ANALYSIS OF DATASET OF TEXTILE MATERIALS MACRO IMAGES TO DETECT DATA LEAKS

Merezhko Yelyzaveta
Bachelor's student
Yurchenko Dmytro
Junior Research Fellow, Postgraduate student
Molchanova Maryna
Ph.D in Computer Science, Senior Lecturer
Mazurets Oleksandr
Ph.D in Engineering Science, Associate Professor
Department of Computer Science
Khmelnytskyi National University, Ukraine

The textile industry is increasingly integrated with digital technologies in production, cataloging, quality control, and decision support. In such systems, digital images become one of the principal sources of information about textile materials because they preserve visible properties of fabrics, including texture, surface structure, color distribution, and local visual irregularities [1]. For this reason, image-based analysis is now an important direction in intelligent textile informatics. At the same time, the practical value of any image-based model depends not only on network architecture or training strategy, but also on the actual quality of the dataset on which the model is trained and evaluated [2]. Specific feature of textile macro images is the repetitive organization of the material itself [3]. Fabrics are formed by repeated

interlacing of fibers and yarns, which creates regular or quasi-regular visual patterns. As a result, two fragments of the same fabric roll may appear almost identical even when captured from slightly different positions [4]. In addition, macro images are strongly influenced by illumination, camera angle, scale, and distance to the object [5]. Therefore, the same material may produce multiple visually similar images, while formally remaining different files inside the dataset. This leads to structural redundancy, that is, to the accumulation of duplicate or near-duplicate samples that do not add substantial new information to the training set.

From the standpoint of machine learning, such redundancy is not a harmless technical detail. When a class is overrepresented by many highly similar samples, the model may learn repeated low-level details instead of more robust material-specific patterns [6]. In cross-validation settings, redundancy also increases the risk that perceptually close samples from the same source material appear in different folds, which can produce leakage-like evaluation effects and inflate the apparent predictive quality of the classifier [7]. Consequently, the problem is not limited to dataset compactness [8]; it directly concerns the validity of the experimental conclusions.

For this reason, the analysis of textile image datasets should include not only class balance and quantity of samples, but also internal perceptual similarity structure. Detecting groups of near-duplicate images allows one to identify leakage-prone redundancy, refine the training corpus, and build a more representative sample distribution. In the context of textile macro-image classification, this preprocessing stage is especially important because the visual domain itself contains repeated micro- and macro-textural patterns.

The use of convolutional neural networks and computer vision methods for textile analysis is justified by the nature of the data itself. Textile macro images contain discriminative information distributed across local edges, weave patterns, repeated texture elements, fiber density, and intensity variations [9]. CNNs are particularly suitable for such data because they learn hierarchical visual representations: early layers detect elementary structures, while deeper layers integrate them into more abstract material descriptors [10]. This makes CNN-based approaches more effective than purely manual or handcrafted inspection in tasks where subtle textural differences must be recognized at scale. At the same time, the reliability of the learned representation depends on whether the dataset reflects true variability rather than repeated visual fragments [11].

Computer vision is also relevant because it provides formal tools for measuring perceptual similarity between images. In the uploaded study, the theoretical review distinguishes several families of approaches for image comparison: perceptual hashing, structural similarity metrics, local feature descriptors, and deep vector representations [12]. Among them, deep embeddings are particularly promising for textile images because they capture both local texture cues and higher-level structural information, while remaining more robust than local handcrafted descriptors in the presence of repetitive patterns. Perceptual hashing, in turn, provides a computationally efficient way to detect near-duplicate images at an earlier stage. Together, these methods form an appropriate basis for analyzing leakage-prone redundancy in textile datasets [13].

An additional reason for the relevance of this direction is methodological. Modern research in computer vision increasingly emphasizes data-centric analysis rather than model-centric optimization only. If the dataset contains semantically redundant samples, then even high values of Accuracy or ROC-AUC may not fully reflect true generalization ability. Therefore, the combination of CNN-based classification with computer-vision methods for similarity detection should be regarded not as auxiliary preprocessing, but as an essential component of trustworthy experimental design in textile image analysis.

The aim of this work is to analyze a dataset of textile material macro images in order to detect redundancy-related data leaks and to evaluate how removing perceptually similar images influences the reliability of convolutional neural network classification.

To achieve this aim, the following objectives were defined:

– to analyze the causes of redundancy in textile image datasets;
– to investigate methods for detecting perceptually similar textile images;
– to apply pHash-based and embeddings-based approaches to dataset cleaning;
– to train and compare a CNN model on the original and cleaned datasets;
– to determine the threshold settings that provide the best balance between dataset reduction and classification performance.

These objectives are consistent with the uploaded study, where the methodological focus is placed on improving the quality and trustworthiness of neural-network classification through redundancy detection based on perceptual similarity.

The experimental study used the Natural and Synthetic Fabrics Dataset published on Kaggle. The dataset contains 3,107 images of textile materials divided into two classes: natural fabrics (1,547 images) and synthetic fabrics (1,560 images). The images vary considerably in texture, scale, and lighting conditions. At the same time, the dataset includes visually close examples representing repeated or almost identical fragments of textile surfaces, which makes it suitable for studying leakage-prone redundancy.

The baseline stage consisted of CNN training on the original dataset using 5-fold cross-validation. The baseline model achieved Accuracy 0.9775, F1-score 0.9769, and ROC-AUC 0.9993, with noticeable standard deviation for Accuracy and F1-score, indicating some instability between folds. This instability was interpreted in the uploaded work as a possible consequence of dataset heterogeneity and the presence of redundant or highly similar samples.

At the second stage, the dataset was cleaned using perceptual hashing (pHash). In this approach, two images were considered duplicates when the distance between their pHash codes did not exceed a selected threshold. Smaller thresholds corresponded to stricter similarity criteria, while larger thresholds allowed more variation within the detected duplicate groups. Three threshold values were examined: 6, 8, and 10. The cleaning process removed 135 images at threshold 6 (4.35%), 216 images at threshold 8 (6.95%), and 376 images at threshold 10 (12.10%). After each cleaning scenario, the same CNN model was retrained and re-evaluated.

At the third stage, redundancy detection was performed using deep embeddings. For this purpose, a pretrained ResNet18 model was used to extract vector

representations of the images. The final fully connected layer was removed, and the resulting feature vectors were normalized. Similarity between images was measured by cosine similarity. In this case, a higher threshold meant a stricter definition of duplication. Three threshold values were studied: 0.985, 0.99, and 0.995. The numbers of removed images were 335 (10.78%), 208 (6.69%), and 81 (2.61%), respectively. The model was again retrained and assessed after each cleaning configuration.

Thus, the experimental design combined dataset audit, similarity-based removal of redundant images, repeated CNN training, and comparative evaluation of the resulting classification quality. Such a design makes it possible to estimate not only whether the dataset contains leakage-prone redundancy, but also whether its removal leads to more stable and more credible model behavior.

Results and Discussion. The baseline experiment showed that the classification task was already relatively easy for the neural model, since ROC-AUC was close to 1. However, the gap between the strong average metrics and the non-negligible variability across folds suggested that the dataset structure deserved additional inspection. In studies of visual datasets, such a pattern often indicates that evaluation quality may be affected not only by class separability, but also by internal redundancy and by the presence of visually repeated samples distributed across training and validation subsets. In the context of textile macro images, this interpretation is especially plausible due to the repeated nature of fabric textures.

The pHash-based experiments demonstrated that moderate cleaning improves classification performance. At threshold 6, the model reached Accuracy 0.9933 and F1-score 0.9932. At threshold 8, the best pHash result was obtained: Accuracy increased to 0.9952, F1-score to 0.9954, and standard deviation dropped to ±0.0028. These values are substantially better than the baseline results. At threshold 10, however, performance declined to Accuracy 0.9810 and F1-score 0.9833. This indicates that excessive removal of samples may damage the informative diversity of the dataset. Therefore, pHash is effective when used with a moderate threshold, but overly aggressive filtering may eliminate not only duplicates but also useful textural variations.

The embeddings-based approach showed an even more stable behavior. At threshold 0.985, the model achieved Accuracy 0.9888 and F1-score 0.9898, which already exceeded the baseline. At threshold 0.99, the best overall result was observed: Accuracy 0.9952, F1-score 0.9955, ROC-AUC 0.9998, and standard deviation ±0.0025. At threshold 0.995, the model retained similarly strong performance, with Accuracy 0.9947 and extremely low standard deviation ±0.001. These findings indicate that deep representations extracted from ResNet18 provide a highly informative basis for identifying leakage-prone redundancy in textile macro-image datasets.

A comparison of the two cleaning strategies suggests that both methods are useful, but they serve slightly different purposes. pHash is computationally simpler and captures global perceptual similarity effectively, which makes it suitable for a fast first-pass audit of the dataset. Deep embeddings are more semantically expressive and appear better suited for preserving informative variation while removing truly redundant samples. This is particularly important for textile images, where repetitive

local patterns may cause handcrafted or purely hash-based methods to overgroup images if thresholds are not selected carefully.

From the standpoint of the paper topic, the obtained results support the interpretation that dataset redundancy can act as a source of data leakage risk. In this study, the removal of perceptually similar images led not only to higher average quality but also to lower metric dispersion, which is a sign of more stable and more trustworthy model behavior. Therefore, the detection of visually redundant samples should be considered an integral part of experimental methodology in textile computer vision, especially when small differences in texture are used for material classification.

Conclusions. Thus, the analysis of the textile materials macro-image dataset has shown that visually redundant and near-duplicate samples represent a significant methodological problem for neural-network classification. In the context of this study, such redundancy can be treated as a leakage-prone factor because it reduces effective dataset diversity and may distort the independence of validation results.

The conducted experiments confirmed that preliminary dataset cleaning based on perceptual similarity improves both classification quality and result stability. The baseline model trained on the original dataset achieved Accuracy 0.9775 and F1-score 0.9769. After pHash-based cleaning, the best configuration was threshold 8, which yielded Accuracy 0.9952 and F1-score 0.9954. The embeddings-based approach produced the most robust results at threshold 0.99, with Accuracy 0.9952, F1-score 0.9955, ROC-AUC 0.9998, and reduced standard deviation. These results indicate that the detection and elimination of perceptually similar textile images is an effective way to increase the trustworthiness of CNN evaluation.

The proposed approach can be used in the preparation of textile datasets for computer vision tasks, in the audit of industrial visual repositories, and in the development of intelligent systems for textile material classification. In further research, it is advisable to expand the analysis to multiclass textile datasets, combine perceptual similarity with cluster-based audit strategies, and investigate whether leakage detection before train-test splitting yields additional gains in generalization quality.

References

1. Ingle, N., Jasper, W. J. (2025). A review of the evolution and concepts of deep learning and AI in the textile industry. Textile Research Journal. DOI: https://doi.org/10.1177/00405175241310632

2. Carrilho, R., Yaghoubi, E., Lindo, J., Hambarde, K., Proença, H. (2024). Toward Automated Fabric Defect Detection: A Survey of Recent Computer Vision Approaches. Electronics, 13(18), 3728. DOI: https://doi.org/10.3390/electronics13183728

3. Seçkin, M., Seçkin, A. Ç., Demircioglu, P., Bogrekci, I. (2023). FabricNET: A Microscopic Image Dataset of Woven Fabrics for Predicting Texture and Weaving Parameters through Machine Learning. Sustainability, 15(21), 15197. DOI: https://doi.org/10.3390/su152115197

4. Molchanova, M., Didur, V., Mazurets, O., Sobko, O., & Zakharkevich, O. (2025). Method for construction and demolition waste classification using two-factor neural network image analysis. In CEUR Workshop Proceedings (Vol. 3970, pp. 168–182).

5. Mazurets, O., Molchanova, M., Klimenko, V., & Prosvitliuk, M. (2024). Practice implementation of neural network model BART-Large-CNN for text annotation. In Prospects of Scientific Research in the Conditions of the Modern World. Proceedings of the XXVII International Scientific and Practical Conference (pp. 97–102). Rotterdam, Netherlands.
6. Mazurets, O., Sobko, O., Vit, R., & Pasternak, V. (2024). Practical approach for detection by deep learning of target objects of subject area based on semantic connectivity indicators in audio database. In Proceedings of the XXIV International Scientific and Practical Conference "Modern Scientific Challenges Are the Driving Force of the Development of Scientific Research" (pp. 91–96). Bruges, Belgium: International Scientific Unity.
7. Yurchenko, D. Yu., Ovcharuk, O. M., Mazurets, O. V., & Shevchuk, P. O. (2025). Metod vykorystannia neiromerezhi hybridnoi arkhitektury dlia vyznachennia emotsiinoi tonalnosti tekstovykh povidomen. Mizhnarodnyi naukovo-tekhnichnyi zhurnal "Vymiriuvalna ta obchysliuchalna tekhnika v tekhnolohichnykh protsesakh", 2, 136–141.
8. Molchanova, M. O., Mazurets, O. V., Sobko, O. V., Klimenko, V. I., & Androshchuk, V. I. (2024). Metod neiromerezhevoho vyiavlennia kiberbulinhu z vykorystanniam khmarnykh servisiv ta obiektno-oriientovanoi modeli. Visnyk Khmelnytskoho natsionalnoho universytetu, seriia: Tekhnichni nauky, 2(333), 200–206.
9. Mushtyn, O., Sobko, O., Molchanova, M., & Mazurets, O. (2025, June 9–11). Convolutional neural network architecture for image-based architectural style recognition. In Proceedings of the 4th International Scientific and Practical Conference: Evolving Science: Theories, Discoveries and Practical Outcomes (pp. 130–143). Zurich, Switzerland.
10. Hladun, O., Zalutska, O., Klimenko, V., & Mazurets, O. (2025, May 12–14). Research on the effectiveness of classifying the remains of destroyed buildings using MobileNetV3 neural network architecture. In Proceedings of the 1st International Scientific and Practical Conference: Innovations in Science: From Theoretical Foundations to Practical Impact (pp. 158–162). Antwerp, Belgium.
11. Zalutska, O., Mazurets, O., & Molchanova, M. (2025, November 21). Efficiency analysis of wrecking waste classification using neural network. In Proceedings of the 12th International Conference: Information Technology and Implementation (Satellite) (pp. 142–143). Kyiv, Ukraine.
12. Didur, V., Molchanova, M., & Mazurets, O. (2025, May 26). Research on the effectiveness of neural network detection of plots with the destroyed buildings remains. In Proceedings of the XXI International Scientific and Practical Conference: Modern technologies and science: Problems, new and relevant developments (pp. 245–251). Zaragoza, Spain.
13. Didur, V. O., Molchanova, M. O., Tyschenko, O. O., & Mazurets, O. V. (2025, May 16). Approach for comparative analysis of effectiveness of using MobileNetV3 and ViT neural network models for graphical localization of destroyed buildings remains areas. In Proceedings of the IX International Scientific and Practical Conference: Formation of Innovative Potential of World Science (pp. 94–97). Waterford, Ireland.