Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Mixing Synthetic Data for Small-Data Document Image Classification

Jonathan Kästner

kaestner.jonathan@protonmail.com

University Leipzig

Martin Rupp

FOM Karlsruhe

Richard Beetz

Provadis Hochschule

Research Article

Keywords:

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11095865/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Mixing Synthetic Data for Small-Data Document Image Classification

Jonathan Kästner[1*], Prof. Dr. Martin Rupp[2†] and Prof. Dr. Richard Beetz[3†]

[1*]faculty for mathematics and computer science, University Leipzig, Augustusplatz 10, Leipzig, 04109, Germany.
[2]faculty for computer science and business informatics, FOM Karlsruhe, Zeppelinstraße 7d, Karlsruhe, 76185, Germany.
[3]faculty for computer science and business informatics, Provadis Hochschule, Industriepark Höchst, Geb. B845, Frankfurt am Main, 65926, Germany.

*Corresponding author(s). E-mail(s): kaestner.jonathan@protonmail.com;
Contributing authors: martin.rupp@fom.de; Richard.Beetz@doz-provadis-hochschule.de;
†These authors contributed equally to this work.

Abstract

This study investigates training and data generation strategies for document error classification in low-data, imbalanced scenarios. We evaluate four models: SVM with SIFT features, a standard CNN, a contrastively pretrained CNN, and a ResNet50 with transfer learning. These are compared across sampled, upsampled, synthetic, and mixed datasets using accuracy and Cohen's Kappa. Contrary to expectations, both contrastive pretraining and transfer learning offered limited benefits, highlighting the challenges of applying large-scale, non-document-specific pretraining to document errors. A supervised CNN trained on mixed data, combining synthetic and upsampled images, achieved the best results, improving accuracy by up to 7% and Kappa score by up to 30%. In contrast, using synthetic data alone resulted in a performance decrease, while increasing dataset size had minimal impact. These findings emphasize that domain-specific data balancing outperforms pretrained models in low-resource document classification. Additionally, mixing synthetic with upsampled data offers a cost-effective and hardware-efficient approach to address data scarcity and class imbalances.

1 Introduction and Research Questions

Document management is a critical operation across industries, impacting areas from HR to finance, legal, healthcare, and logistics. Every year, human resources (HR) departments at large companies process thousands of employee-related documents. The demand for automation in document processing has never been greater. According to a recent market report, "The global intelligent document processing market size was estimated at USD 2.30 billion in 2024 and is projected to reach USD 12.35 billion by 2030,..." [1]. This surge reflects a broader trend toward integrating machine learning (ML) into core business operations.

However, document datasets used in research typically do not resemble real-world industry data. They are often single-page datasets with many examples. In contrast, industry datasets are typically smaller, as data for automating smaller, niche use cases can be scarce. Additionally, classes such as document errors are less common in industry data. For example, forms are often submitted without errors, or only final documents were kept in the past. This results in imbalanced datasets that are difficult to train on. In research, we have limited resources to address these challenges, which makes bridging the gap between research and industry especially critical [2].

To overcome these limitations, we propose a data generation and upsampling framework in the following chapters. We describe the components and test environments, and establish a benchmark by comparing multiple models with various data sampling approaches. The results are evaluated using accuracy and Cohen's kappa score.

2 Results

A dataset of real-world company documents was collected. The dataset consisted of images and scans of standard HR documents, each containing two pages, which were combined side-by-side to create a single image for processing. Some of the documents contained errors that the models were tasked with classifying. The dataset was highly imbalanced, with some classes containing as few as one example. These classes, containing only one example, were exclusively included in the test set. In total, the dataset comprised only 235 images, which represents a typical industry scenario that is challenging to handle due to both class imbalances and data scarcity.

To evaluate the impact of synthetic data generation, four different data distributions were created: the first contained only 88 real images (sampled), the second contained only synthetic images (synthetic), the third consisted of upsampled real images to balance the class distribution (upsampled), and the fourth combined real upsampled and synthetic images in a 1:1 ratio (mixed). These synthetic, upsampled, and mixed distributions were created for three different dataset sizes: 700, 7,000, and 40,000 images. On each distribution, four different models were trained: a SIFT/Bag of Visual Words/SVM classifier, a 7-layer ResNet trained from scratch or with contrastive pretraining, and a ResNet50 model with transfer learning. Contrary to expectation neither transfer learning nor contrastive pretraining improved the metrics significantly. The best overall performance was achieved by the 7-Layer ResNet trained from scratch (CNN, Figure 1).

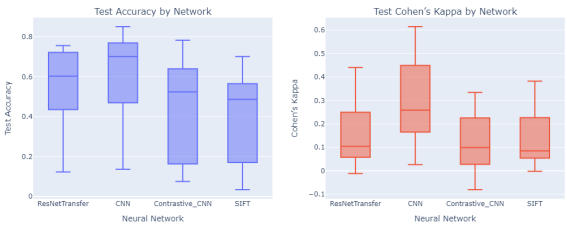


Figure 1 Performance of models aggregated by data distribution and data amount

To assess the effect of these different data distributions, the four models were trained on each dataset and then compared. Accuracy was used to evaluate the impact of data distribution and dataset size on prediction performance. Since the dataset was heavily imbalanced, Cohen's kappa score was also calculated to better account for

the prediction of minority classes. The aggregated model performances were then statistically tested using the Friedman test, followed by the Bonferroni-Dunn post-hoc test as recommended by [3].

2.1 Data Distribution

When comparing data distributions, the mixed distribution achieved the highest maximum accuracy and the best overall Cohen's Kappa score. Notably, the Kappa score showed a more than +20% increase in the median and a +30% increase in the maximum compared to the second-best values. The sampled distribution achieved the highest median accuracy but had the worst Kappa score. Upsampling resulted in a decrease in accuracy (-17%) compared to the sampled distribution, but a (+7%) increase in the Kappa score. The synthetic distribution exhibited the lowest accuracy but slightly outperformed the sampled distribution in terms of the Kappa score.

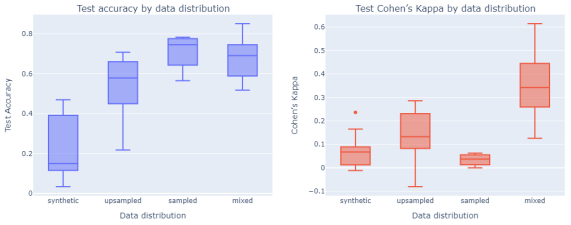


Figure 2 Performance of all models by data distribution

Since the number of observations per group must be equal, the data were aggregated. The upsampled, synthetic, and mixed distributions were reduced, resulting in more samples being evaluated for different dataset sizes. To test statistical significance, these distributions were aggregated for both median and maximum performance across the three different dataset sizes. This approach resulted in four values per data distribution, with each of the four models evaluated for each metric.

The effects of data distributions were analyzed across various aggregation methods. When metrics were aggregated by maximum, the results showed statistically significant effects for both accuracy ($Q = 10.385$, $p = 0.016$) and Cohen's Kappa ($Q = 11.100$, $p = 0.011$). Similarly, with median aggregation, accuracy ($Q = 9.300$, $p = 0.026$) and Cohen's Kappa ($Q = 11.100$, $p = 0.011$) also showed significant differences. Since all p-values were below the standard threshold of 0.05, the null hypothesis (H0) was rejected in favor of the alternative hypothesis (H1), suggesting that data distribution significantly influenced model performance. To identify which specific distributions contributed to this effect, a post-hoc analysis was conducted by comparing each distribution to the sampled baseline.

The upsampled and mixed distributions showed statistically significant improvements over the baseline in Cohen's Kappa score for both aggregation methods. In contrast, the synthetic distribution performed significantly worse in terms of accuracy for both aggregation methods. However, a statistically significant decline in Kappa score was only observed for the maximum aggregation method. The median aggregation showed a similar trend, but without reaching statistical significance.

Table 1 P-values of post-hoc Bonferroni–Dunn test: comparing data distributions with sampled baseline (significant values in bold).

	Median		Maximum	
Model	**Accuracy**	**Cohen's κ**	**Accuracy**	**Cohen's κ**
upsampled	0.083	**0.021**	0.149	**0.021**
synthetic	**0.021**	0.387	**0.021**	**0.043**
mixed	0.773	**0.021**	1.000	**0.021**

3 Data Amount

In the results, all distributions, except the sampled distribution, were compared based on the amount of data. In general, increasing the dataset size improved both metrics in most cases. The performance ranking was as follows: mixed, upsampled, and synthetic, across all dataset sizes

for both metrics. For the synthetic data, accuracy decreased at 7,000 data points but increased again at 40,000 data points. This performance dip was also observed in the mixed distribution, with a slightly decreased median accuracy. Nevertheless, the highest maximum accuracy and Kappa score were achieved in the mixed distribution at 7,000 data points. Maximum performance for both mixed and upsampled distributions slightly decreased at 40,000 data points. However, the median performance continued to increase. The synthetic distribution exhibited the largest performance improvement at 40,000 data points in both metrics.



Figure 3 Performance of all models by data amount and distribution

To evaluate whether the data amount alone had a significant effect across all distributions, the Friedman test was conducted. For accuracy ($Q = 0.167$, $p = 0.920$) and Cohen's Kappa ($Q = 0.667$, $p = 0.717$), the resulting p-values were well above the 0.05 significance threshold. Therefore, the null hypothesis was not rejected, indicating that data amount did not have a statistically significant effect on performance. Consequently, no post-hoc tests were performed.

4 Discussion

This work demonstrates that data composition can have a greater impact on document image classification performance than model complexity. By combining synthetic and upsampled document images, we propose a simple yet effective strategy that significantly improves performance on small, imbalanced, multipage datasets. This finding challenges the common assumption that transfer learning or contrastive pretraining are necessary for strong document classification and instead emphasizes the effectiveness of domain-specific data balancing.

Contrary to similar studies, neither transfer learning nor pretraining significantly improved prediction accuracy [4]. The ImageNet domain used for transfer learning may have been too dissimilar from document images. Contrastive pretraining might have underperformed due to poor augmentation combinations, smaller batch sizes, or insufficient training times, as noted by [5]. Frameworks like MoCo might perform better in future studies with less resources [6].

Mixed data outperformed all other approaches. This result demonstrates that synthetic and upsampled data address complementary weaknesses: synthetic data adds variability, while upsampling strengthens the real distribution of data. Together, they create a more balanced feature space. This insight could potentially be generalized to other image classification tasks.

Unlike prior work, which treated synthetic and real data augmentation as separate strategies, this study identifies their combination as a novel and empirically validated approach to addressing class imbalance in document datasets. In contrast to [7], we present the combination of synthetic and upsampled data as a more effective strategy. This combination leads to a greater improvement than using either upsampled, real, or synthetic data alone. The results were evaluated on an industry-level dataset containing multipage documents, a small dataset size, and imbalanced data. This approach aligns with the recommendations of [2] for the real-world evaluation of document image classification methods.

5 Methods

5.1 Data

A total of 235 document images were collected, all of which are standardized HR-related forms from a large German IT and Telco company, each consisting of two pages. Due to data retention policies, only documents from the past two years were available, limiting the dataset size. The documents were either scanned or photographed

to ensure that handwritten signatures were preserved. As a result, the images vary in quality, angle, and lighting conditions. All documents were initially stored as PDF files and later converted to JPEG format for further processing. The documents share a consistent structure, each containing a PDF417 barcode, a handwritten signature, and information about both the person recommending and the person being recommended. More than 20% of the collected documents contained errors.

The data was manually collected and labeled by members of the company HR team. As these individuals are responsible for the manual validation of such documents as part of their professional duties, they can be considered domain experts. Each document image was annotated with a single error label or marked as correct.

The dataset is highly imbalanced: the majority class, labeled "correct," contains 181 samples. The next most frequent classes—missing barcode, missing signature, and low resolution—each contain only 16–18 instances. The rarest categories—wrong order, invalid barcode, and missing data—are each represented by a single sample. The following is a description of each label: "correct" denotes error-free documents; "missing signature" refers to documents with absent or computer-generated signatures; "low resolution" marks documents whose image quality is too poor for reliable further processing; "missing barcode" identifies cases where the barcode is absent or replaced by a gray placeholder, often due to XFA-PDF conversion; "invalid barcode" refers to damaged or unreadable barcodes; "wrong order" flags documents with their first two pages swapped; and "missing data" is used when the recommended person's name is not provided.

5.2 Data Generation

To address the challenges of limited and imbalanced data, we employed synthetic data generation and image augmentation techniques.

Synthetic images were generated based on the original template. Name data was generated using ChatGPT (**referenced in the AI directory**), with 50% of the names being of German origin, representing diverse ethnic backgrounds, and 50% being international names. This approach aimed to introduce a slight bias toward German names while avoiding overfitting to specific ethnic groups. Each name was sampled independently for both the recommender and the recommended individual.

A unique SAP number, a fixed-length random numeric string, was then generated for the recommender. These names and SAP numbers were used to generate a personalized PDF417 barcode for each document. To simulate signature diversity, the signature dataset from [8] was utilized. Synthetic errors were introduced to reflect the error types observed in the real data. These errors were distributed uniformly across the different error categories.

5.3 Model

To process each document as a single entity, the two pages were merged side by side into one composite image. This approach enabled the use of a single image classification model, simplifying the implementation and reducing model complexity.

Four different classification approaches were evaluated. The first was a Support Vector Machine (SVM) using scale-invariant feature transform (SIFT) and Bag of Visual Words (BoVW) features. The second was a Convolutional Neural Network (CNN) trained from scratch using supervised learning. The third approach involved training a CNN with contrastive pretraining. Finally, a ResNet50 model, trained with transfer learning using ImageNet weights, was evaluated.

The SVM/SIFT classifier served as the traditional baseline. It utilized an SVM trained on SIFT features combined with a Bag of Visual Words (BoVW) representation. This method was chosen based on its strong performance in prior document image classification research [9–11].

A convolutional neural network (CNN) was trained from scratch. The architecture was inspired by ResNet [12] but was scaled down to meet computational constraints. The implemented model included a 7×7 convolutional layer

with a stride of 2, followed by max pooling (to reduce the input size), three convolutional blocks (each reducing spatial dimensions while increasing filter depth), and skip connections between blocks. These skip connections, consisting of a single network layer added between blocks, facilitated the training process, as described by [12]. The network concluded with an average pooling layer and a fully connected layer with a softmax output over the seven error categories.

This CNN was trained both in a supervised fashion and with contrastive pretraining. Due to resource limitations, contrastive pretraining was applied only to this architecture. The contrastive pretraining was performed as described by [5], with an additional network layer added before the similarity calculation during pretraining. This final layer and the network head were subsequently removed. A logistic regression classifier was trained on the network's output. The supervised training used the same data distribution, but with labeled data.

In addition, a ResNet50 model was employed with pretrained ImageNet weights. Its final layer was replaced with a softmax layer to classify the seven document error types. This transfer learning setup leverages large-scale image representations and was chosen due to its strong performance in related work [4].

5.4 Experimental Setup

The experimental process consisted of data preparation, dataset splitting, model training, and evaluation using appropriate metrics to assess performance on real-world document data.

All documents, each containing two pages, were manually annotated by HR experts. To prepare the input for the models, the two pages of each document were padded if necessary, concatenated side by side, resized to 256×256 pixels, and normalized. This unified format enabled each document to be processed as a single image. As noted by [2], multi-page document classification remains a relatively underexplored area, partly due to the limited availability of publicly accessible datasets.

To create a representative test set, labels with only a single occurrence in the entire dataset were included exclusively in the test set. This ensured that model performance could be assessed on all error types—even rare ones—by including at least one example of each in the evaluation data. The final test set consisted of 147 images. Its size was determined using standard sample size estimation techniques for proportions, assuming a 95% confidence level and a 5% margin of error, as outlined in [13]. This sample size is sufficient to draw reliable inferences about model performance on the complete dataset of 235 images.

The remaining images were used to construct different training distributions:

- **Sampled:** All real images not included in the test set. This distribution contained only the four error labels with more than one instance.
- **Upsampled:** The sampled dataset was upsampled with replacement so that each class reached the same number of samples, addressing class imbalance.
- **Synthetic:** Training data was generated using the synthetic generation pipeline described earlier.
- **Mixed:** A combination of synthetic and upsampled real images. The real images were upsampled to match the number of synthetic samples per class.

To evaluate the effect of dataset size, synthetic and mixed datasets were generated at three scales: 700, 7,000, and 40,000 samples—corresponding to 100, 1,000, and 5,714 samples per class across seven labels. These scales were chosen to represent different dataset sizes while being constrained by computational limitations.

Each training distribution was further split into training and validation subsets. The validation set was used for hyperparameter tuning, while the test set remained untouched until the final evaluation, ensuring unbiased performance estimates on real-world data.

Image augmentations were randomly applied during training to enhance generalization and simulate document variations encountered in practice. An end-to-end training pipeline was implemented to streamline preprocessing, augmentation, model training, and evaluation. The models evaluated were:

- **SIFT**: Support-Vector-Machine-Klassifikator auf Basis eines Bag-of-Visual-Words-Modells aus SIFT-Merkmalen (im Text als „SIFT“ bzw. „SIFT/SVM“ bezeichnet).
- **CNN**: Von Grund auf neu trainiertes, überwachtes CNN mit 7 Schichten.
- **Contrastive_CNN**: Dieselbe CNN-Architektur, zunächst kontrastiv vortrainiert, anschließend mit einem Logistic-Regression-Klassifikationskopf auf den Ziel Daten supervidiert feinabgestimmt.
- **ResNet50**: Transfer-Learning mit einem ResNet50, bei dem die letzte Schicht durch eine neue Klassifikationsschicht ersetzt und das Netz auf ImageNet-Gewichten vortrainiert wurde.

Each model was trained on the respective training distribution, and the best-performing version (based on validation set performance) was selected for final evaluation on the test set.

Two evaluation metrics were used: accuracy, a standard metric in classification tasks, and Cohen's Kappa. Accuracy reflects the proportion of correctly classified samples but can be misleading in imbalanced datasets, as high accuracy may simply reflect dominance by the majority class. Therefore, Cohen's Kappa was used as a secondary metric. It accounts for the agreement expected by chance, providing a more balanced evaluation of model performance across uneven class distributions. This is particularly important in this context, where some error classes are rare. The Kappa score was used to assess the quality of predictions for minority classes [14, 15].

Both metrics were then tested for statistical significance. Since normality could not be assumed, a non-parametric test was chosen. The Friedman test was selected, as it allows the comparison of more than two models, as recommended by [3]. Additionally, a post-hoc analysis using the Dunn test with Bonferroni correction was applied to identify which methods were statistically significant. All methods were compared to the baseline, rather than with each other, to avoid artificially inflated p-values due to cross-comparison of approaches, as suggested by [3]. The baseline model for comparison was the SVM/SIFT classifier, and for data distributions and data amount, the sampled distribution served as the baseline.

The null hypothesis H_0 stated that no difference in the metric being compared would be found, given the independent variable. The alternative hypothesis H_1 posited that a difference would exist. A significance threshold of $\alpha = 0.05$ was used for rejecting the null hypothesis, a commonly accepted threshold in the literature. Details of these calculations can be found in [3, 16, 17].

Declarations

The authors have no affiliation to commercial products connected to this research. The image data resolution was compressed such that personal identification is not possible. Data, code and other materials are available to eligible parties on request.

Appendix A Experiment Results Data

net	data_distribution	data_amount	train_accuracy	test_accuracy	test_cohens_kappa
ResNetTransfer	synthetic	40000	0.4375	0.4558	0.0884
CNN	synthetic	40000	0.7188	0.4694	0.2363
Contrastive_CNN	synthetic	40000	0.4547	0.0748	0.0274
SIFT	synthetic	40000	0.4399	0.1701	0.0617
CNN	upsampled	7000	1	0.6939	0.2855
ResNetTransfer	upsampled	7000	0.9375	0.6054	0.1179
Contrastive_CNN	upsampled	7000	0.8869	0.2177	-0.0803
SIFT	upsampled	7000	0.7995	0.4694	0.107
ResNetTransfer	upsampled	40000	0.9375	0.5986	0.1733
SIFT	upsampled	40000	0.7973	0.5034	0.1179
CNN	upsampled	40000	1	0.7075	0.2811
SIFT	sampled	147	0.7536	0.5646	0.0628
Contrastive_CNN	upsampled	40000	0.9742	0.6395	0.2252
Contrastive_CNN	upsampled	700	0.8839	0.5578	0.147
CNN	upsampled	700	1	0.6803	0.2364
ResNetTransfer	upsampled	700	1	0.4286	0.0579
SIFT	upsampled	700	0.7339	0.3469	0.0541
Contrastive_CNN	mixed	40000	0.8916	0.517	0.1257
SIFT	mixed	40000	0.5818	0.7007	0.3824
ResNetTransfer	mixed	40000	0.8125	0.7551	0.4404
CNN	mixed	40000	0.875	0.8231	0.5864
ResNetTransfer	mixed	7000	0.7188	0.6122	0.2494
CNN	mixed	7000	0.8125	0.8503	0.6145
Contrastive_CNN	mixed	7000	0.6936	0.5306	0.2697
SIFT	mixed	7000	0.5711	0.6803	0.3512
Contrastive_CNN	sampled	147	0.7681	0.7823	0
Contrastive_CNN	synthetic	700	0.3536	0.1088	0.0555
Contrastive_CNN	synthetic	7000	0.4532	0.1633	0.0731
Contrastive_CNN	mixed	700	0.5929	0.6395	0.3342
SIFT	mixed	700	0.4696	0.5646	0.2264
SIFT	synthetic	7000	0.4304	0.034	-0.0019
SIFT	synthetic	700	0.3643	0.1361	-0.0019
CNN	mixed	700	0.6875	0.7347	0.4495
CNN	sampled	147	0.6	0.7687	0.0263
CNN	synthetic	700	0.6875	0.1361	0.0881
CNN	synthetic	7000	0.8125	0.3469	0.1649
ResNetTransfer	synthetic	7000	0.4375	0.1224	-0.0114
ResNetTransfer	synthetic	700	0.4375	0.4354	0.0904
ResNetTransfer	sampled	147	1	0.7211	0.0483
ResNetTransfer	mixed	700	0.625	0.7347	0.2714

Figure A1 Experimental Results

References

[1] Grand-View-Research: Intelligent Document Processing Market Size, Share & Trends Analysis Report By Component (Solution, Services), By Technology (ML, NLP), By Deployment, By Organization Size, By End-use (Manufacturing, Retail), By Region, And Segment Forecasts, 2025 - 2030 (2024)

[2] Van Landeghem, J., Biswas, S., Blaschko, M., Moens, M.-F.: Beyond Document Page Classification: Design, Datasets, and Challenges. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2962–2972 (2024)

[3] Demsar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets (2006)

[4] Muhammad Zeshan Afzal, Afzal, M., Muhammad Zeshan Afzal, Afzal, M.Z., Andreas Kölsch, Kolsch, A., Sheraz Ahmed, Ahmed, S., Marcus Liwicki, Liwicki, M.: Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification, 883–888 (2017) https://doi.org/10.1109/icdar.2017.149

[5] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A Simple Framework for Contrastive Learning of Visual Representations. In: Proceedings of the 37th International Conference on Machine Learning, pp. 1597–1607. PMLR, ??? (2020)

[6] Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020)

[7] Fuente, N., Majó, M., Luzko, I., Córdova, H., Fernández-Esparrach, G., Bernal, J.: Enhancing Image Classification in Small and Unbalanced Datasets Through Synthetic Data Augmentation. In: Drechsler, K., Oyarzun Laura, C., Freiman, M., Chen, Y., Wesarg, S., Erdt, M. (eds.) Clinical Image-Based Procedures, pp. 11–21. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-73083-2_2

[8] Sathish Kumar: Offline Handwritten Signature Database Based on Age Annotation (OHSDA). Mendeley (2023). https://doi.org/10.17632/C5MFHR2XCZ.1

[9] Liu, L., Wang, Z., Qiu, T., Chen, Q., Lu, Y., Suen, Y.: Document Image Classification: Progress over Two Decades. Neurocomputing **453** (2021) https://doi.org/10.1016/j.neucom.2021.04.114

[10] V.V. Arlazarov, Vladimir V. Arlazarov, E.I. Andreeva, Elena Andreeva, K.B. Bulatov, Konstantin B. Bulatov, D.P. Nikolaev, Dmitry P. Nikolaev, O.O. Petrova, Olga O. Petrova, B.I. Savelev, B.I. Savelev, O.A. Slavin, Oleg Slavin: Document image analysis and recognition: A survey. Computer Optics **46**(4), 567–589 (2022) https://doi.org/10.18287/2412-6179-co-1020

[11] Rusiñol, M., Chazalon, J., Ogier, J.-M., Lladós, J.: A comparative study of local detectors and descriptors for mobile document classification. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 596–600 (2015). https://doi.org/10.1109/ICDAR.2015.7333831

[12] He, K., Xiangyu Zhang, Zhang, X., Shouxin Ren, Ren, S., Jian Sun, Sun, J.: Deep Residual Learning for Image Recognition, 770–778 (2016) https://doi.org/10.1109/cvpr.2016.90

[13] Ahmed, S.K.: How to choose a sampling technique and determine sample size for research: A simplified guide for researchers. Oral Oncology Reports **12**, 100662 (2024) https://doi.org/10.1016/j.oor.2024.100662

[14] Jude Chukwura Obi: A comparative study of several classification metrics and their performances on data. World Journal of Advanced Engineering Technology and Sciences **8**(1), 308–314 (2023) https://doi.org/10.30574/wjaets.2023.8.1.0054

[15] Rainio, O., Teuho, J., Klén, R.: Evaluation metrics and statistical tests for machine learning. Scientific Reports **14**(1), 6086 (2024) https://doi.org/10.1038/s41598-024-56706-x

[16] Dunn, O.J.: Multiple Comparisons Among Means. Journal of the American Statistical Association **56**(293), 52–64 (1961) https://doi.org/10.2307/2282330 2282330

[17] Haynes, W.: Bonferroni Correction. In: Encyclopedia of Systems Biology, pp. 154–154. Springer, ??? (2013). https://doi.org/10.1007/978-1-4419-9863-7_1213