download count *(downloadCount)*, and usability ratings *(usabilityRating)* of potential duplicate datasets on Kaggle. For detailed definitions of these fields, please see Appendix A, Table IV. Usability ratings serve as indicators of dataset quality, reflecting aspects such as completeness, credibility, and compatibility, key factors for users when selecting datasets. As illustrated in Figure 2, the investigation reveals a trend where datasets with higher download counts tend to receive higher usability ratings. This suggests that datasets that are downloaded more frequently are perceived as being of higher quality and more valuable to the

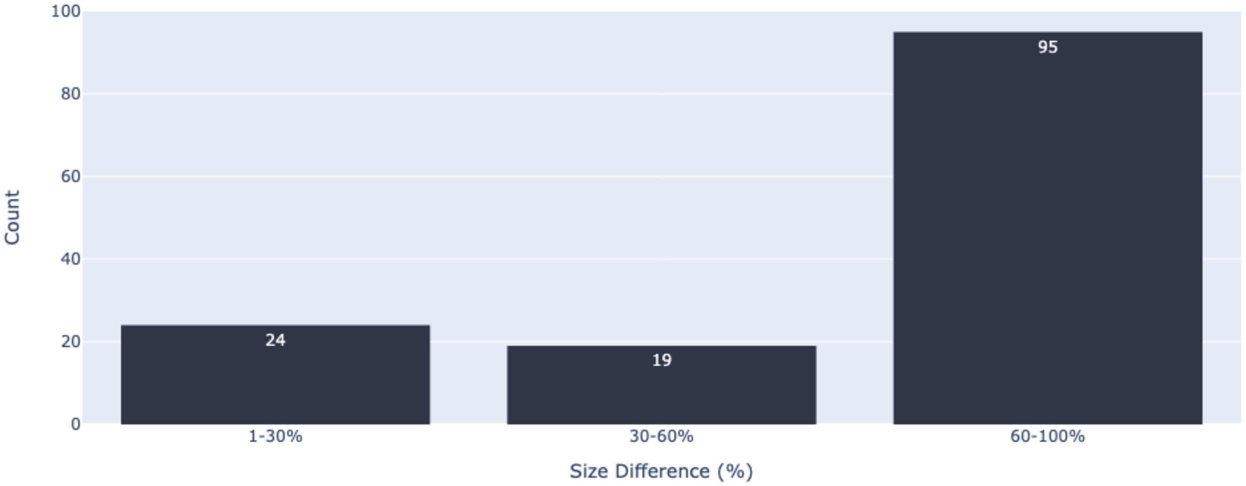


Fig. 1: Distribution of *"Size Difference (%)"* on content size between original and duplicate datasets on Kaggle (matched on year in the title).

Acronym	Kaggle Dataset Name	Kaggle Size (MB)	Original Dataset Name	Original Size (MB)	Size Diff. (%)	Size Diff. Cat. (%)
Data 2017	ISIC-2017-1-FOLD	6144.0	ISIC-2017-Challenge	12358.0	50.28	30-60%
Data 2018	ISIC-2018	13312.0	ISIC-2018-Challenge	16261.0	18.14	1-30%
Data 2019	ISIC-2019-Preprocessed-Dataset	3072.0	ISIC-2019-Challenge	13154.0	76.65	60-100%

TABLE I: Characteristics of selected datasets from Kaggle for detailed comparison, categorized by *"Size Difference (%)"* and aligned with their corresponding matched official ISIC challenge dataset for comparison.

community, which must not always be the case thus a careful detailed check of the dataset needs to be done before being picked for usage.

IV. Methods

Image comparison methodologies vary widely in their complexity and effectiveness. In our study, we used five distinct image comparison techniques to evaluate their efficacy in identifying duplicates within ISIC datasets. Each technique contributes unique insights into the characteristics of the datasets, ultimately combining the insights to provide a comprehensive understanding of the similarity between selected duplicate datasets from Kaggle and the official ISIC challenge datasets. A methodology flowchart illustrating how this approach was applied is shown in Figure 3. The flowchart outlines the process described in Section II-A, where the datasets are divided into three sets: *"train"*, *"test"*, and *"val"* with the *"all"* set representing the entire dataset. Experiments are conducted by comparing each corresponding set (*"train"*, *"test"*, *"val"*) of one dataset with the matching sets of another dataset to assess their similarities. We sequentially apply four methods: first, image filename comparison (Method IV-A), second, image dimension comparison (Method IV-B), third, image size comparison (Method IV-C), and fourth, pixel-by-pixel comparison (Method IV-D). The final method, the hashing technique (Method IV-E), is applied to the entire dataset, referred to as the *"all"* set.

A. Method 1 - Image filename comparison

This method involves basic checks, such as comparing image filenames across the datasets. It serves as a preliminary filter to identify potential duplicates or differences. Images found during this method will be called as *"Common Images"* across this research. A primary challenge of this technique is that different naming conventions can cause identical images to have different filenames, especially when dataset creators use different naming for unknown reasons. This issue will be mitigated by usage of another technique that does not rely on filenames, thereby overcoming this limitation.

B. Method 2 - Image dimension comparison

This method builds upon Method IV-A, where images identified as potential duplicates based on matching filenames (*"Common Images"*) are further filtered for detailed analysis. Specifically, we check the dimensions of these images to make sure they can be compared using the Pixel-by-Pixel Comparison (Method IV-D), which only works with images that have the same dimension.

Acronym	Sets	Kaggle (Duplicate) Image Count	ISIC (Original) Image Count
Data 2017	Train	821	4000
	Test	279	1200
	Val	274	300
	Total	**1374**	**5500**
Data 2018	Train	2594	12609
	Test	1000	2512
	Val	100	293
	Total	**3694**	**15414**
Data 2019	Train	25331	25331
	Test	N/A	8238
	Val	N/A	N/A
	Total	**25331**	**33569**

TABLE II: Image counts across sets for selected Kaggle datasets and their official ISIC challenge counterparts.

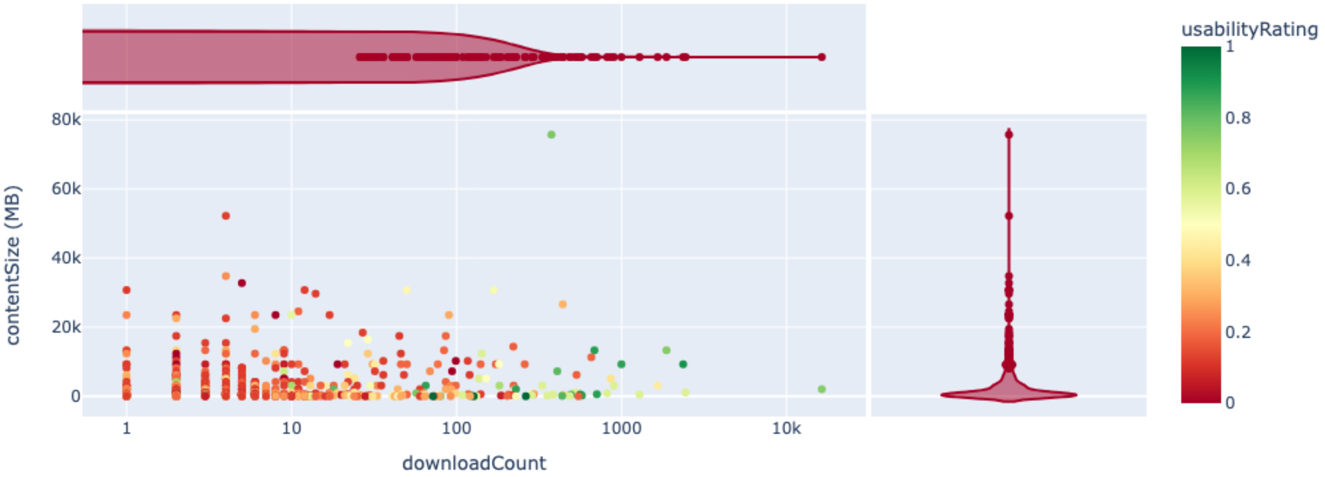


Fig. 2: Relationship between content size, download count and usability rating of the potential ISIC duplicates on Kaggle.

C. Method 3 - Image size comparison

Identical to Method IV-B the image size comparison extends on Method IV-A where we check the sizes of the *"Common Images"* to see if there are any differences which would indicate potential modifications done. These potential modifications could include compression, resolution changes, cropping, color depth adjustment, etc. Each of these modifications can pose problems by altering the image's quality, content, context, or authenticity, which can be significant depending on their intended use in contexts such as legal or academic publications.

D. Method 4 - Pixel-by-pixel comparison

Building on the baseline comparisons described in Method IV-A and IV-B, this technique delves deeper by analyzing images at the pixel level. It is designed to detect exact matches and find differences that are not visible through basic comparisons alone. By focusing on the *"Common Images"* identified by the filename comparison that share identical dimensions found during dimensions checks, this method explores whether images that appear similar based on filenames and dimensions are truly identical when going through a detailed pixel-by-pixel comparison. Since the bit depth of the images used is 8 bits, allowing for 256 distinct shades per color, we have enough detail to spot differences effectively. However, this bit depth might not catch the tiniest gradations that higher bit depths could reveal. Despite this limitation, the 8-bit depth is quite adequate for uncovering significant differences, offering a clearer view of how similar the datasets really are.

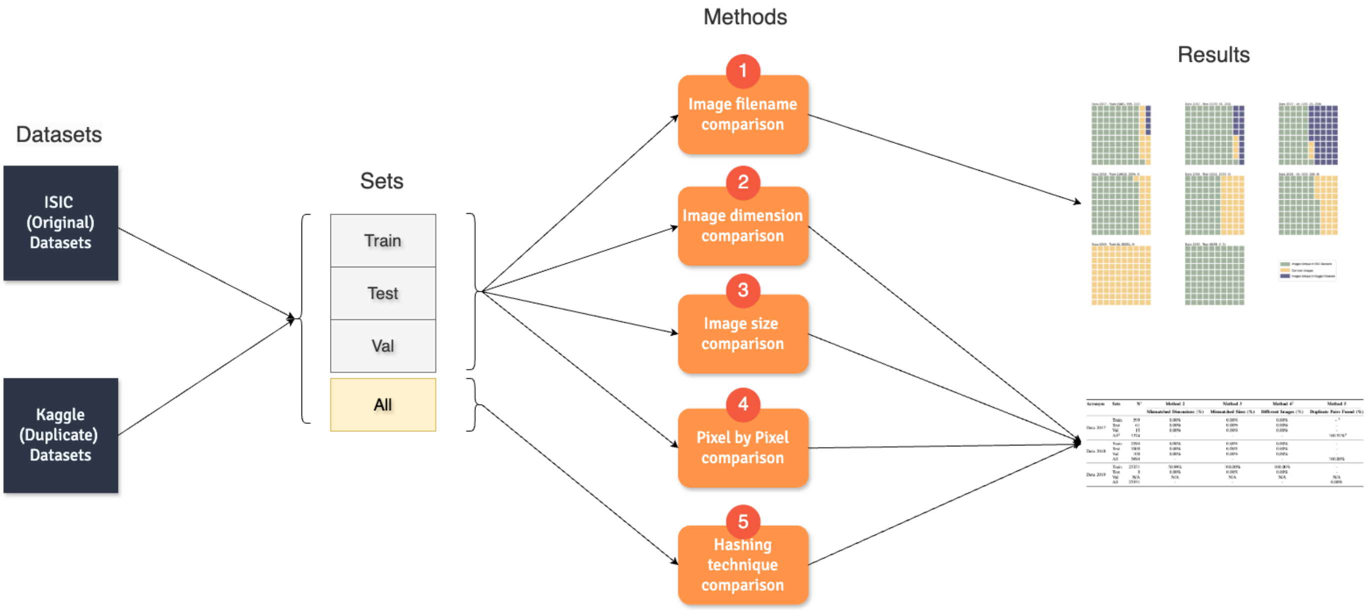

Methods
Results
Datasets
ISIC (Original) Datasets
Kaggle (Duplicate) Datasets
Sets
Train
Test
Val
All
1
Image filename comparison
2
Image dimension comparison
3
Image size comparison
4
Pixel by Pixel comparison
5
Hashing technique comparison

Fig. 3: Flowchart of the process of how the methods were run, with the numbers in circles indicating the order.

E. *Method 5 - Hashing Technique comparison*

The hashing technique uses a cryptographic function to identify exact duplicates across datasets by calculating MD5 hashes for each image. This method detects identical images regardless of their filenames or sets because it is applied to the entire dataset, as illustrated in the flowchart in Figure 3. It's important to note that applying the method to the entire dataset is an intentional choice, not an inherent feature of the method itself. For each image, the hash is recorded along with its dataset name, subfolder (*"train"*, *"test"*, *"val"*), file path, and filename, enabling a detailed comparison of content across all datasets.

Moreover, if we apply the hashing technique in the same way as the pixel-by-pixel comparison to the same sets, we would expect to see the same results from both methods since they are both effective at finding duplicates. This means there isn't a clear advantage in choosing one method over the other based on performance alone. However, the decision might come down to factors like computational efficiency or how resource-intensive one method is compared to the other.

This approach completed our suite of methods each one of which tends to prove and find insights that help us understand the similarity of the duplicate datasets.

V. Results

Results of the methods used in this study are presented in Figure 5 for Method 1 the image filename comparison method and in Table III for the other four Methods IV-B, IV-C, IV-D, and IV-E. The following subsections titles provide conclusions based on the results of each method.

A. *Common and Missing Images*

Figure 5 shows the results of the image filename comparison, where we attempt to identify images that are unique to the original ISIC challenge datasets and their corresponding duplicate datasets on Kaggle. The figure also includes the count of *"Common Images"* across the *"train"*, *"test"*, and *"val"* sets. Our results reveal that, based on filename comparison across all three sets, a somehow low number of *"Common Images"* were identified. The "Data 2017" pair datasets showed a low number of *"Common Images"* in all three sets, with an average of only 7% of all images being common based on filenames. This raises a concern that either the images are named differently or misplaced across the sets, which this will be discussed in Subsection V-C.

In contrast, the "Data 2018" dataset reveals that all images in the duplicate dataset are also available in the original dataset. For the "Data 2019" dataset, the results follow the same approach, with the exception that the "test" set does not exist for the duplicate dataset.

Although this method does not provide an exact measure of how similar the duplicate datasets are, it directly shows the number of *"Common Images"* based on matched filenames. Additionally, as indirectly indicated by the numbers in brackets in the title of Figure 5, the duplicate datasets are missing a significant number of images compared to the original dataset. These findings from this experiment reveal the first potential differences among the duplicate datasets.

B. *Mismatched dimensions and size (preprocessed images)*

Table III presents the results of several methods, focusing specifically on Methods IV-B, IV-C, and IV-D, which involve image dimension comparison, size comparison, and pixel-by-pixel comparison, respectively. This table shows the percentage of mismatched dimensions and sizes among *"Common Images"* across all datasets and sets.

For the “Data 2017” and “Data 2018” datasets, there were no mismatches in either dimensions or sizes among the *“Common Images”* with a result 0.00%. This indicates that the *“Common Images”* in these datasets are identical in both their spatial properties and file sizes. However, in “Data 2019”, out of 25,331 *“Common Images”* listed in Table III, 50.99% (12,917 images) had different dimensions despite having the same filenames. Additionally, all 25,331 images (100.00%) in “Data 2019” differed in size. These differences suggest that the images in the duplicate dataset may have been resized, cropped, compressed, or otherwise edited before being uploaded to Kaggle.

Furthermore, Table III includes also the results of the pixel-by-pixel comparison (Method 4) for the *“Common Images”* that had matching dimensions identified by the image dimension comparison. The results confirm that all *“Common Images”* with matching dimensions in “Data 2017” and “Data 2018” datasets are exact duplicates at the pixel level as well. In contrast, for the “Data 2019” “train” set, although 49.01% (12,414 images) of the *“Common Images”* had matching dimensions, none of these images were identical when compared pixel-by-pixel. This indicates that the images in the duplicate dataset were likely preprocessed with such adjustments as brightness, contrast, or saturation by the dataset’s creator. An example comparing an original dataset image with its duplicate is shown in Figure 4, where the duplicate image clearly shows edits that were done.

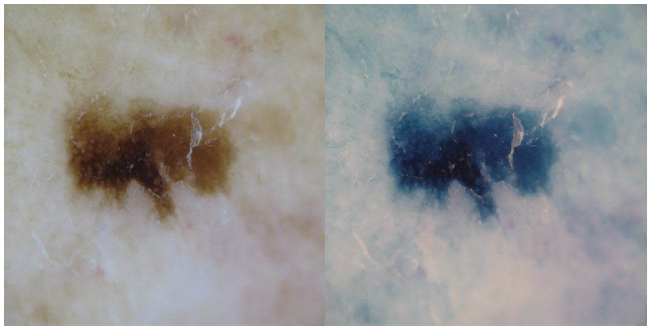
The results from these three methods show that duplicate datasets can potentially include images with the same filenames but different dimensions, sizes, and preprocessing. These differences confirm that the images are not identical, which can have a significant impact on machine learning applications in medical imaging. This variability can lead to inconsistent research outcomes, a major concern in the medical field where the accuracy and reliability of diagnostic tools are crucial. Furthermore, although some machine learning models may handle these differences effectively, others may not, so, this stresses the importance of addressing this issue to increase the carefulness around this topic.

C. Misplaced Images

Lastly, Table III presents the results for the final method applied, the hashing technique comparison which is conducted independently of filenames and sets. This method served as a definitive tool for identifying exact duplicates across the datasets. The analysis was performed on the entire dataset for a reason as mentioned in Subsection V-A, where the first method identified a very small number of *“Common Images”* raising concerns that images might be misplaced or named differently.

The analysis reveals the following findings:

- **“Data 2017”**: A total of 1,381 duplicates were identified, representing 100.51% of the 1,374 images in the duplicate folder. This indicates that 5 images appear more than once.
- **“Data 2018”**: The results confirm that all 3,694 images in the duplicate folder are exact duplicates of those in the original dataset.
- **“Data 2019”**: No exact duplicates were found. This absence is attributed to preprocessing modifications identified by the previous methods.



(a) Original dataset image (b) Duplicate dataset image

Fig. 4: One of many different images (ISIC_0071212) found by pixel-by-pixel comparison in “Data 2019” datasets which have the same filename but different edits applied to the duplicate image.

The number of images compared using the hashing technique referred as “N” in Table III is determined by the smaller dataset either the original or the duplicate whichever has fewer images. In this case, all numbers correspond to the total number of images in the duplicate datasets, which contain the least images due to many missing as can be seen in Table II. Therefore, the purpose was to determine if these already contained images are still at least part of the original dataset.

The results of this experiment confirmed the concern raised by the image filename comparison that images might be misplaced across the sets. This is evidenced by the *“Data 2017”* results, where the filename comparison identified only 7% of the *“Common Images”* as shown in Figure 5. However, the hashing technique revealed 1,381 duplicates, including 5 additional images beyond the total, indicating that the images in the duplicate datasets may have been mixed across sets, leading to data leakage.

VI. Discussion

In this project, our primary focus has been to identify how similar the ISIC datasets available on Kaggle are compared to the official ISIC challenge datasets. Additionally, we aimed to demonstrate the effectiveness of various image comparison techniques in identifying exact and near-exact duplicates within the sample we chose.

We presented a comprehensive methodology flowchart detailing how these methods were applied and how each one was used to identify image duplicates.

By integrating image comparison techniques, our findings reveal that duplicate datasets often contain missing, misplaced, and preprocessed images. This gives us a better insight into the nature of these duplicates. It’s especially important as it

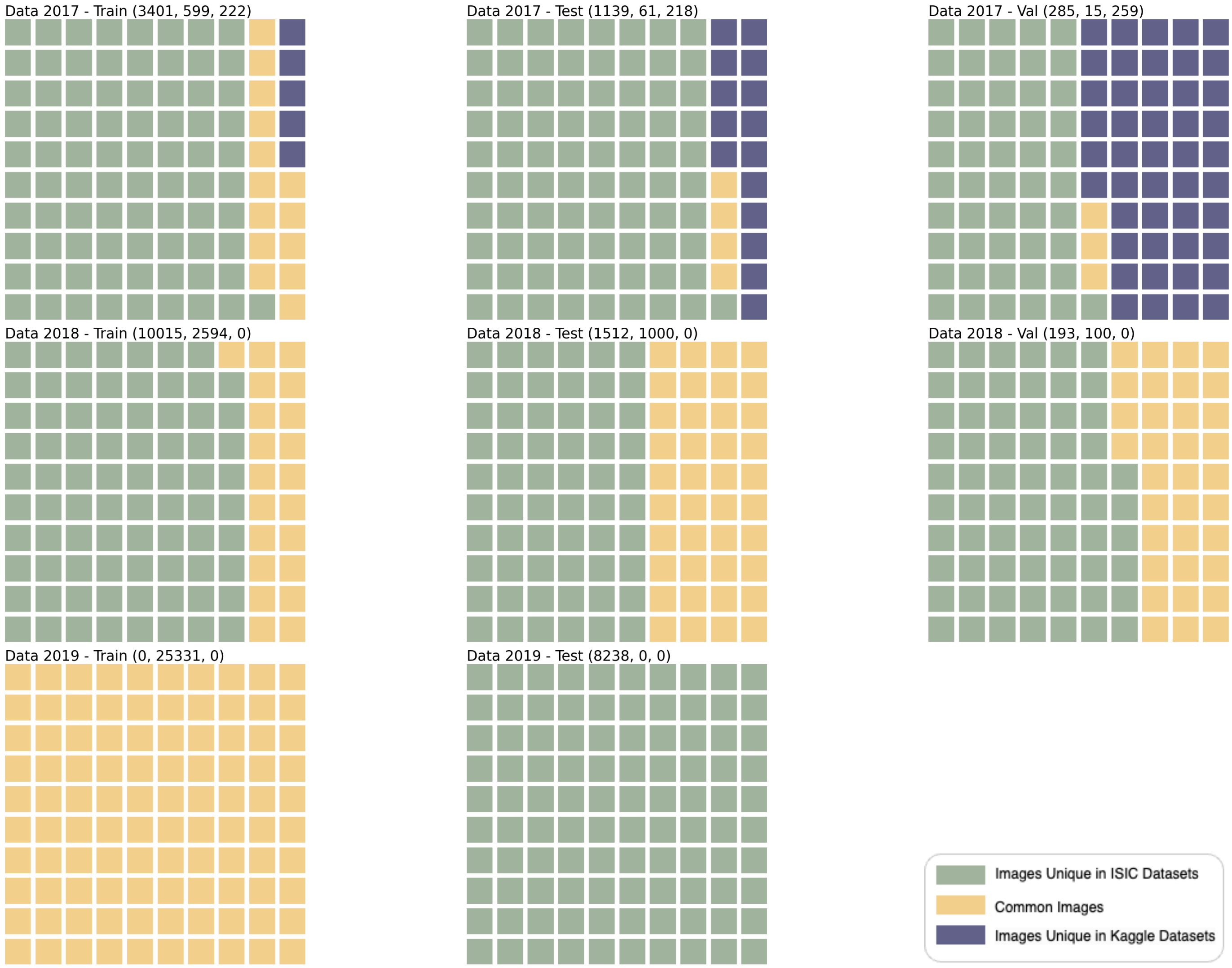


Fig. 5: 10x10 Waffle charts for the number of unique images in the original ISIC challenge datasets and the corresponding duplicate datasets on Kaggle, along with the count of common images across the *"train"*, *"test"*, and *"val"* sets, identified by image filename comparison. The numbers in brackets in the title signify the total number of unique images in original ISIC challenge datasets, the number of images common, and the total number of unique images in duplicate datasets on Kaggle, respectively. *"Val"* set does not exist in "Data 2019".

reminds users to be cautious when choosing datasets, ensuring they select ones with high data quality. Since data quality depends on how the data is expected to be used, the most important data quality dimensions depend on the context of use and industry needs, however, some mainly used dimensions are accuracy, completeness, and consistency of the data [17]. Moreover, poor data quality can lead to poor patient care which introduces a significant risk of misdiagnoses, leading to delayed or inappropriate treatment plans [18].

The comparative analysis of the methods we used within our framework offers researchers valuable insights into the performance of available techniques for detecting duplicates. This enables a deeper understanding of their effectiveness and limitations.

However, it is essential to acknowledge certain limitations associated with the use of all this, which will be elaborated upon in the next subsections.

A. Limitations in Keyword Based Search

Our method of identifying potential duplicate datasets was constrained by the use of "ISIC" word as a primary keyword in dataset titles during the initial metadata retrieval phase. This keyword-based approach, based only on the title of the dataset, may overlook datasets that do not specifically mention "ISIC" in their titles but are nonetheless relevant and possibly duplicate versions of the official datasets. This limitation suggests that our study might not capture the full landscape of dataset duplication, highlighting that a potential search of the word "ISIC" being mentioned anywhere in the

Acronym	Sets	N[1]	Method 2	Method 3	Method 4[2]	Method 5
			Mismatched Dimensions (%)	**Mismatched Sizes (%)**	**Different Images (%)**	**Duplicate Pairs Found (%)**
Data 2017	Train	599	0.00%	0.00%	0.00%	-[5]
	Test	61	0.00%	0.00%	0.00%	-
	Val	15	0.00%	0.00%	0.00%	-
	All[3]	1374	-	-	-	100.51%[4]
Data 2018	Train	2594	0.00%	0.00%	0.00%	-
	Test	1000	0.00%	0.00%	0.00%	-
	Val	100	0.00%	0.00%	0.00%	-
	All	3694	-	-	-	100.00%
Data 2019	Train	25331	50.99%	100.00%	100.00%	-
	Test	0	0.00%	0.00%	0.00%	-
	Val	N/A[6]	N/A	N/A	N/A	N/A
	All	25331	-	-	-	0.00%

[1] N represents the number of common images used for comparisons in Methods 2, 3, and 4 found during Method 1 in Figure 5, represented by the middle number in the brackets in the title of each subplot.

[2] For Method 4, N varies based on the percentage result from "Mismatched Dimensions". If the mismatch is 0.00%, N is the full value. Otherwise, **N = N - Mismatched Dimensions (%)**, which gives the number of images with matching dimensions used for Method 4. In this case, N for Method 4 is 12,414.

[3] "All" represents the whole dataset, which includes all images of the dataset with the lowest number of total images out of both datasets being compared, as shown in Table II.

[4] The value exceeds the maximum of 100.00% because the method found multiple duplicates of the same image.

[5] "-" indicates that the experiments were not conducted for that set.

[6] "N/A" stands for "Not Applicable," meaning that the set does not exist.

TABLE III: Summary of Results by Method

dataset on Kaggle might be an option to be investigated in the future.

B. Assumptions in Dataset Matching

Our comparative analysis relies heavily on the assumption that datasets matched on the titles as a potential duplicate contain same year tags, potentially accurately reflecting their content relation. This approach presupposes that all matched datasets explicitly including same year tags in their titles are potential duplicates, which may not always be the case. Some datasets potentially representing exact or near-exact duplicates might not have the year mentioned in the title, thereby escaping our inspection. This assumption could lead to underestimating the extent of duplicates or misrepresenting their characteristics.

C. Sampling Bias

The initial phase of our research was primarily focused on defining the scope of the dataset analysis. A sampling bias was introduced by selecting only three random matched datasets for detailed examination. This selection does not comprehensively represent the entire range of potential duplicates, and thus, the findings from this study should be interpreted with caution. The decision to limit the scope to three matched datasets was driven by time constraints, which may influence the generalizability of our results. However, our algorithmic code is designed to be easily applied to other datasets, allowing for broader investigations if the study is expanded in the future. This limitation highlights the need for careful consideration when assuming these findings to the broader spectrum of ISIC duplicates on Kaggle and their similarity to the official ISIC challenge datasets.

D. Experimental Setup and Method Application

A notable limitation in our methodology involves the application of comparison techniques. For most methods, we conducted comparisons within the same sets (*"train"* to *"train"*, *"test"* to *"test"*, and *"val"* to *"val"*) to mimic practical use. However, for hashing, we analyzed the entire dataset, and this for the only reason to identify if there are cases when images are misplacement across different sets, crucial for detecting potential *"train"*/*"test"*/*"val"* data leakage.

This difference makes it challenging to directly compare the efficacy of each technique head-to-head without acknowledging this limitation.

VII. Future Work

a) Expanding Dataset Analysis: The scope was limited to analyzing a subset of potentially duplicated ISIC datasets on Kaggle as mentioned in Subsection VI-C. For future research, we aim to extend this analysis to include a comprehensive examination of all ISIC related datasets on Kaggle. This expanded study will allow us to uncover patterns and differences across the datasets, increasing our understanding of how these duplicates compare to the original ISIC datasets.

b) Using Advanced Computational Resources: Future investigations could include advanced computational resources, such as high-performance computing clusters, to handle the increased scale of data analysis by comparing all the potential duplicates. It's noteworthy that the current methods can complete a study on our sample datasets in approximately 1.5 hours on a standard machine (if the results are not saved). Therefore, conducting a comprehensive investigation of all duplicates would be impractical on basic computing infrastructure and using advanced computational resources would make this easier and enable a more detailed comparison.

c) Incorporating Additional Image Comparison Techniques: Incorporating more image comparison techniques could benefit the research by enabling a more comprehensive evaluation of the effectiveness of specific methods in detecting duplicates. One such technique could potentially be the computer vision approach for identifying near-duplicate images [19]. This works by extracting relevant features from images, either pixel-based, area-based, or point-based features to create a representation of each image. These features are then compared against other image features to identify perceptually similar images, despite potential alterations like rotation, cropping, scaling, etc. Although this technique may not be the most suitable for detecting exact duplicates, it offers a good opportunity for a possible comparison. It would be interesting to compare near-duplicate datasets identified by this method with exact duplicates detected by another method using a machine-learning model. This comparison could determine if the outcomes are the same, and see how risky those alternations in duplicates are.

d) Addressing Ethical Considerations: Exploring the ethical aspects of duplicate datasets would be beneficial, especially regarding data licenses and documentation. Machine Learning (ML) and Natural Language Processing (NLP) conferences have already started requiring ethics statements and various checklists for submissions [20, 21, 22, 7]. These checklists typically include questions about data licenses and documentation. To examine these aspects, we can analyze the license types of datasets on platforms like Kaggle. A lack of a license often indicates a dataset may not be legitimate. Moreover, each dataset is expected to have documentation, typically found in the "Data" section on platforms such as Kaggle, where the characteristics of the dataset are explained. This process of examination encourages not only data sharing but also taking care of the license and documentation of any new data shared or used in research papers, promoting the use of officially documented versions.

VIII. Conculsion

This study provided an insightful analysis of the similarities between the duplicates of ISIC datasets available on Kaggle and those from the official ISIC challenge using various techniques to identify both exact and near-exact duplicates despite the constraints imposed by the sample size.

Our findings indicate that duplicate Kaggle datasets often miss some images present in the official datasets. This is potentially attributed to creators tailoring the Kaggle datasets for specific ISIC challenge tasks, which do not require the full set of images.

Furthermore, our pixel-by-pixel comparison showed that many duplicates on Kaggle had been modified, for example through adjusting the color histogram. These alterations are likely intended to increase the performance of machine learning models on specific tasks within the ISIC challenges. Moreover, the hashing technique was crucial for identifying the misplacement of images across different sets. This issue was not detected by simpler methods such as filename, dimension, and size comparison, which were performed on a set-to-set basis, however, this is not a limitation of those methods, the decision to be performed on a set-to-set basis was an experimental choice. The hashing comparison, conducted across the entire dataset, found these misplacements, suggesting that such cases of misplacements of images might also be prevalent in other Kaggle datasets.

Among all the methods tested, hashing comparison proved to be the most effective in identifying duplicates, outperforming the other methods. The accuracy of the results from the hashing comparison was confirmed by the metric used to count the number of duplicate images across the datasets.

In summary, this project contributed as a framework for analyzing the substantial degree of similarity between the Kaggle and official ISIC datasets and how effective different methods can be used in finding those duplicates. While acknowledging the documented limitations, we suggest that the findings might have broader implications for other duplicate datasets on Kaggle. However, further research is needed to confirm whether these results can be generalized to additional contexts. Moreover, this work provides a foundation for future improvements and a more in-depth understanding of duplicates of ISIC datasets not just in Kaggle but also other platforms where datasets can be hosted.

References

[1] J. Li, P. Jiang, Q. An, G.-G. Wang, and H.-F. Kong, "Medical image identification methods: a review," *Computers in Biology and Medicine*, p. 107777, 2023.

[2] L. M. Prevedello, S. S. Halabi, G. Shih, C. C. Wu, M. D. Kohli, F. H. Chokshi, B. J. Erickson, J. Kalpathy-Cramer, K. P. Andriole, and A. E. Flanders, "Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions," *Radiology: Artificial Intelligence*, vol. 1, no. 1, p. e180031, 2019.

[3] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[4] A. Alabduljabbar, S. U. Khan, A. Alsuhaibani, F. Almarshad, and Y. N. Altherwy, "Medical imaging datasets, preparation, and availability for artificial intelligence in medical imaging," *Journal of Alzheimer's Disease Reports*, vol. 8, no. 1, pp. 974–986, 2024.

[5] ISIC Archive. [Online]. Available: https://www.isic-archive.com/

[6] ISIC Challenge. [Online]. Available: https://challenge.isic-archive.com/data/

[7] A. Jiménez-Sánchez, N.-R. Avlona, D. Juodelyte, T. Sourget, C. Vang-Larsen, A. Rogers, H. D. Zając, and V. Cheplygina, "Copycats: the many lives of a publicly available medical imaging dataset," 2024. [Online]. Available: https://arxiv.org/abs/2402.06353

[8] Kaggle datasets documentation. [Online]. Available: https://www.kaggle.com/docs/datasets

[9] M. Akhtar, O. Benjelloun, C. Conforti, P. Gijsbers, J. Giner-Miguelez, N. Jain, M. Kuchnik, Q. Lhoest, P. Marcenac, M. Maskey *et al.*, "Croissant: A metadata format for ml-ready datasets," in *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning*, 2024, pp. 1–6.

[10] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1605.01397*, 2016.

[11] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[12] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[13] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[14] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[15] ISIC 2024 Challenge. [Online]. Available: https://challenge2024.isic-archive.com/

[16] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the isic image datasets: Usage, benchmarks and recommendations," *Medical image analysis*, vol. 75, p. 102305, 2022.

[17] S. Juddoo, C. George, P. Duquenoy, and D. Windridge, "Data governance in the health industry: Investigating data quality dimensions within a big data context," *Applied System Innovation*, vol. 1, no. 4, p. 43, 2018.

[18] B. Ehsani-Moghaddam, K. Martin, and J. A. Queenan, "Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data," *Health Information Management Journal*, vol. 50, no. 1-2, pp. 88–92, 2021.

[19] K. Thyagharajan and G. Kalaiarasi, "A review on near-duplicate detection of images using computer vision techniques," *Archives of Computational Methods in Engineering*, vol. 28, no. 3, pp. 897–916, 2021.

[20] A. Rogers, T. Baldwin, and K. Leins, "Just what do you think you're doing, dave?'a checklist for responsible data use in nlp," *arXiv preprint arXiv:2109.06598*, 2021.

[21] A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, "Introducing the neurips 2021 paper checklist," https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500b, 2021.

[22] M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, "Responsible nlp research checklist," https://aclrollingreview.org/responsibleNLPresearch/, 2021.

APPENDIX

APPENDIX A
TABLES

Field	Description
@context.@language	Language in which the dataset was created.
@context.@vocab	URL to schema.org which Croissant vocabulary is an extension to.
@type	Type of retrieved data, in our case all "Datasets".
name	Official title or name of the dataset retrieved as given by the creator.
alternateName	Subtitle or alternate name of the dataset.
description	Description of the dataset.
url	URL to the dataset page on Kaggle.
identifier	ID of the dataset.
creator.@type	The type who created the dataset in our case all "Person".
creator.name	The name of the dataset creator.
creator.url	URL of the creator account.
license.@type	Type of the license used.
license.name	Name of the license used.
keywords	Main keywords which represent the dataset.
dateModified	Last date when the dataset was modified.
isAccessibleForFree	"True" if the dataset can be accessed for free and "False" if it cannot be accessed for free.
distribution_0.@type	Type of the distribution.
contentUrl	Link to download the dataset immediately.
contentSize	The size of dataset in MB.
encodingFormat	Format of the encoding (zip etc.).
isPrivate	"True" if the dataset is private and "False" if the dataset is public.
downloadCount	Integer telling how many times the dataset was downloaded.
viewCount	Integer telling how many times the dataset was viewed.
voteCount	Integer telling how many times the dataset was voted.
usabilityRating	Usability rating of the dataset given in a value between 0 and 1 with 1 being the best.
conformsTo	URL to Croissant format specifications.

TABLE IV: Data fields collected from Kaggle API with Crossiant format.