IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 May 2024, accepted 5 June 2024, date of publication 14 June 2024, date of current version 1 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3414651

RESEARCH ARTICLE

Rethinking Deep CNN Training: A Novel Approach for Quality-Aware Dataset Optimization

BOHDAN RUSYN[1,2], OLEKSIY LUTSYK[1], ROSTYSLAV KOSAREVYCH[1], OLEG KAPSHII[3], OLEKSANDR KARPIN[3], TARAS MAKSYMYUK[3,4], (Member, IEEE), AND JURAJ GAZDA[5]

[1]Department of Remote Sensing Information Technologies, Karpenko Physico-Mechanical Institute, NAS of Ukraine, 79060 Lviv, Ukraine
[2]Department of Informatics and Teleinformatics, University of Radom, 26-600 Radom, Poland
[3]Advanced Systems Research Group, Infineon Technologies, 79000 Lviv, Ukraine
[4]Department of Telecommunications, Lviv Polytechnic National University, 79000 Lviv, Ukraine
[5]Department of Computers and Informatics, Technical University of Kosice, 040 01 Košice, Slovakia

Corresponding author: Juraj Gazda (juraj.gazda@tuke.sk)

This work was supported in part by the National Academy of Science of Ukraine under Project #0117U000519, in part by Ukrainian Government "Development the Innovative Methods and Means for the Deployment of the Dual-Purpose Information Infrastructure for the Digital Transformation of Ukraine" under Project # 0123U100232, and in part by Slovak Research and Development Agency under Project APVV-18-0214 and Project VEGA 1/0685/23.

ABSTRACT The informativeness of data has always been of great interest within the machine learning community. Nowadays, with the skyrocketing advancement of artificial intelligence and massive volumes of noisy data, it becomes even more essential to develop robust and effective methods for training data optimization. Existing approaches are mostly based on empirical trial and error, with either stochastic or deterministic data reduction strategies. The key limitation of such solutions is that they do not consider the overall informativeness of the resulting training dataset. In this paper, a novel approach for quality-aware dataset optimization by initial assessment of its informativeness is proposed. As a metric of informativeness, entropy values are calculated over the target dataset. To alleviate the computational complexity, an initial clustering of the dataset is performed, and the entropy of each cluster is calculated independently. The dataset is then optimized by dynamic programming to find a sequence of subsets with low overall entropy according to imposed size limitations. The experimental evaluation shows that the proposed approach improves over current best alternatives in terms of accuracy, precision, recall, and F1-score metrics. Moreover, the proposed approach provides excellent interclass discrimination even for a large number of classes.

INDEX TERMS Deep learning, dataset reduction, training optimization, CNN.

I. INTRODUCTION

Artificial intelligence (AI) applications are currently skyrocketing in various industries, gaining momentum driven by the tremendous growth of average computational capacity. So far, the semiconductor industry is experiencing a boost in investment in so-called AI-capable hardware such as GPUs, TPUs, and neural accelerators. Nowadays, a wide range of products are developed, from high-end computing systems to tiny microcontrollers with different types of neural accelerators [1], [2].

Consequently, the software side of the AI world is growing even faster, with new algorithms being released every single day. Thus, we are now observing a paradigm shift in AI development. Many new neural networks are explicitly designed and optimized for specific tasks and hardware, rather than relying on a typical universal "training black-box for any task" approach [3].

In this context, choosing a proper dataset to train a neural network is of pivotal importance to achieve target performance. Irrelevant or noisy data samples can jeopardize

The associate editor coordinating the review of this manuscript and approving it for publication was Weiren Zhu.

the model's ability to generalize information. In addition, all classes in the dataset should be balanced to avoid biases, which could lower the accuracy for minor classes [4], [5]. Moreover, to avoid overfitting, the training dataset must be sufficiently large to comprehensively reflect the addressed problem.

Particular attention should be given to the zero-data training problem, where training data are not available for some classes [6], [7]. Typically, the data for these classes should be represented by some alternative numerical descriptions in the form of embedding vectors. A typical example is the face recognition problem, where the classifier is trained on a large subset of persons but can then be effectively applied to recognize other persons who never appeared in the training dataset [8]. This is achieved by exploiting the natural architecture of deep neural networks, which consist of a feature extractor and a classifier [9]. In computer vision problems (e.g., face recognition), the feature extractor is commonly based on a convolutional neural network (CNN), which extracts unique feature embeddings from each image. These are then classified in the dense neural network classifier (e.g., multi-layer perceptron) during training [10]. For deployment, the classifier can be removed, and the feature extractor can be used to compare feature embeddings of known and unknown classes using various distance metrics such as Euclidean, cosine, etc. In the case of zero-data training, artificial feature embeddings of the zero-class can be generated and added directly to the classifier, along with the embeddings extracted by the feature extractor from the raw data. Additionally, some augmentation methods can be employed to enhance the dataset volume or mitigate class imbalance [11], [12], [13].

The widely accepted approaches for training dataset selection in deep learning usually involve a combination of normalization, balancing, and tuning. However, these approaches are based on empirical trial and error and do not assess the informativeness of the training dataset. Informativeness refers to the quality and relevance of the information in the dataset for solving a specific machine learning task. An informative dataset should effectively represent the key patterns and variations that the model can encounter in its target application domain [14], [15]. Such a dataset should include diverse, representative examples covering different scenarios while maintaining minimal volume to satisfy the required performance for a given task.

In this paper, a novel approach for quality-aware dataset optimization is proposed through the initial assessment of its informativeness. The motivation for this research is based on the possibility that two datasets of equal size may differ in informativeness. While both datasets may have the same training complexity, training with the more informative one would be more efficient. To measure informativeness, the entropy of the dataset is proposed as an evaluation metric. Entropy, which measures uncertainty or randomness in a dataset, can be used to quantitatively assess the informativeness of training samples [16]. The entropy of a training dataset measures how much information, on average, is needed to describe or predict the outcome of a random sample from this dataset.

In essence, entropy in this context serves as a metric for evaluating how much unique information each element of the dataset contributes to the overall understanding of the data distribution. Too high entropy indicates a greater level of uncertainty or diversity within the dataset, suggesting that the data contain a wide array of features and patterns that decrease informativeness. Conversely, too low entropy implies less diversity and excessively high informativeness of samples. This can result in overfitting during training and decrease the robustness of the neural network to noisy data in real-world tasks.

To alleviate this problem, the whole dataset is clustered into several subsets, and the entropy values in each subset are controlled independently. This approach reduces the probability of too low entropy values in the entire dataset. Another important advantage of the proposed approach is that entropy is calculated among the feature embeddings, after the feature extractor, making it suitable for the above-mentioned zero-data training problem.

The major contributions of this paper are as follows:

1) A novel approach for dataset optimization is proposed, using entropy as a metric of informativeness.
2) Dataset clustering into multiple subsets is employed, and intra-cluster entropy is calculated for each subset independently.
3) Simulations are conducted to compare the efficiency of the proposed approach with alternative solutions.

The remainder of this paper is organized as follows. In Section II, existing research on dataset optimization for a better trade-off between cost and performance is reviewed. In Section III, the proposed quality-aware dataset optimization approach with entropy analysis is described. Section IV presents comparative simulation results and discussion. Section V offers additional discussion on the results and practical applications of the proposed approach. Finally, the article is concluded in Section VI.

II. RELATED WORK

A. AN OVERVIEW OF DATA INFORMATIVENESS AND GENERALIZATION CAPABILITIES OF AI MODELS

Understanding the informativeness and complexity of datasets in deep learning has gained significant attention, particularly in relation to the generalization capabilities of neural network models. In [17], Zhang et al. explored a remarkable phenomenon of deep learning models. According to numerous experiments, an arbitrary neural network can easily fit data with completely random labels, even when there is no relation between the real class of the sample and its label in the dataset. In practice, this means that regardless of mismatch and imbalance between data and labels, a neural network of sufficient size can achieve 100%

training accuracy. Furthermore, the authors investigated that after replacing real images with complete Gaussian noise and assigning them random labels, the neural network is still capable of achieving 100% training accuracy. Essentially, this means that the effective capacity of neural networks is sufficient for memorizing the entire dataset even if it is statistically meaningless. This work emphasizes the need to reconsider our understanding of generalization in the context of deep learning, indirectly touching upon the intrinsic complexities and informativeness of training datasets.

Parallel to these insights, the authors in [18] propose methodologies for the quantitative assessment of dataset complexity and informativeness by examining the performance of deep learning models with variable settings. Their approach provides a framework for evaluating how different types of data affect learning processes, thus serving as a valuable tool for researchers aiming to optimize training modes and model architectures depending on data characteristics.

The impact of individual data samples in a dataset as a measure of informativeness has been addressed by Ghorbani and Zou through the concept of Data Shapley value [19]. Deriving from cooperative game theory, the authors proposed an approach to quantify the contribution of each data point to the achieved accuracy of a deep learning model. In this context, Shapley values effectively serve as a metric of informativeness for training purposes.

In the broader discussion on the underlying mechanics of machine learning models, the authors in [20] argue against the prevailing notion of interpretability and informativeness, pointing out that these concepts are often misapplied or misunderstood within the machine learning community. The authors emphasize the importance of clarity of the informativeness concept in model development and evaluation.

Further explorations investigate whether specific directions in the activation space of deep networks are crucial for generalization. The findings in [21] suggest a systematic methodology to assess the generalization capability of machine learning-based solutions via a novel feature extraction pipeline.

In most cases, studies [22], [23] establish a correlation between the size of the training dataset and the probability of correct classification. These approaches reveal certain properties of the training dataset as well as the specific learning model. If the model's parameters are appropriately chosen, there generally exists a relationship where increasing the size of the training dataset leads to improved classification accuracy. This is attributed to the fact that a larger training set allows for the model to be trained with more representative features, reducing the likelihood of overfitting. Among various solutions for the extraction of more informative features, principal component analysis (PCA) is the most widely used, along with its various derivatives [24], [25], [26].

B. ENTROPY-BASED ASSESSMENT OF DATA INFORMATIVENESS IN DEEP LEARNING

Recently, the use of entropy as a measure of informativeness in machine learning has been extensively studied, reflecting its significance in improving model performance and decision-making processes. The integration of entropy in active learning methodologies is well-illustrated by work exploring representativeness and informativeness for active learning [27]. This approach utilizes entropy measures to select samples that could provide the most informative data for model training, thereby optimizing learning efficiency and effectiveness.

The paper [28] introduces methods to estimate bounds on entropy to find the contribution of different variables within a model to the overall informativeness. This method provides an accurate model evaluation and enables new applications for data-driven systems, particularly in medical diagnostics and personalized medicine.

The impact of entropy on real-world datasets and its implications for machine learning are discussed in [29]. The authors proposed an entropy-based measure to capture the nonredundant, noncorrelated core information from the data by using well-known algorithms from the classification domain to investigate the quality of the proposed solution. The paper highlights the practical challenges and benefits of applying entropy in diverse environments, emphasizing its role in feature selection, model optimization, and performance enhancement.

The utilization of entropy in classification tasks is addressed in [30]. The authors proposed to utilize changes in entropy-based features for the classification of different types of DDoS (Distributed Denial of Service) attacks.

Lastly, the broader application of entropy was found in [31], where the authors have proven that higher entropy increases a lower bound on a robust objective in reinforcement learning tasks. These findings can be used to learn robust policies that can handle various disturbances in the learning dynamics and the reward function.

The works summarized in Table 1, serve as a foundational guide for understanding the role of entropy in improving the decision-making processes of various machine learning algorithms. Nevertheless, there are still some research gaps in assessing dataset informativeness, particularly in the computer vision domain, which is dominated by unstructured image data.

C. UNRESOLVED CHALLENGES IN ENTROPY-BASED ASSESSMENT OF DATA INFORMATIVENESS

The entropy of a sample is used as a metric of uncertainty. When the entropy is low, it implies a nearly uniform distribution of possible training vectors, with low uncertainty. In this case, the training dataset is more informative as the possible outcomes are more predictable and carry more specific information. Models can be trained and generalized more easily with such training vectors. On the other hand,

TABLE 1. Overview of the existing solutions for training optimization and dataset quality estimation in machine learning.

Reference	Main contribution	Limitations
[17], [21]	Reveal the generalization capabilities of neural networks	Do not assess dataset informativeness
[18]	Quantitative assessment of dataset complexity and quality	No clear measure to estimate quality of arbitrary dataset. Many different approaches for different datasets.
[19]	Cooperative game theory approach for evaluation of informativeness	Limited to single samples, complex for large datasets
[20]	Rethink the concept of model interpretability and transparency	Focus on the models, rather than data
[22], [23]	Investigate the relation between dataset size and classification accuracy	Focused more on the number of samples rather than quality
[24]–[26]	Investigate the feature engineering to improve the training performance of the model	Evaluates the importance of the features, not informativeness of the sample
[27]–[30]	Entropy-based measures to select the most informative data samples	Limited to structured data
[31]	Utilize entropy to learn robust policies in reinforcement learning tasks	Not suitable for classification tasks

if the entropy of the sample is high, it indicates a more dispersed and uncertain distribution of possible vectors, making the training dataset less informative. In addition, samples with high entropy are often noisy, ambiguous, and lack clear distinctions between different classes.

Thus, it's essential to keep the entropy of the training samples low to preserve valuable information for the efficient training of machine learning models. Otherwise, training with high entropy data can be less informative and may require larger data volumes, longer training times, and even additional input context to achieve the same performance of the machine learning models.

The conventional approach to calculate the entropy of a training sample involves the following steps:

- Identifying the event of interest from the perspective of deep learning models, which in this case would be the prediction the model has to make for a given input vector.
- Calculating the probabilities of model predictions for each possible vector based on the training dataset. For a training sample, this involves estimating the probability for each element of the sample.
- Computing the entropy for the aggregate set based on the previously calculated probabilities.

The main limitation of this approach is that it considers a general case where entropy is calculated at a global level and can only deal with structured data, such as:

- Tabular data represented in structures with rows and columns. Each row represents a sample element, and each column represents a characteristic or feature of that element. This type of data is often stored in CSV format.
- Training samples based on data providing a hierarchical structure using tags. Such data with relationships between different elements is commonly found in XML format, facilitating the easy formation of a feature vector.
- Training samples can be effectively formed based on other types of data where information is organized in a structured form with specific rules and relationships among the elements.

To evaluate the informativeness of unstructured data such as images, preliminary processing and feature extraction are necessary [32], [33], [34]. Preliminary processing can yield features better suited to estimate the informativeness of image data.

D. FEATURE EXTRACTION FROM UNSTRUCTURED IMAGE DATA

Currently, several highly effective preprocessing methods for feature formation are recognized, each with its own strengths and weaknesses: Scale-Invariant Feature Transform (SIFT) is used for keypoint detection and descriptor extraction. This algorithm identifies keypoints that are invariant to image scaling, rotation, and changes in illumination [35].

Histogram of Oriented Gradients is an effective method for object detection in images. It calculates gradient orientations and magnitudes in local image areas, forming histograms to capture shape information [36]. It is useful for feature extraction from unstructured data.

Color Histograms are based on the distribution of color information in images. Features extracted using this method are sensitive to color and consider statistical moments (mean, variance, skewness) of color distribution [37].

Local Binary Patterns are used for texture analysis. The algorithm encodes local texture information by comparing the intensity values of a central pixel with its neighboring pixels, building chains of features [38].

Gabor Filters are employed for extracting texture features from images. They are sensitive to different orientations and frequencies, useful for texture information analysis [39].

Local Phase Quantization is based on the quantization of local phase. It is used as a texture descriptor that encodes information about the local phase in images, robust to noise and changes in illumination [40].

Autoencoders are neural network models based on unsupervised learning principles. They can be used to obtain a more compact and generalized image representation, useful for data dimensionality reduction, feature extraction, or pre-training deep learning models [41].

Convolutional Neural Networks (CNNs) are deep learning models [42]. They can obtain hierarchical image characteristics during training, making them versatile and adaptable to various training samples and learning algorithms.

In the context of data informativeness evaluation, CNN-based methods are the most suitable. Additionally, the CNN feature detection method has several advantages over others, primarily in its adaptive feature determination directly from raw input data. This is especially useful when working with complex and multidimensional data such as images. CNNs can establish hierarchical feature dependencies, capturing both low-level and high-level semantic features, like object shape characteristics. They adapt well to the characteristics of the data they are trained on. By adjusting the model architecture and dataset size, CNNs can efficiently learn to extract features most relevant to a specific problem. They can also be trained on large datasets like ImageNet and used for feature extraction in related tasks. Thanks to their good generalization properties, CNN models can easily adapt to work with other training samples, crucial for real-world transfer learning tasks.

In terms of computational capabilities, CNNs can be efficiently parallelized on graphics processors, making them computationally effective for large-scale datasets. Overall, using CNNs for feature extraction provides an efficient approach that can significantly enhance accuracy and reliability in image analysis tasks compared to traditional feature extraction methods.

A pre-trained model like ResNet50, which was trained on the ImageNet training dataset, can be used for CNN-based feature extraction. Since the model is already pre-trained on a specific image dataset, the images for feature extraction should match the size of those on which the model was trained, namely 224 × 224 pixels. During the forward pass through the model, input data are propagated layer by layer to the last convolutional layer, after which it is frozen. These activations are the desired features obtained from the input data, which can be used as feature vectors for various purposes, including entropy-based informativeness assessment. Depending on the target goal, the convolutional feature vector can be used not only based on the last convolutional layer but also on any layer of the network that may represent particular interest.

III. A QUALITY-AWARE DATASET OPTIMIZATION BASED ON ENTROPY ANALYSIS

A. COMMON SOLUTIONS FOR DIMENSIONALITY REDUCTION

In the area of machine learning, dimensionality reduction is a powerful process that transforms intricate, high-dimensional data into a more manageable, low-dimensional format without sacrificing the essence of the original information. The task of dimensionality reduction is not only to reduce the volume of data but also to preserve the core characteristics that represent the value and meaning of the data.

This becomes necessary when working with large datasets, where training can be computationally expensive or time-consuming. Several approaches can address this issue.

1) RANDOM SUBSET SELECTION

Involves randomly selecting a percentage of data from the training sample. This method is quick and simple but may not retain all critical elements in the data, as it is unguided in terms of informativeness.

2) STRATIFIED SAMPLING

This approach ensures that the reduced dataset preserves the class distribution of the original data [43]. It is more suitable if the dataset is unbalanced.

3) CLUSTERING

Used to group similar subsets of data. Then, the redundant elements, which do not contribute additional informativeness, are removed to obtain a smaller but representative dataset.

4) IDENTIFYING ACTIVE ELEMENTS

Focuses on the most crucial elements in the training process. This allows determining the most informative instances in the dataset and removing the uninformative redundant samples [44].

5) ACTIVE LEARNING

Unlike previous approaches, which focus on reducing the dataset by removing non-informative samples, active learning initially trains the model on a small dataset and iteratively adds additional samples to improve the model's accuracy and precision [45]. This approach allows finding the smallest subset of data that satisfies the target performance indicators of the trained model.

B. OPTIMAL SELECTION OF LOW-ENTROPY DATA SUBSETS

In the diverse space of existing solutions for dimensionality reduction and their shortcomings, a new approach is proposed, combining advantages and alleviating existing limitations. In computer vision, most datasets are extremely large and diverse, complicating the assessment of informativeness by entropy calculation. To tackle this problem, the training dataset is divided into many smaller clusters based on the similarity of samples. Entropy is then calculated within each cluster to assess its informativeness [16]:

$$H\left(X\right) = -\sum_{i=1}^{n} p\left(x_i\right) log_2 P\left(x_i\right), \tag{1}$$

where $X = \{x_1, x_2, \ldots, x_n\}$ is a discrete random array with a probability distribution $\{p(x_1), p(x_2), \ldots, p(x_n)\}$.

The important factor here is to ensure that the number of clusters is larger than the number of classes in the dataset, so that the minimization of entropy within each cluster will not result in an excessively clean dataset.

The straightforward approach would be to compare all entropy values and compose a dataset from the clusters with lower entropy. To achieve better stability, a pairwise assessment of individual clusters with the dynamic composition of smaller datasets from particular clusters is proposed, as described in Fig. 1.

The workflow of the proposed method is described below.

Step 1. Raw images are processed by the pretrained convolutional neural network to obtain the feature map $Fc\left(X_i\right)$ of each image X_i.

Step 2. Obtained feature maps are clustered into many subsets Y_i.

Step 3. For each cluster Y_k, the entropy is calculated to assess its informativeness in general $H(Y_k)$.

Step 4. The absolute difference of the pairwise entropies of all possible pairs of clusters is calculated to create the mutual similarity matrix:

$$S = \begin{bmatrix} 0 & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & 0 & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & 0 \end{bmatrix} \tag{2}$$

where $s_{i,j}$ is an element of the mutual similarity matrix. The mutual similarity matrix is symmetric with zeros on the diagonal elements and represents the similarity of entropy between the i-th and j-th subsets of data.

Step 5. After obtaining the mutual similarity matrix, dynamic programming is applied to determine the longest subsequence of clusters. The main goal here is to find the longest subsequence of clusters with the lowest overall entropy, considering the imposed dataset limitations. This process starts at the bottom-left corner of the matrix S, with further transition to the next smallest element. When there are few alternative paths that lead to the smallest values, a movement occurs in the direction that maximizes the length of the subsequence. As a result, a dynamic matrix D is obtained, where $D_{i,j}$ is the size of the longest sequence ending at element $D_{i,j}$.

Initially, elements of D are defined as follows:

$$D_{i,1} = 0, \quad D_{1,j} = 0. \tag{3}$$

Let's suppose that A and B are two different sequences. Then, in a two-cycle calculation, the result will be the following:

$$D_{i,j} = D_{i-1,y-1} + 1, \quad if\ A_{i-1} = B_{j-1} \tag{4}$$

$$D_{i,j} = max\left(D_{i-1,y}, D_{i,y-1}\right) \tag{5}$$

Step 6. Based on the obtained longest subsequence, the corresponding clusters are selected to form the optimal dataset.

Step 7. Finally, the dataset is checked for inter-class imbalance. If the imbalance does not exceed the allowed threshold, the dataset can be rearranged within the determined subset of optimal clusters.

The advantage of this approach is its lower computational cost and optimized selection during the iterative analysis of the entropy of data aggregates. Since some stages of the algorithm, such as dividing into sub-samples or choosing the initial approximation for finding the maximum length of the sequence from the mutual similarity matrix, are randomly selected, the result may vary, but the informativeness of the obtained dataset will be approximately in the same range.

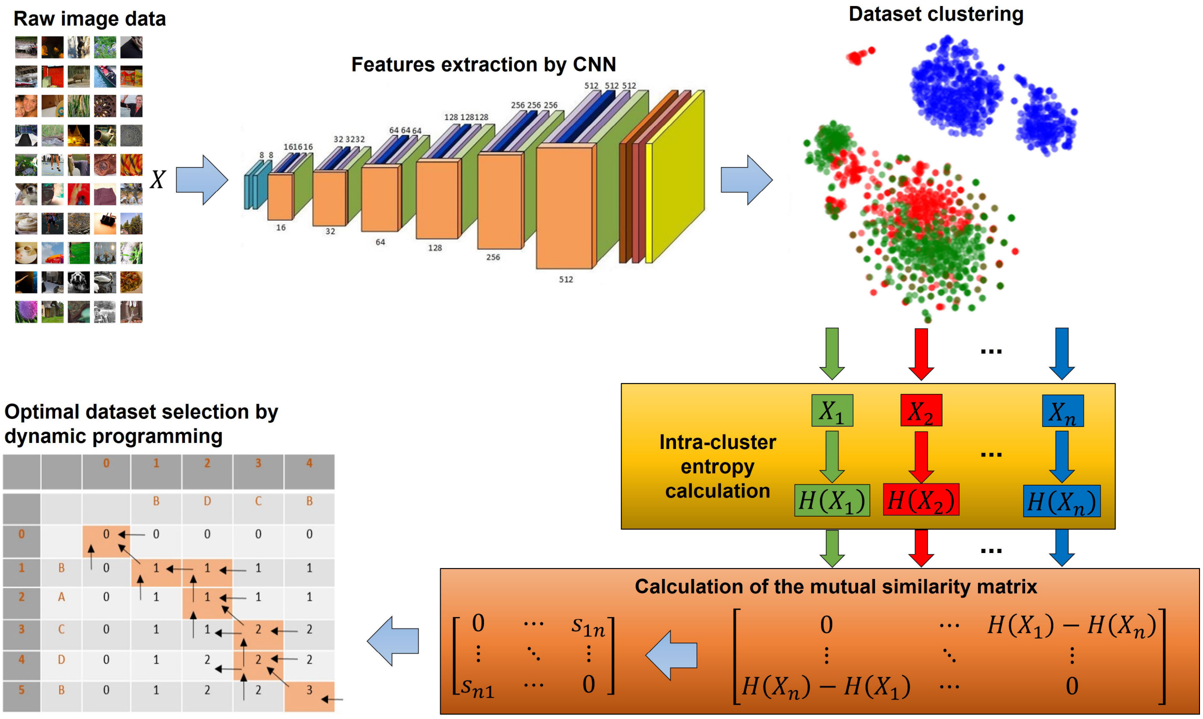
IV. NUMERICAL RESULTS

For validation of the proposed approach, the experimental workflow is designed according to the most widely adopted practices in building computer vision training and testing pipelines [46].

First, two well-known training datasets are selected: MiniImageNet [47] and MNIST [48]. The original MNIST database contains images of handwritten characters, with a total of 60,000 training images and 10,000 test images, covering 10 classes. The image size is 28 × 28 pixels. MiniImageNet, on the other hand, is a simplified version of the famous ImageNet database for training models. MiniImageNet includes 60,000 images, covering 100 classes with 600 images per class. The image size is 84 × 84 pixels. Both training datasets evenly cover the respective classes, and there is no imbalance in class representation.

According to the experimental workflow, a miniResNet CNN model is defined for training and testing. Initially, training is performed on the original dataset. Subsequently, a series of trainings are conducted on various optimized datasets. To assess the performance of the proposed quality-aware dataset optimization in a comparative landscape, several alternative solutions are selected. The baseline for comparison is random subset selection, and the other two alternatives are clustering and stratified sampling. The most common metrics used to evaluate model performance are precision, recall, accuracy, and F1-score. For brevity, not all metrics are presented for both MiniImageNet and MNIST datasets due to their differences. Since the MNIST dataset has only 10 classes, which are equally balanced, the precision, recall, and F1-score metrics are less informative, allowing reliance solely on accuracy values. Conversely, MiniImageNet covers 100 classes, making the accuracy value potentially misleading, while precision, recall, accuracy, and F1-score become much more important metrics.

To ensure that the entropy of the obtained optimized subsets is not too low, the cumulative distribution function is compared before and after applying the proposed approach, as shown in Fig. 2. For clarity, all entropy values are normalized from 0 to 1, so that the entropy distribution among all clusters of the original training dataset follows a normal probability distribution with a mean entropy value of 0.5. After optimization of each cluster, the cumulative density function of normalized entropy moves towards lower values, with a mean at 0.4, and a quasi-uniform distribution within the target range of normalized entropy values (0.2-0.6).


Raw image data
X
Features extraction by CNN
8 8
16
161616
32
32 3232
64
64 64 64
128
128 128128
256
256 256 256
512
512 512 512
Dataset clustering
X_1
X_2
X_n
Intra-cluster entropy calculation
$H(X_1)$
$H(X_2)$
$H(X_n)$
Calculation of the mutual similarity matrix
Optimal dataset selection by dynamic programming

$$\begin{bmatrix} 0 & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \cdots & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & \cdots & H(X_1) - H(X_n) \\ \vdots & \ddots & \vdots \\ H(X_n) - H(X_1) & \cdots & 0 \end{bmatrix}$$

		0	1	2	3	4
			B	D	C	B
0		0	0	0	0	0
1	B	0	1	1	1	1
2	A	0	1	1	1	1
3	C	0	1	1	2	2
4	D	0	1	2	2	2
5	B	0	1	2	2	3

FIGURE 1. Overall workflow of the proposed quality-aware dataset optimization method.

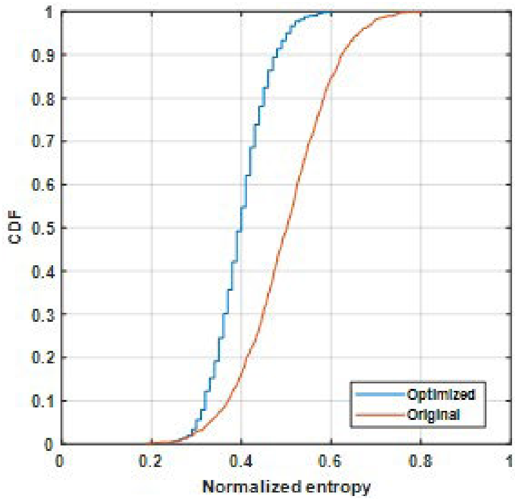


FIGURE 2. Cumulative distribution functions of entropy in the MiniImageNet dataset before and after optimization.

Thus, all compared approaches have been tested using the following procedure.

1) Both MiniImageNet and MNIST datasets are split into training, validation, and testing parts with approximate proportions of 70/15/15, respectively.
2) The model is trained on the original training part of the MiniImageNet dataset, and the precision, recall, and F1-score metrics are evaluated.
3) The model is trained on the original training part of the MNIST dataset, and the accuracy metric is evaluated.
4) The training dataset is optimized by random element selection, clustering, stratified sampling, and the proposed quality-aware optimization method. To determine the boundaries of possible reduction, different size limitations for the dataset, ranging from 2,000 to 15,000 samples, are used.
5) For each obtained training subset, training and evaluation are conducted as described in steps 2 and 3.

Figs. 3 and 4 present the results of the model's precision and recall, respectively, after training on all optimized subsets of MiniImageNet. The number of images per class is evenly balanced.

When the dataset size is large enough, i.e., at least 15,000 images, the difference in precision between all tested approaches is negligible. As seen from Fig. 3, all methods can minimize false positive classifications if the dataset size is larger than 10,000 images. However, with further reduction of the dataset size, the advantage of the proposed approach increases up to 4% in precision for approximately 2,000 images in the dataset. On the other hand, in terms of recall in Fig. 4, it is clearly seen that the proposed approach outperforms all competitive solutions.

When comparing other alternatives, it is observed that the stratified sampling method has an advantage over random selection in terms of balance between classes. The downside of clustering is the difficulty of accurately predicting the training dataset size because it is not possible to influence the cluster detection process of the algorithm. However, due to its nature, clustering can maintain good precision even for 6,000 images in the dataset because the loss of informativeness is not severe.

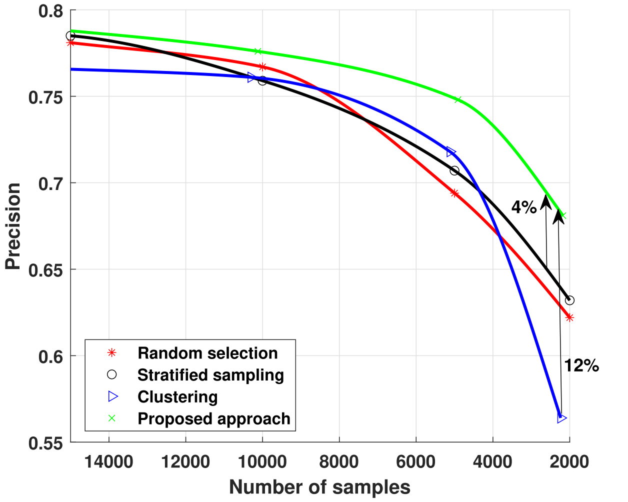


FIGURE 3. Precision versus the number of training samples with different optimization methods on the MiniImageNet dataset.

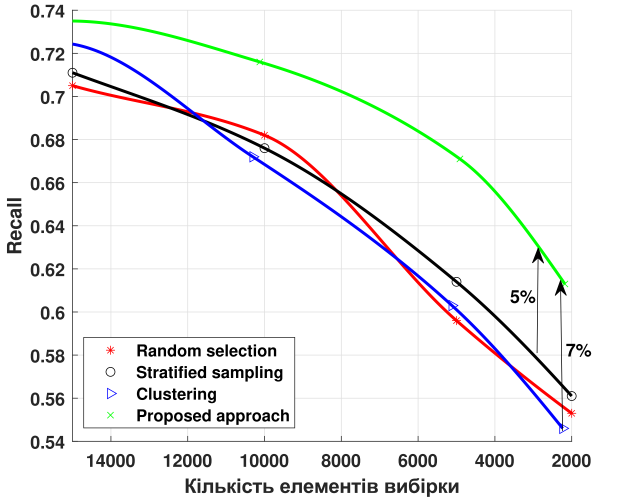


FIGURE 4. Recall versus the number of training samples with different optimization methods on the MiniImageNet dataset.

The results of accuracy on the MNIST dataset are presented in Fig. 5. Here, the random selection performs the worst, which is expected due to complete randomness. The strategic sampling and clustering perform almost identically, while the proposed quality-aware dataset optimization outperforms them by 2% for subsets of small size. The similarity of accuracy performance between the different methods is explained by the low number of classes in the MNIST dataset.

For a more precise assessment of the limitations of the proposed approach, numerical values of the accuracy, precision, recall, and F1-score are compared with other dataset optimization strategies. Corresponding results for the accuracy on the MNIST dataset for the different size limitations are presented in Table 2. Similarly, the results for precision, recall, and F1-score on the MiniImageNet dataset are presented in Tables 3, 4, and 5, respectively.

The noticeable advantage of our solution is clearly visible for the small dataset size (bold numbers), which highlights much better informativeness of selected samples.

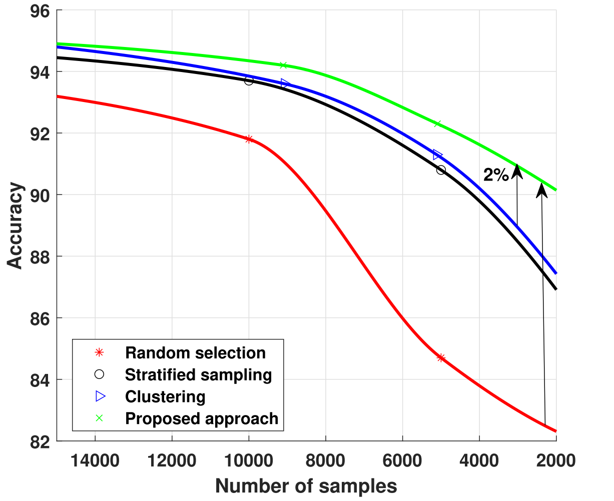


FIGURE 5. Accuracy versus the number of training samples with different optimization methods on the MNIST dataset.

TABLE 2. Aggregated results of the accuracy metric with different size limitations of the optimized MNIST dataset.

N	Random selection	Stratified sampling	Clustering	Proposed approach
2000	0.82	0.87	0.87	0.90
2500	0.83	0.88	0.88	0.91
3000	0.83	0.89	0.89	0.91
5000	0.85	0.90	0.90	0.92
15000	0.93	0.94	0.95	0.95
25000	0.95	0.95	0.96	0.95
30000	0.95	0.95	0.96	0.96
45000	0.96	0.97	0.97	0.97

TABLE 3. Aggregated results of the precision metrics with different size limitations of the optimized MiniImageNet dataset.

N	Random selection	Stratified sampling	Clustering	Proposed approach
2000	0.62	0.63	0.56	0.68
2500	0.64	0.65	0.58	0.69
3000	0.65	0.66	0.62	0.71
5000	0.69	0.71	0.72	0.75
15000	0.78	0.78	0.77	0.79
25000	0.80	0.80	0.78	0.81
30000	0.81	0.81	0.79	0.82
45000	0.82	0.83	0.82	0.83

With the further increase of the dataset size, the difference between different strategies diminishes, because the number of training samples becomes closer to the original MNIST or MiniImageNet (50,000 training samples).

To determine a balanced performance, the F1-score metric is provided, which represents a normalized tradeoff between precision and recall, useful for imbalanced datasets where some classes may be represented better than others. As observed from Table 5, the F1 score correlates with other metrics, indicating that the model features 15% more false negatives for the smallest training set compared to the largest training set.

For a deeper analysis of the performance, several additional metrics are calculated to provide more insights into the limitations and capabilities of the model in distinguishing

TABLE 4. Aggregated results of the recall metric with different size limitations of the optimized MiniImageNet dataset.

N	Random selection	Stratified sampling	Clustering	Proposed approach
2000	0.55	0.56	0.55	0.61
2500	0.55	0.57	0.56	0.62
3000	0.56	0.58	0.56	0.63
5000	0.60	0.61	0.60	0.67
15000	0.71	0.76	0.72	0.74
25000	0.75	0.95	0.72	0.74
30000	0.76	0.76	0.73	0.74
45000	0.77	0.77	0.74	0.76

TABLE 5. Aggregated results of the F1-score metric with different size limitations of the optimized MiniImageNet dataset.

N	Random selection	Stratified sampling	Clustering	Proposed approach
2000	0.58	0.59	0.55	0.64
2500	0.59	0.61	0.57	0.65
3000	0.60	0.62	0.59	0.67
5000	0.64	0.66	0.65	0.71
15000	0.74	0.77	0.74	0.76
25000	0.77	0.87	0.75	0.77
30000	0.78	0.78	0.76	0.78
45000	0.79	0.80	0.78	0.79

TABLE 6. Aggregated results of the different performance metrics for the proposed approach with different size limitations of the optimized MiniImageNet dataset.

N	TPR	TNR	PPV	NPV	FPR	FNR	FDR
2000	0.583	0.996	0.615	0.996	0.004	0.417	0.385
2500	0.644	0.996	0.665	0.996	0.004	0.356	0.335
3000	0.537	0.995	0.720	0.995	0.005	0.463	0.280
5000	0.662	0.997	0.749	0.997	0.003	0.338	0.250
15000	0.711	0.997	0.795	0.997	0.003	0.289	0.205
25000	0.767	0.997	0.804	0.998	0.002	0.233	0.196
30000	0.780	0.998	0.818	0.998	0.002	0.220	0.181
45000	0.784	0.998	0.835	0.998	0.002	0.216	0.164

TABLE 7. Aggregated results of the different performance metrics for the proposed approach with different size limitations of the optimized MiniImageNet dataset.

N	Top-5 Acc	Cohen's Kappa	AUROC	TFLOPS per epoch	Number of parameters
2000	0.838	0.578	0.978	1.36	4.4 M
2500	0.859	0.640	0.980	1.7	
3000	0.839	0.533	0.980	2.04	
5000	0.882	0.659	0.987	3.40	
15000	0.903	0.708	0.989	10.2	
25000	0.932	0.764	0.993	17.0	
30000	0.944	0.778	0.995	20.4	
45000	0.944	0.782	0.994	30.6	

images among the 100 classes of the MiniImageNet dataset (Table 6). These metrics include True Positive Rate (TPR), True Negative Rate (TNR), Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), and False Discovery Rate (FDR).

TPR and TNR offer insights into the model's ability to correctly identify true positives and true negatives, reflecting its sensitivity and specificity. PPV and NPV measure the precision of positive and negative predictions, respectively, providing a detailed view of the model's predictive accuracy. FPR and FNR, which are inversely related to TNR and TPR, highlight the proportion of misclassified negatives and positives, revealing potential areas for improvement in model precision and recall. Finally, FDR quantifies the proportion of false positives among all positive predictions, indicating the reliability of positive classifications.

Top-5 Accuracy is a widely accepted metric to evaluate performance in multi-class classification problems, especially when the number of classes is very large like in our case of MiniImageNet with 100 classes. In the original accuracy metric (i.e., Top-1 Accuracy), a 20% drop is observed between the largest and the smallest training sets, which is seen from the TPR and FNR indicators in Table 5. However, for Top-5 Accuracy, only a 10% drop is observed, which means that the majority of misclassified instances are still among the top 5 predicted classes. This result indicates that core essential informativeness is preserved even for a very small training set, and the trained model can narrow down the correct label to a small set of likely candidates, which is important in real-world applications.

A similar trend is observed with Cohen's Kappa indicator, which shows the agreement between the predicted and true labels while adjusting for the possibility of agreement occurring by chance, which is important in multi-class problems with class imbalance. The observed similarity between accuracy and Cohen's Kappa indicates that the proposed approach maintains a good balance between classes within the training set, regardless of the imposed size limitations.

In addition, the Receiver Operating Characteristic (ROC) is evaluated to estimate the overall ability of the model to discriminate among many classes. Since ROC is typically plotted for binary classification problems, it is computed using the one-vs-rest method in this case. To avoid plotting 100 curves, the Area Under the ROC Curve (AUROC) is calculated, providing a numerical understanding of the overall performance. The AUROC metric represents the area covered by the TPR vs. FPR curve, where AUROC $=$ 1 indicates an ideal case, and AUROC $= 0.5$ indicates the worst case. The average AUROC is computed over all possible combinations (one vs. rest) in the MiniImageNet dataset.

As seen from Table 7, the trained models show excellent performance in AUROC regardless of the training set size, indicating the ability to distinguish any class from all other classes. This result correlates with low FPR values in Table 5 and demonstrates the good informativeness of small, optimized training datasets.

Finally, the total computational complexity per epoch for model training on each optimized training set is estimated. For simplicity, the constant backpropagation part related to specific hyperparameters such as dropout rates and regularization is excluded. Instead, the complexity of the forward propagation during one epoch is calculated, which is directly proportional to the number of training samples.

As can be seen from the results in Tables 2-5, the proposed approach achieves the same performance for 2,000 samples as other studied solutions can achieve with approximately 5,000 samples. Thus, as seen from Table 7, the proposed approach requires 2 times fewer floating point operations (FLOPS) to get the same performance of the trained model, i.e., 1.36 TFLOPS vs 3.40 TFLOPS. Assuming training on commercial cloud infrastructure with a pay-per-use subscription, this advantage can be converted to noticeable cost savings in many practical applications.

V. DISCUSSION

The experimental results demonstrate the advantages of the proposed approach in its ability to reduce training data while preserving the maximum possible informativeness. The workflow of the proposed approach is applicable to both dimensionality reduction and data augmentation problems, significantly widening its applicability. The applicability of the proposed approach is indicated by the excellent performance in the discrimination between a very large number of classes, e.g., 100 and beyond. With this feature, the proposed solution can be a good candidate for usage in combination with other approaches or as a baseline for transfer learning in specific computer vision applications in various domains.

For problems where large volumes of training data are available, the training workflow can be significantly optimized by assessing data informativeness, improving overall cost-efficiency. This approach helps tackle the challenge of big and noisy datasets.

Another type of problem relates to the limited data challenge, observed in various specific domains such as healthcare, remote sensing, experimental physics, and chemistry. In these domains, the common solution is data augmentation. The proposed approach enables smarter data augmentation by ensuring that each augmented sample positively contributes to the overall informativeness of the dataset.

Finally, the findings in this paper can help to save costs for training on rented cloud infrastructure. This provides more flexibility to the developers of machine learning solutions, who will be able to afford more training capacity and develop more competitive products.

The possible future development of this work is in exploring the possibilities of replacing neural network-generated features with specific similarity metrics between two images, which could be designed to account not only color differences but also structural and textural aspects of the image. Here, the mean square error, peak-signal-to-noise-ratio, structural similarity index measure (SSIM), as well as their various derivatives, are of great interest. For example, SSIM is now widely used to assess how well two images match each other from a human perception perspective, as it uses brightness, contrast, and the structure of the image.

VI. CONCLUSION

In this paper, a novel approach for quality-aware dataset optimization through entropy analysis has been proposed. The distinguishing feature of this approach is in the combination of clustering and active learning. Such a workflow allows splitting the entire dataset into many subsets and analyzing the informativeness by calculating the individual entropy of each subset. The optimal dataset is composed by dynamic programming to find a sequence with the lowest overall entropy while ensuring balanced representativeness of all classes.

Experimental evaluation of the proposed dataset optimization solution on MNIST and MiniImageNet datasets proves its advantage over current best practices by 4% in precision, by 5% in recall, by 5% in F1-score, and by 2% in accuracy. Despite the marginal improvement, it is worth noting that the slope of performance characteristic curves is much lower and nearly constant compared to other studied alternatives. Moreover, the excellent AUROC results indicate that the proposed approach has very good discriminative capability even among a large number of classes. Therefore, in practice, the proposed approach is suitable for a wide range of real-world applications from data reduction in very large datasets to data augmentation in very small datasets.

Further research in this direction could include applications of various similarity metrics between images, which may indicate how much unique information is present in each sample of data to complement the entropy evaluation.

REFERENCES

[1] S. Wei, X. Lin, F. Tu, Y. Wang, L. Liu, and S. Yin, "Reconfigurability, why it matters in AI tasks processing: A survey of reconfigurable AI chips," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 70, no. 3, pp. 1228–1241, Mar. 2023.

[2] T. Goethals, B. Volckaert, and F. D. Turck, "Enabling and leveraging AI in the intelligent edge: A review of current trends and future directions," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 2311–2341, 2021.

[3] A. Banafa, *9 Narrow AI Vs. General AI Vs. Super AI*. Denmark, Europe: River, 2024, pp. 55–60.

[4] N. Xu, "The application of deep learning in image processing is studied based on the reel neural network model," *J. Phys., Conf. Ser.*, vol. 1881, no. 3, Apr. 2021, Art. no. 032096.

[5] B. Rusyn, R. Kosarevych, O. Lutsyk, and V. Korniy, "Segmentation of atmospheric clouds images obtained by remote sensing," in *Proc. 14th Int. Conf. Adv. Trends Radioelecrtronics, Telecommun. Comput. Eng. (TCSET)*, Feb. 2018, pp. 213–216.

[6] M. Bustreo, J. Cavazza, and V. Murino, "Enhancing visual embeddings through weakly supervised captioning for zero-shot learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1298–1307.

[7] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[8] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.

[9] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi, "GhostFaceNets: Lightweight face recognition model from cheap operations," *IEEE Access*, vol. 11, pp. 35429–35446, 2023.

[10] Y. LeCun and Y. Bengio, *Convolutional Networks for Images, Speech, and Time Series*, 1995, pp. 3361–3369.

[11] W.-W. Fan and C.-H. Lee, "Classification of imbalanced data using deep learning with adding noise," *J. Sensors*, vol. 2021, pp. 1–18, Nov. 2021.

[12] M. Iman, H. R. Arabnia, and K. Rasheed, "A review of deep transfer learning and recent advancements," *Technologies*, vol. 11, no. 2, p. 40, Mar. 2023.

[13] L. Wan, R. Liu, L. Sun, H. Nie, and X. Wang, "UAV swarm based radar signal sorting via multi-source data fusion: A deep transfer learning framework," *Inf. Fusion*, vol. 78, pp. 90–101, Feb. 2022.

[14] Q. Ferro, S. Graillat, T. Hilaire, F. Jezequel, and B. Lewandowski, "Neural network precision tuning using stochastic arithmetic," in *Proc. Int. Workshop Numer. Softw. Verification*, 2022, pp. 164–186.

[15] B. Shubyn and T. Maksymyuk, "Intelligent handover management in 5G mobile networks based on recurrent neural networks," in *Proc. 3rd Int. Conf. Adv. Inf. Commun. Technol. (AICT)*, Jul. 2019, pp. 348–351.

[16] S. Host, *Information Measures*. Piscataway, NJ, USA: IEEE, 2019, pp. 37–68.

[17] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, Mar. 2021.

[18] Y. Gong, G. Liu, Y. Xue, R. Li, and L. Meng, "A survey on dataset quality in machine learning," *Inf. Softw. Technol.*, vol. 162, Oct. 2023, Art. no. 107268.

[19] A. Ghorbani and J. Zou, "Data Shapley: Equitable valuation of data for machine learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2242–2251.

[20] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.

[21] A. Gorji, H.-U.-R. Khalid, A. Bourdoux, and H. Sahli, "On the generalization and reliability of single radar-based human activity recognition," *IEEE Access*, vol. 9, pp. 85334–85349, 2021.

[22] J. Wibbeke, P. Teimourzadeh Baboli, and S. Rohjans, "Optimal data reduction of training data in machine learning-based modelling: A multidimensional bin packing approach," *Energies*, vol. 15, no. 9, p. 3092, Apr. 2022.

[23] M. Manthiramoorthi, M. Mani, and A. Murthy, "Application of Pareto's principle on deep learning research output: A scientometric analysis," in *Proc. ICMLST*, 2021, pp. 1–10.

[24] R. Zhang, T. Du, and S. Qu, "A principal component analysis algorithm based on dimension reduction window," *IEEE Access*, vol. 6, pp. 63737–63747, 2018.

[25] A. Zare, A. Ozdemir, M. A. Iwen, and S. Aviyente, "Extension of PCA to higher order data structures: An introduction to tensors, tensor decompositions, and tensor PCA," *Proc. IEEE*, vol. 106, no. 8, pp. 1341–1358, Aug. 2018.

[26] R. M. Terol, A. R. Reina, S. Ziaei, and D. Gil, "A machine learning approach to reduce dimensional space in large datasets," *IEEE Access*, vol. 8, pp. 148181–148192, 2020.

[27] B. Du, Z. Wang, L. Zhang, L. Zhang, W. Liu, J. Shen, and D. Tao, "Exploring representativeness and informativeness for active learning," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 14–26, Jan. 2017.

[28] F. Saad, M. Cusumano-Towner, and V. Mansinghka, "Estimators of entropy and information via inference in probabilistic models," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2022, pp. 5604–5621.

[29] P. Juszczuk, J. Kozak, G. Dziczkowski, S. Głowania, T. Jach, and B. Probierz, "Real-world data difficulty estimation with the use of entropy," *Entropy*, vol. 23, no. 12, p. 1621, Dec. 2021.

[30] M. H. Nguyen, Y.-K. Lai, and K.-P. Chang, "An entropy-based DDoS attack detection and classification with hierarchical temporal memory," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2021, pp. 1942–1948.

[31] B. Eysenbach and S. Levine, "Maximum entropy Rl (provably) solves some robust Rl problems," in *Proc. 10th Int. Conf. Learn. Represent. ICLR*, 2022, pp. 1–12.

[32] A. Mishra, N. Raj, and G. Bajwa, "EEG-based image feature extraction for visual classification using deep learning," in *Proc. Int. Conf. Intell. Data Sci. Technol. Appl. (IDSTA)*, Sep. 2022, pp. 181–188.

[33] B. Rusyn, O. Lutsyk, R. Kosarevych, T. Maksymyuk, and J. Gazda, "Features extraction from multi-spectral remote sensing images based on multi-threshold binarization," *Sci. Rep.*, vol. 13, no. 1, p. 19655, Nov. 2023.

[34] R. Kosarevych, O. Lutsyk, B. Rusyn, O. Alokhina, T. Maksymyuk, and J. Gazda, "Spatial point patterns generation on remote sensing data using convolutional neural networks with further statistical analysis," *Sci. Rep.*, vol. 12, no. 1, p. 14341, Aug. 2022.

[35] L. Tang, S. Ma, X. Ma, and H. You, "Research on image matching of improved SIFT algorithm based on stability factor and feature descriptor simplification," *Appl. Sci.*, vol. 12, no. 17, p. 8448, Aug. 2022.

[36] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 172–176.

[37] S. A. Mohseni, H. R. Wu, J. A. Thom, and A. Bab-Hadiashar, "Recognizing induced emotions with only one feature: A novel color histogram-based system," *IEEE Access*, vol. 8, pp. 37173–37190, 2020.

[38] K.-C. Song, Y.-H. Yan, W.-H. Chen, and X. Zhang, "Research and perspective on local binary pattern," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 730–744, Mar. 2014.

[39] R. R. Isnanto, A. A. Zahra, A. L. Kurniawan, and I. P. Windasari, "Face recognition system using feature extraction method of 2-D Gabor wavelet filter bank and distance-based similarity measures," in *Proc. 7th Int. Conf. Informat. Comput. (ICIC)*, Dec. 2022, pp. 1–4.

[40] A. Durmusoglu and Y. Kahraman, "Face expression recognition using a combination of local binary patterns and local phase quantization," in *Proc. Int. Conf. Commun., Control Inf. Sci. (ICCISc)*, vol. 1, Jun. 2021, pp. 1–5.

[41] A. Gogna and A. Majumdar, "Discriminative autoencoder for feature extraction: Application to character recognition," *Neural Process. Lett.*, vol. 49, no. 3, pp. 1723–1735, Jun. 2019.

[42] A. Alem and S. Kumar, "End-to-end convolutional neural network feature extraction for remote sensed images classification," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022.

[43] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[44] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart, "Importance sampling: Intrinsic dimension and computational cost," *Stat. Sci.*, vol. 32, no. 3, pp. 405–431, Aug. 2017.

[45] Z. Zhao, Z. Zeng, K. Xu, C. Chen, and C. Guan, "DSAL: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 10, pp. 3744–3751, Oct. 2021.

[46] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.

[47] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*. Red Hook, NY, USA: Curran Associates, Dec. 2016, pp. 3637–3645.

[48] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

BOHDAN RUSYN received the D.Sc. degree in information technologies, in 1997. He is currently a Full Professor and the Head of the Remote Sensing Information Technologies Department, Karpenko Physico-Mechanical Institute, NAS of Ukraine. His research interests include computer vision, pattern recognition, 3D reconstruction, data compression, biometric identification, and digital signal processing. He serves as an Editor for *International Journal of Computing* and *Mathematical Modeling and Computing* Journal.

OLEKSIY LUTSYK received the Ph.D. degree in information technologies from the Karpenko Physico-Mechanical Institute, NAS of Ukraine, in 2014. He is currently a Senior Researcher with the Remote Sensing Information Technologies Department, Karpenko Physico-Mechanical Institute, NAS of Ukraine. His research interests include computer vision, pattern recognition, 3D reconstruction, and deep learning.

ROSTYSLAV KOSAREVYCH received the D.Sc. degree in information technologies from the Karpenko Physico-Mechanical Institute, NAS of Ukraine, in 2017. He is currently the Head of the Remote Sensing Laboratory, Karpenko Physico-Mechanical Institute, NAS of Ukraine. His research interests include computer vision, binary point patterns, and deep learning.

OLEG KAPSHII received the Ph.D. degree in mathematical modeling and numerical methods from Lviv Polytechnic National University, in 2006. He is currently a Lead Principal Engineer with the Advanced System Research Group, Infineon Technologies. His research interests include signal processing, embedded systems, wireless communication technologies, and machine learning.

OLEKSANDR KARPIN received the Ph.D. degree in mathematical modeling and numerical methods from Lviv Polytechnic National University, in 2007. He is currently a Senior Principal Engineer with the Advanced System Research Group, Infineon Technologies. He is also an Associate Professor with the Sensor and Semiconductor Electronics Department, Ivan Franko National University of Lviv. His research interests include signal processing, capacitive sensing, machine learning, and embedded systems.

TARAS MAKSYMYUK (Member, IEEE) received the Ph.D. degree in telecommunication systems and networks from Lviv Polytechnic National University, Lviv, Ukraine, in 2015. He is currently an Associate Professor with the Telecommunications Department, Lviv Polytechnic National University. He completed a Postdoctoral Fellowship with the Internet of Things and Artificial Intelligence Laboratory, Korea University. He is also acting as a Senior Systems Engineer with the Advanced System Research Group, Infineon Technologies. His research interests include 5G/6G mobile networks, the Internet of Things, computer vision, and artificial intelligence. He serves as an Editor for the Internet of Things Series in *IEEE Communications Magazine* and *Wireless Communications and Mobile Computing*.

JURAJ GAZDA was a Guest Researcher with Ramon Llull University, Barcelona, and the Technical University of Hamburg-Harburg. He was involved in the development of Nokia Siemens Networks (NSN). In 2017, he was recognized as a Best Young Scientist with TUKE. He is currently a Full Professor with the Faculty of Electrical Engineering, Technical University of Koåice (TUKE), Slovakia. His research interests include spectrum pricing, techno-economic aspects of 5G networks, coexistence of HetNets, and machine learning in 5G networks. He serves as an Editor for *KSII Transactions on Internet and Information Systems* and a Guest Editor for *Wireless Communications and Mobile Computing* (Wiley).