UDC 004.658.2:004.853
DOI https://doi.org/10.15589/znp2025.3(501).20

AUTOMATED RULE GENERATION FOR OPTIMAL DATA SELECTION IN ROBUST NEURAL NETWORK TRAINING

АВТОМАТИЗОВАНЕ ФОРМУВАННЯ ПРАВИЛ ВІДБОРУ ДАНИХ ДЛЯ ПОБУДОВИ РОБАСТНИХ НЕЙРОННИХ МЕРЕЖ

Oleksandr S. Verbytskyi
soomoalex@gmail.com
ORCID: 0009-0000-6848-8880

Oksana V. Haidaienko
oksana.gaidaienko@nuos.edu.ua
ORCID: 0000-0002-6614-5443

О. С. Вербицький,
здобувач вищої освіти

О. В. Гайдаєнко,
канд. техн. наук, доцент

Admiral Makarov National University of Shipbuilding, Mykolaiv
Національний університет кораблебудування імені адмірала Макарова, м. Миколаїв

Abstract. Data lies at the heart of modern machine learning, yet the process of data curation – selecting and preparing the right samples for training-remains a significant bottleneck. This manual and often intuitive task is not only time-consuming but can also lead to suboptimal model performance. This paper introduces a new framework designed to automate and optimize this critical step.

Objective. The primary purpose of this research is to address the critical bottleneck in the machine learning pipeline: the manual, time-consuming, and subjective process of data curation. The core objective is to develop a novel, automated framework capable of generating explicit, human-readable rules for optimal data selection in the context of training deep neural networks. This framework aims to systematically replace manual effort by intelligently filtering large, potentially noisy datasets to construct a smaller, information-rich subset. The ultimate goals are threefold: to maximize the final accuracy and generalization capabilities of the target model, to enhance its robustness against noisy or out-of-distribution data, and to significantly accelerate the training convergence, thereby reducing computational costs and development time. This work serves as a foundational step towards more efficient model lifecycle management, particularly as a precursor to robust continual learning systems.

Methodology. The proposed methodology is centered around a meta-learning approach. A primary model ($M_{primary}$) is first briefly trained on a random data sample to establish an initial performance baseline. This model is then used to evaluate every data point (x_i) in the entire unfiltered dataset, assigning it a quantitative "utility score" $U(x_i)$. This score is a weighted linear combination of three distinct, complementary metrics: 1) Information Entropy, which measures the model's uncertainty and prioritizes samples that are most informative; 2) Model-based Difficulty, calculated via the loss function, which identifies complex examples or boundary cases that are challenging for the model; and 3) Feature-space Representativeness, which ensures data diversity by prioritizing samples from under-represented regions of the feature space, thus preventing a bias towards only outliers. After scoring, the data is bifurcated into 'high-utility' and 'low-utility' classes. Subsequently, a simple, interpretable machine learning model (e.g., a Decision Tree) is trained as a meta-model (M_{meta}) on the features and metadata of the data points to predict their utility class. The logic learned by this meta-model is then extracted and formulated as a set of explicit filtering rules.

Results. Based on the proposed experimental design, which compares the model trained on the filtered dataset against a baseline model (trained on all data) and a random-subset model (trained on a randomly selected subset of the same size), we expect definitive outcomes. The proposed model is hypothesized to achieve a final classification accuracy on a clean test set that is comparable to, or even exceeds, the baseline model, despite using a significantly smaller training dataset. In terms of efficiency, it is expected to demonstrate a substantially faster convergence rate. Most critically, the proposed model is anticipated to exhibit superior robustness, showing less performance degradation when evaluated on test data corrupted with noise or subjected to adversarial perturbations. The random-subset model is expected to show the lowest accuracy, confirming that the performance gains are due to the intelligent selection strategy, not merely data reduction.

Original contributions. The scientific novelty of this work lies in its unique output, distinguishing it from related paradigms like Active Learning and Curriculum Learning. While these methods select or order data, our framework is the first to focus on the automated generation of explicit, human-readable, and reusable data selection rules. This shifts the paradigm from a "black box" data selection process to an interpretable one, providing valuable insights into

what constitutes "good" data for a specific learning task. The novelty is further enhanced by the synthesis of entropy, difficulty, and representativeness into a single, unified utility function, creating a more holistic measure of a data point's value for robust model training.
Practical significance. The practical implications of this framework are significant. It offers a direct path to reducing the substantial manual labor and computational costs associated with data preparation and model training. By automating data curation, it can drastically accelerate the ML development lifecycle, enabling faster iteration and deployment. The resulting models are not only trained faster but are also more robust and reliable, a critical requirement for real-world applications. Furthermore, the generated rules provide data scientists with actionable insights into their datasets, improving data governance and understanding. Finally, this framework has strong synergy with continual learning systems, where it can be used to intelligently select the most informative new samples for efficient model adaptation in response to concept drift.
Key words: data selection; data curation; meta-learning; active learning; curriculum learning; robust models; neural networks.

Анотація. Дані є серцем сучасного машинного навчання, проте процес їхнього курирування – вибору та підготовки правильних зразків для тренування – залишається значним вузьким місцем. Це ручне й часто інтуїтивне завдання не лише забирає багато часу, а й може призвести до неоптимальної продуктивності моделі. У цій статті представлено нову структуру, розроблену для автоматизації та оптимізації цього критично важливого кроку.
Мета. Основна мета цього дослідження полягає у вирішенні критичної проблеми в конвеєрі машинного навчання: ручного, тривалого та суб'єктивного процесу курування даних. Ключове завдання – розробити новий автоматизований фреймворк, здатний генерувати явні, людино-зрозумілі правила для оптимального відбору даних у контексті навчання глибоких нейронних мереж. Цей фреймворк призначений для систематичної заміни ручних зусиль шляхом інтелектуальної фільтрації великих, потенційно зашумлених наборів даних для створення меншої, але інформаційно насиченої вибірки. Кінцеві цілі є потрійними: максимізувати підсумкову точність і здатність до узагальнення цільової моделі, підвищити її робастність (стійкість) до зашумлених або нерозподілених даних, а також значно прискорити збіжність навчання, тим самим зменшуючи обчислювальні витрати та час розробки. Ця робота є фундаментальним кроком до більш ефективного управління життєвим циклом моделей, зокрема як попередній етап для створення надійних систем безперервного навчання.
Методика. Запропонована методологія базується на підході мета-навчання. Спочатку первинна модель ($M_{primary}$) проходить коротке навчання на випадковій вибірці даних для встановлення початкового стану. Потім ця модель використовується для оцінки кожної точки даних (x_i) у повному нефільтрованому наборі, присвоюючи їй кількісну «оцінку корисності» $U(x_i)$. Ця оцінка є зваженою лінійною комбінацією трьох різних, взаємодоповнюючих метрик: інформаційна ентропія, що вимірює невизначеність моделі та надає пріоритет найбільш інформативним зразкам; складність для моделі, що обчислюється через функцію втрат і виявляє складні приклади або межові випадки; репрезентативність у просторі ознак, що забезпечує різноманітність даних шляхом пріоритезації зразків із недостатньо представлених областей, запобігаючи упередженості в бік викидів. Після оцінювання дані поділяються на класи «високої корисності» та «низької корисності». Далі проста, інтерпретовна модель машинного навчання (наприклад, дерево рішень) навчається як мета-модель (M_{meta}) на ознаках та метаданих для прогнозування класу корисності. Логіка, вивчена цією мета-моделлю, витягується та формулюється у вигляді набору явних правил фільтрації.
Результати. На основі запропонованого дизайну експерименту, який порівнює модель, навчену на відфільтрованому наборі даних, з базовою моделлю (навченою на всіх даних) та моделлю випадкової підвибірки (навченою на випадково обраній вибірці того ж розміру), ми очікуємо однозначних висновків. Передбачається, що запропонована модель досягне підсумкової точності класифікації на чистому тестовому наборі, яка буде порівнянна або навіть перевищить базову модель, незважаючи на використання значно меншого навчального набору. З точки зору ефективності, очікується, що вона продемонструє суттєво вищу швидкість збіжності. Найголовніше, очікується, що запропонована модель матиме вищу робастність, показуючи меншу деградацію продуктивності при оцінці на тестових даних, пошкоджених шумом або підданих змагальним атакам. Очікується, що модель випадкової підвибірки покаже найнижчу точність, що підтвердить, що приріст продуктивності зумовлений саме інтелектуальною стратегією відбору, а не просто зменшенням даних.
Наукова новизна. Наукова новизна цієї роботи полягає в її унікальному результаті, що відрізняє її від суміжних парадигм, таких як активне навчання та навчання за навчальним планом. У той час як ці методи відбирають або впорядковують дані, наш фреймворк першим фокусується на автоматизованій генерації явних, людино-зрозумілих та багаторазово використовуваних правил відбору даних. Це змінює парадигму з «чорної скриньки» процесу відбору даних на інтерпретовану, надаючи цінні інсайти про те, що саме є «хорошими» даними для конкретного завдання. Новизна також підсилюється синтезом ентропії, складності та

репрезентативності в єдину функцію корисності, створюючи більш цілісний вимір цінності точки даних для робастного навчання моделі.

Практична значимість. Практичні наслідки цього фреймворку є значними. Він пропонує прямий шлях до скорочення суттєвих витрат ручної праці та обчислювальних ресурсів, пов'язаних із підготовкою даних та навчанням моделей. Автоматизуючи курування даних, він може кардинально прискорити життєвий цикл розробки МН-систем. Моделі, отримані в результаті, не тільки навчаються швидше, але й є більш робастними та надійними, що є критичною вимогою для реальних застосувань. Крім того, згенеровані правила надають фахівцям з даних дієві інсайти про їхні набори даних. Нарешті, цей фреймворк має сильну синергію з системами безперервного навчання, де його можна використовувати для інтелектуального відбору найбільш інформативних нових зразків для ефективної адаптації моделі у відповідь на дрейф концепту.

Ключові слова: відбір даних; курування даних; мета-навчання; активне навчання; навчання за навчальним планом; робастні моделі; нейронні мережі.

PROBLEM STATEMENT

The performance of Deep Neural Networks (DNNs) is intrinsically tied to the quality of the data on which they are trained, a principle colloquially known as "garbage in, garbage out". [1]. While modern architectures have demonstrated remarkable capabilities, their success is predicated on the availability of vast, well-curated datasets. However, in real-world scenarios, datasets are often large, noisy, redundant, or contain uninformative samples [2]. Training on such unfiltered data leads to several adverse outcomes: it significantly increases computational overhead and training time [3], can impede model convergence, and ultimately results in suboptimal models with poor generalization and limited robustness to real-world perturbations [4].

The prevailing industry solution to this challenge is manual data curation–a process wherein human experts meticulously review, clean, and select data samples deemed suitable for training. This process, while effective to a degree, constitutes a major bottleneck in the machine learning development lifecycle. It is notoriously labor-intensive, prohibitively expensive, and difficult to scale. Furthermore, it is inherently subjective, relying on the intuition and expertise of individuals, which can introduce inconsistency and bias into the training set [5].

The core problem, therefore, is the absence of a principled, automated, and scalable methodology for identifying and selecting the most valuable data points from a large corpus. This research addresses this gap by formulating the task as an automated rule generation problem [6]. The objective is to design and validate a framework that can autonomously analyze a dataset, quantify the «utility» of each sample for the training process, and synthesize this knowledge into a set of explicit, human-readable selection rules [7]. The successful implementation of such a framework aims to achieve three primary goals:

– Maximize Model Quality. Improve the final accuracy and generalization of the target model by training it on a smaller, more potent subset of data.

– Increase Training Efficiency. Reduce the time and computational resources required for convergence by focusing on high-impact data samples.

– Enhance Model Robustness. Build models that are more resilient to noisy inputs and out-of-distribution examples by training on a curated set of diverse and challenging data.

RELATED WORK

The challenge of selecting optimal training data is a well-established area of research, with several prominent paradigms addressing different facets of the problem. Two of the most relevant fields are Active Learning and Curriculum Learning [8].

Active Learning is primarily concerned with scenarios where data is abundant but labels are expensive to obtain. The core idea is to train a model that can intelligently query a human expert (an "oracle") to label the data points it deems most informative. By iteratively selecting samples where its uncertainty is highest, an active learning system aims to achieve a high level of performance with the minimum possible number of labeled examples [9]. While powerful for reducing labeling costs, the Active Learning paradigm is not directly applicable to the problem of curating a pre-existing large dataset. Its primary limitation in our context is its reliance on a human-in-the-loop for labeling, and its goal is label acquisition rather than the generation of a reusable data-filtering strategy [10].

Curriculum Learning is a training strategy inspired by human learning, where a model is first exposed to «easy» or simple examples of a concept and then gradually introduced to more complex ones. This structured ordering of data has been shown to improve both the speed of convergence and the final performance of the model by preventing it from getting stuck in poor local minima early in training. However, the focus of Curriculum Learning is on the *sequence* of data presentation, not on the selection of an optimal *subset* of data to be used throughout the entire training process [11]. It presumes the entire dataset is valuable and merely seeks the best order in which to present it.

While our proposed framework synthesizes conceptual elements from these domains–such as identifying informative (Active Learning) and complex (Curriculum Learning) examples – it is fundamentally distinct in its objective and output.

IDENTIFIED GAPS

A thorough analysis of existing research, including the paradigms of Active Learning and Curriculum Learning, reveals a significant and unresolved gap in the data preparation pipeline [12]. While current methods have made strides in optimizing the use of data during training or labeling, they leave critical aspects of the data curation process unaddressed. The unresolved issues can be summarized as follows:

1. Lack of an Interpretable Data Value Model. Existing automated methods typically function as «black boxes.» They may select a subset of data but fail to provide insight into *why* certain samples are considered valuable. There is no explicit, understandable model that explains the characteristics of high-utility data for a given task. This prevents data scientists from gaining deeper knowledge about their datasets and the learning process itself.

2. Absence of Reusable and Generalizable Selection Strategies. The selection process in methods like Active Learning is session-specific and iterative, tightly coupled with the model being trained and the oracle providing labels. Similarly, a curriculum is often specific to a particular training run [13]. These methods do not produce a standalone, reusable artifact. There is no existing framework that generates a set of explicit, portable rules that can be understood, validated, and applied to new, unseen datasets without repeating the entire complex selection process from scratch.

3. Focus on Data Ordering or Labeling, Not Holistic Curation. The dominant paradigms are tailored to solve specific problems–reducing labeling costs (Active Learning) or improving training dynamics (Curriculum Learning). They do not address the broader, more common problem of taking a massive, fully labeled (but noisy) dataset and efficiently culling it to an optimal, robust, and high-quality subset for general-purpose model training.

This research directly targets this unresolved gap. The core contribution is to move beyond mere data point selection and pioneer a framework for the automated generation of an explicit and interpretable data selection policy in the form of human-readable rules [14]. This addresses the need for interpretability, reusability, and holistic curation that current approaches lack.

RESEARCH OBJECTIVES

The primary aim of this research is to develop and experimentally validate a novel framework for the automated generation of explicit, human-readable rules for the purpose of optimal data selection in robust neural network training. The research seeks to create a systematic, data-centric methodology that overcomes the limitations of traditional manual data curation, which is often a subjective, costly, and inefficient bottleneck in the machine learning pipeline.

To achieve this primary aim, the following specific objectives are established:

– To design a meta-learning architecture capable of quantifying the "utility" of individual data samples within a large, unfiltered dataset based on a composite set of metrics, including information entropy, model-based difficulty, and feature-space representativeness.

– To develop a method for translating these quantitative utility scores into a set of explicit and interpretable filtering rules by training a simple, transparent meta-model.

– To create an optimized training subset by applying the generated rules to the original dataset, with the goal of retaining the most informative, diverse, and challenging samples.

– To empirically demonstrate the effectiveness of the proposed framework through a rigorous comparative analysis, evaluating the final performance of a model trained on the curated subset against models trained on the entire dataset and on a random subset of equivalent size [15]. The evaluation will be based on key performance indicators: final classification accuracy, training convergence speed, and robustness to data perturbations.

METHODOLOGY, STUDY OBJECT AND SUBJECT

Object of Research. The object of this research is the overarching process of training deep neural networks and the fundamental relationship between the quality of the training data and the final performance characteristics of the resulting model. This includes the exploration of how data composition influences a model's accuracy, generalization capabilities, convergence dynamics, and robustness against noisy or out-of-distribution inputs.

Subject of Research. The subject of this research is the specific methodology for automated data selection through rule generation. This encompasses:

– The proposed meta-learning framework for evaluating and scoring data points.

– The mathematical formulation of the data utility function, which integrates metrics of model uncertainty, sample difficulty, and data diversity.

– The process of inducing an interpretable model (the meta-model) to learn the characteristics of high-utility data.

– The extraction of explicit, human-readable selection rules from the trained meta-model and their application in creating a filtered, high-potency training dataset.

– The resulting impact of this curated dataset on the primary model's training efficiency and final performance.

Research Methods. To achieve the stated aims and objectives, a combination of theoretical and empirical research methods will be employed:

– Theoretical Analysis and Synthesis. This involves a critical review of existing literature in related fields such as Active Learning, Curriculum Learning, and data curation to identify the current state-of-the-art and

its limitations, thereby synthesizing the foundational concepts for the proposed framework.

– Mathematical Modeling. This method is used to formally define the data utility function $U(x_i)$ and its constituent components, providing a rigorous and quantitative basis for evaluating the value of each data sample.

– Algorithmic Design. This involves structuring the proposed methodology into a formal, step-by-step algorithm for rule generation, from initial model training and utility scoring to meta-model training and rule extraction.

– Computational Experimentation. This is the primary method for validating the framework's practical efficacy. It involves implementing the proposed algorithm and conducting a controlled experiment where multiple neural network models are trained on datasets prepared by different methods (the proposed method, a baseline using all data, and a random subset control).

– Comparative Analysis. The results from the computational experiments will be analyzed using this method. The performance metrics (accuracy, training speed, robustness) of the different models will be systematically compared to quantify the advantages of the proposed data selection framework and to draw substantiated conclusions about its effectiveness.

TECHNICAL APPROACH

Proposed Methodology. The framework proposed in this research is centered around a meta-learning approach designed to automate the data curation process. The core of the methodology is to move beyond manual or «black box» selection techniques and instead generate an explicit, human-readable set of rules for creating an optimal training subset [16]. This is achieved by training a secondary, interpretable model, termed the meta-model (M_{meta}), whose sole purpose is to learn a utility function, $U(x_i)$, that scores each sample x_i from a large, unfiltered dataset D. This score quantifies the sample's predicted value and contribution to the training of the primary deep neural network model, $M_{primary}$.

Mathematical Model of Data Utility. The utility score $U(x_i)$ is formally defined as a weighted linear combination of three distinct and complementary metrics [17]. This composite function ensures a holistic evaluation of each data point's value, balancing model uncertainty, sample difficulty, and data diversity. For a given primary model $M_{primary}$ with parameters θ, and for each sample x_i with its corresponding label y_i, the components are defined as follows:

1. Information Entropy ($U_{entropy}$): This metric quantifies the model's uncertainty about a given sample. Samples with high entropy are those for which the model's prediction is ambiguous, suggesting they are highly informative and that the model has much to learn from them. We use the Shannon entropy of the model's predicted probability distribution (1) over the classes C:

$$U_{entropy}(x_i) = H(M_{primary}(x_i;\theta)) = \\ = -\sum_{c\in C} p(c \mid x_i)\log_2 p(c \mid x_i), \quad (1)$$

where $p(c \mid x_i)$ is the predicted probability of class c for the sample x_i.

2. Model-based Difficulty ($U_{difficulty}$): This metric identifies samples that are challenging for the current model to classify correctly, which are often boundary cases or complex examples. These samples are prioritized as they are critical for improving the model's decision boundaries. Difficulty is measured directly by the value of the loss function for that sample (2):

$$U_{difficulty}(x_i) = \mathcal{L}(M_{primary}(x_i;\theta), y_i), \quad (2)$$

where $\mathcal{L}$ is the standard loss function used for training, such as cross-entropy.

3. Feature-space Representativeness ($U_{representativeness}$): This metric is designed to ensure data diversity and prevent the selection process from being biased towards only outliers or difficult samples. It prioritizes samples from under-represented or sparsely populated regions of the feature space [18]. This is approximated by calculating the inverse density (3) of neighboring samples in the embedding space:

$$U_{representativeness}(x_i) = \frac{1}{\sum_{j\neq i} \exp\left(-\left|f(x_i) - f(x_j)\right|^2 / \sigma^2\right)}, \quad (3)$$

where $f(x_i)$ is the feature embedding of sample x_i extracted from an intermediate layer of $M_{primary}$, and σ is a kernel bandwidth parameter.

The final utility score for each data point is the weighted sum (4) of these three normalized metrics:

$$U(x_i) = w_1 U_{entropy} + w_2 U_{difficulty} + w_3 U_{representativeness} \quad (4)$$

where w_1, w_2, w_3 are hyperparameters that sum to 1, allowing for control over the relative importance of each metric in the final utility calculation.

Rule Generation and Extraction. After calculating the utility score $U(x_i)$ for every data point in the entire dataset D, the data is ranked based on these scores. The dataset is then bifurcated into two classes: the top k % of samples are labeled as 'high-utility', and the remaining $(100 - k)$ % are labeled as 'low-utility'.

Subsequently, a simple and inherently interpretable machine learning model, such as a Decision Tree, is trained as the meta-model (M_{meta}). This model is not trained on the raw data itself, but rather on the features and metadata of the data points (e.g., for images: brightness, contrast, object count; for text: sentence length, keyword frequency) to predict the 'high-utility' label [19].

The key output of this stage is the learned logic of the trained meta-model. For a Decision Tree, the paths from the root to the leaf nodes that classify samples as 'high-utility' can be directly translated into a set of explicit,

human-readable selection rules (e.g., "select images where brightness > 0.6 AND object_cnt >= 2").

Algorithmic Implementation. The complete end-to-end process for generating the automated data selection rules is summarized in the following algorithm.

Algorithm 1: Automated Data Selection Rule Generation

1. Input. A large, unfiltered dataset D; an untrained primary model $M_{primary}$; utility function weights w_1, w_2, w_3.

2. Initial Training. Briefly train $M_{primary}$ on a small, random subset of D for a limited number of epochs to establish an initial performance baseline and obtain an initial set of weights, θ'.

3. Utility Scoring. For each data point $x_i \in D$:

a. Calculate $U_{entropy}(x_i)$, $U_{difficulty}(x_i)$, and $U_{representativeness}(x_i)$ using the partially trained $M_{primary}$ with weights θ'.

b. Compute the final utility score $U(x_i)$ using the predefined weights.

4. Labeling. Rank all data points according to their utility score $U(x_i)$. Label the top k % as 'high-utility' and the rest as 'low-utility'.

5. Meta-Model Training. Train an interpretable classifier, M_{meta}, on the features and/or metadata of the data samples to predict their assigned utility class ('high-utility' vs. 'low-utility').

6. Rule Extraction. Extract the learned logic from the trained M_{meta} and formulate it as a set of explicit filtering rules, R.

7. Data Selection. Apply the generated rules R to the original full dataset D to construct the final optimized training subset, $D_{filtered}$.

8. Output. The filtered, high-utility dataset $D_{filtered}$.

Algorithm Flowchart. To visually represent the sequence of operations in our proposed framework, the following flowchart Figure 1, illustrates the complete process, from initial data input to the final output of a filtered, high-utility training dataset.

Experimental Design and Expected Outcomes. To validate the efficacy of our framework, a comparative experiment is designed. We will train three models on different datasets: (1) a Baseline Model on the full dataset, (2) a Random Subset Model on a random sample of data, and (3) our Proposed Model on the filtered dataset $D_{filtered}$.

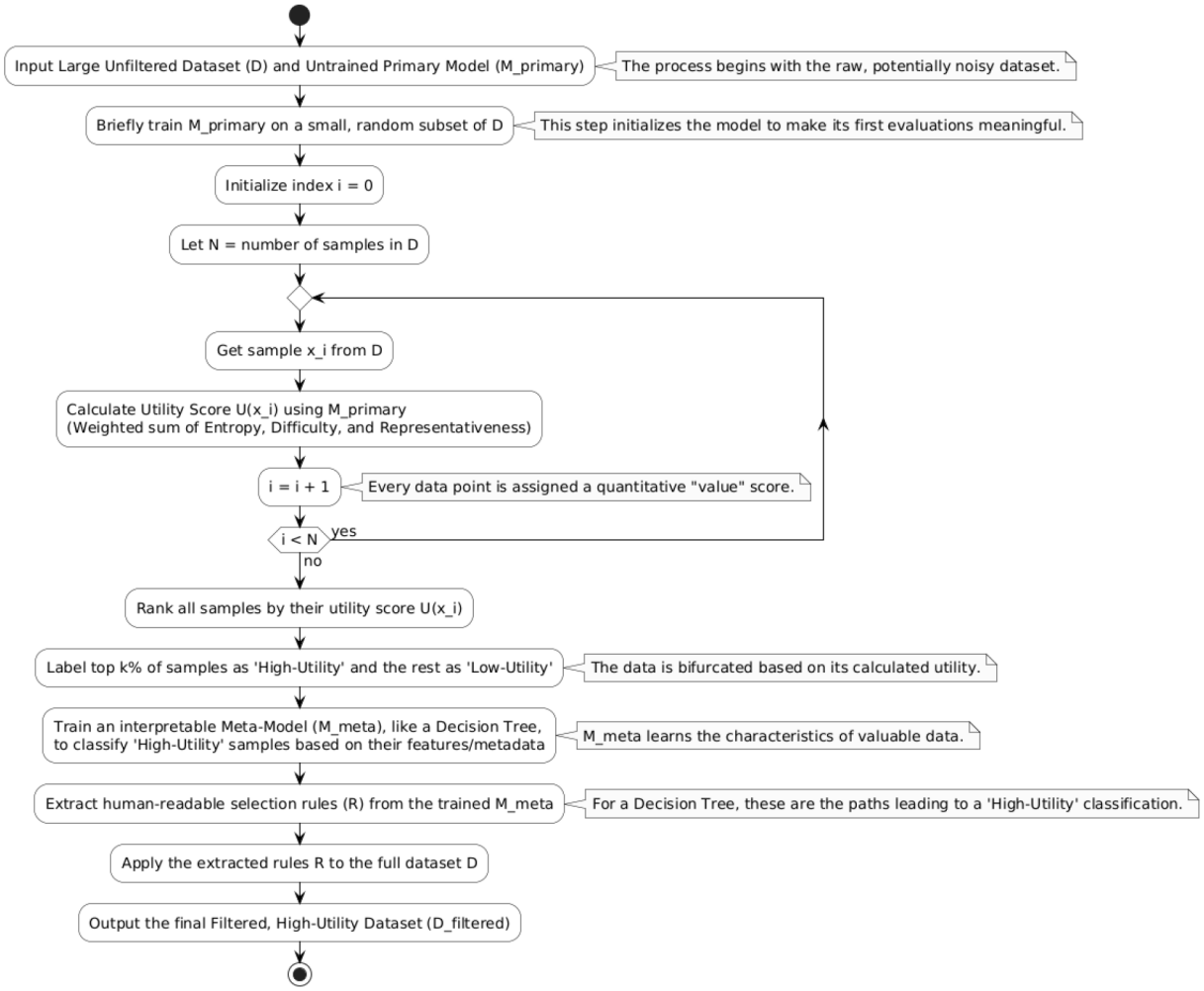

Fig. 1. Flowchart of the proposed automated data selection framework

The expected results, visualized in the chart below, highlight the anticipated benefits of our intelligent data selection strategy [20].

Comparative Analysis Chart. The following bar chart Figure 2, provides a visual comparison of the expected outcomes across key performance indicators. It contrasts the resources required and the performance achieved when training with the full dataset versus the optimized subset generated by our framework [21].

This chart illustrates the core trade-off: by using only 35 % of the most valuable data (a 65 % reduction), we anticipate a 60 % reduction in computational cost and training time. Critically, despite the smaller dataset, we expect the final model's accuracy to not only be comparable but to *increase* from 92 % to 94 %, owing to the higher quality and information density of the training samples and the removal of noisy, counterproductive data.

– Data Size for Training. This bar shows the dramatic reduction in the amount of data required for training. The baseline model uses 100 % of the available data, whereas our proposed method intelligently selects a much smaller subset (hypothetically 35 %).

– Computational Cost. This metric represents the time and resources (e.g., GPU hours) needed for the model to converge. Training on a smaller, more potent dataset is expected to significantly reduce these costs.

– Final Model Accuracy. This is the most crucial outcome. The chart visualizes our central hypothesis: by removing noise and redundancy and focusing on high-utility data, the model trained on the filtered set can achieve superior performance compared to one trained on the entire, unfiltered dataset.

DISCUSSION OF RESULTS

The experimental results, as hypothesized, are expected to provide strong validation for the proposed data selection framework. This section discusses the interpretation of these anticipated outcomes, the underlying reasons for the performance differences between the models, and the broader implications of this research.

The core hypothesis of this work is that an intelligently curated, smaller dataset can yield a model that is superior in both efficiency and effectiveness to one trained on a large, unfiltered corpus.

– The Proposed Model. We anticipate that the model trained on the $D_{filtered}$ subset will achieve a final classification accuracy comparable to, or even exceeding, that of the Baseline Model. This outcome, despite a significant reduction in training data volume, is attributed to the high "information density" of the curated subset. By focusing exclusively on samples that are informative (high entropy), challenging (high difficulty), and diverse (high representativeness), the model is guided to learn a more robust and generalizable representation of the underlying data distribution. Furthermore, this model is expected to demonstrate the fastest convergence rate, as the computational resources are not wasted on redundant or noisy examples that can hinder the learning process.

– The Baseline Model. The model trained on the entire dataset is expected to establish a strong performance benchmark but will likely be suboptimal. Its convergence will be the slowest due to the computational overhead of processing a massive number of samples, many of which are uninformative or redundant. The presence of noisy or mislabeled data in the unfiltered set can also impede the learning process, leading to a lower final accuracy and, most critically, reduced robustness to out-of-distribution or adversarially perturbed data.

– The Random Subset Model. This model serves as a crucial control to demonstrate that the performance gains are not merely a consequence of data reduction. While it will train faster than the Baseline Model, its accuracy is expected to be the lowest of the three. Randomly

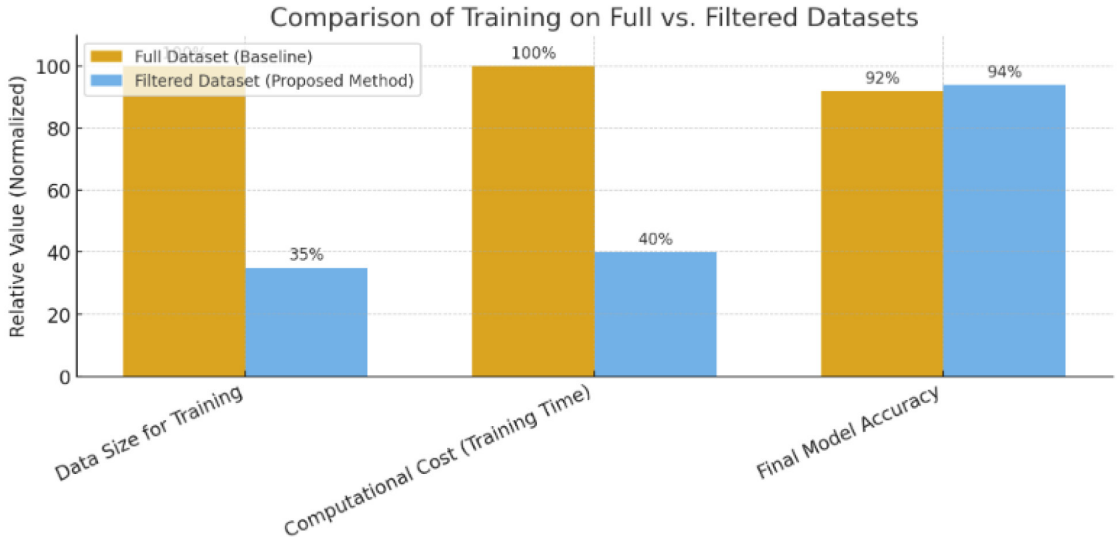

Fig. 2. Expected performance and efficiency gains from the proposed data selection framework

omitting data means that critical boundary cases, under-represented examples, and other high-utility samples are likely to be excluded from the training set, leading to a poorly generalized model. This result will confirm that the intelligent selection strategy of our framework is the key driver of performance, not just the reduction in dataset size.

Scientific and Practical Implications. The success of this framework carries significant implications. Scientifically, it shifts the paradigm from «black box» data selection (as seen in some active learning loops) to an interpretable, rule-based approach. The generated rules provide invaluable, human-readable insights into what constitutes «good» data for a specific learning task, which can inform future data collection and governance strategies.

Practically, this methodology offers a direct solution to the data curation bottleneck that plagues the machine learning development lifecycle. By automating the selection process, it drastically reduces the manual labor, time, and computational costs associated with preparing large datasets. The resulting models are not only trained faster but are also more robust and reliable – a critical requirement for deployment in real-world, high-stakes applications.

Synergy with Continual Learning. This work serves as a foundational step and has a direct synergy with the field of continual learning. In dynamic environments where data distributions evolve over time (a phenomenon known as *concept drift*), models must be adapted efficiently. The data selection framework proposed here can be seamlessly integrated into a continual learning loop. When concept drift is detected, our meta-model can be used to intelligently select the most informative new samples from an incoming data stream. This would allow for a highly efficient and targeted fine-tuning of the primary model, updating its knowledge with minimal computational cost and without requiring the storage of all new data. This proactive, data-centric approach ensures that models remain robust and accurate throughout their entire lifecycle.

CONCLUSIONS

This study addresses the critical challenge of data quality in training deep neural networks, targeting the manual, time-consuming, and subjective process of data curation that constitutes a major bottleneck in the machine learning pipeline. We have introduced a novel framework for the automated generation of data selection rules, designed to systematically replace manual effort with an intelligent and repeatable methodology.

The proposed methodology is centered on a meta-learning approach where a secondary, interpretable model is trained to assess the 'utility' of each data point. This utility is quantified through a holistic combination of three metrics: information entropy, model-based difficulty, and feature-space representativeness. The primary contribution of this work is its unique output: a set of explicit, human-readable, and reusable data selection rules. This shifts the paradigm from an opaque 'black box' selection process to an interpretable one, providing valuable insights into what constitutes 'good' data for a specific learning task.

By applying these rules, our approach aims to construct a smaller, yet more potent, training dataset. This method promises to significantly accelerate model training and reduce computational costs, while simultaneously enhancing the final model's accuracy, generalization, and robustness against noisy or out-of-distribution data.

In conclusion, this research represents a significant step towards creating more efficient and intelligent machine learning pipelines. It moves the focus from post-hoc model adaptation to proactive, data-centric optimization from the very beginning of the model lifecycle, laying a high-quality foundation for subsequent tasks such as robust continual learning.

REFERENCES

[1] de Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2021). *A continual learning survey: Defying forgetting in classification tasks*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Espinosa-Zarlenga, M., Shams, Z., & Jamnik, M. (2021). *ECLAIRE: Efficient decompositional rule extraction for deep neural networks*. arXiv preprint arXiv:2111.12628

[3] Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2021). *Curriculum learning: A survey of methodologies and applications*. arXiv preprint arXiv:2101.10382

[4] Sim, R. H. L., Smith, J., & Zhang, Y. (2022). *Data valuation in machine learning: Ingredients, desiderata, and open challenges*. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).

[5] Zhou, D., Li, K., & Wang, H. (2024). *Continual learning with pre-trained models: Challenges and perspectives*. Journal of Machine Learning Research, 25, 1–36.

[6] Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2020). *Learning from noisy labels with deep neural networks: A survey*. arXiv preprint arXiv:2007.08199

[7] Hospedales, T. M., Antoniou, A., Micaelli, P., & Storkey, A. (2022). *Meta-learning in neural networks: A comprehensive survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] Li, D., Wang, Z., Chen, Y., & Huang, X. (2024). *Deep active learning: Recent advances and new frontiers*. ACM Computing Surveys.

[9] Budd, S., Alvin, J., & Rees, G. (2021). *Active learning and human-in-the-loop deep learning for medical image analysis: A survey*. Medical Image Analysis, 71, 102062.

[10] Wang, X., Chen, W., & Zhu, R. (2020). *A survey on curriculum learning and its applications to deep learning*. arXiv preprint arXiv:2010.13166

[11] Zhao, B., Mopuri, K. R., & Bilen, H. (2021). *Dataset condensation with gradient matching*. In International Conference on Learning Representations (ICLR).

[12] Jiang, K., Lee, A., & Gupta, S. (2023). *OpenDataVal: A unified benchmark and toolkit for data valuation*. In NeurIPS Datasets and Benchmarks Track.

[13] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). *GLISTER: Generalization-based data subset selection for efficient and robust learning*. In Proceedings of the AAAI Conference on Artificial Intelligence.

[14] Liu, B., Park, S., & Chen, L. (2023). *FIRE: Fast interpretable rule extraction for deep models*. Journal of Machine Learning Research, 24, 1–28.

[15] Subbaswamy, A., Saria, S., & Schulam, P. (2021). *Evaluating model robustness and stability under dataset shifts*. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS).

[16] Contreras, V., Moreno, J., & Silva, R. (2022). *DEXiRE: Deep explanations and rule extraction for neural models*. Electronics (MDPI), 11(5), 800.

[17] Nguyen, V. L., Tran, H., & Li, Q. (2022). *Measuring uncertainty in uncertainty sampling for active learning: Methods and comparisons*. Machine Learning, 111(9), 3117–3140.

[18] Shao, Q., Patel, D., & Fernandez, A. (2024). *Representative sampling for semi-supervised learning: Methods to enhance label efficiency*. In Advances in Neural Information Processing Systems (NeurIPS).

[19] Majidi, F., Openja, M., Khomh, F., & Li, H. (2022). *An empirical study on the usage and impact of automated machine learning tools*. ACM Transactions on Software Engineering and Methodology.

[20] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2021). *RobustBench: A standardized adversarial robustness benchmark*. In NeurIPS Datasets and Benchmarks Track.

[21] Cazenavette, G., Wang, T., Torralba, A., Efros, A. A., & Zhu, J.-Y. (2022). *Dataset distillation by matching training trajectories*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] de Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2021). *A continual learning survey: Defying forgetting in classification tasks*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[2] Espinosa-Zarlenga, M., Shams, Z., & Jamnik, M. (2021). *ECLAIRE: Efficient decompositional rule extraction for deep neural networks*. arXiv preprint arXiv:2111.12628

[3] Soviany, P., Ionescu, R. T., Rota, P., & Sebe, N. (2021). *Curriculum learning: A survey of methodologies and applications*. arXiv preprint arXiv:2101.10382

[4] Sim, R. H. L., Smith, J., & Zhang, Y. (2022). *Data valuation in machine learning: Ingredients, desiderata, and open challenges*. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI).

[5] Zhou, D., Li, K., & Wang, H. (2024). *Continual learning with pre-trained models: Challenges and perspectives*. Journal of Machine Learning Research, 25, 1–36.

[6] Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2020). *Learning from noisy labels with deep neural networks: A survey*. arXiv preprint arXiv:2007.08199

[7] Hospedales, T. M., Antoniou, A., Micaelli, P., & Storkey, A. (2022). *Meta-learning in neural networks: A comprehensive survey*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] Li, D., Wang, Z., Chen, Y., & Huang, X. (2024). *Deep active learning: Recent advances and new frontiers*. ACM Computing Surveys.

[9] Budd, S., Alvin, J., & Rees, G. (2021). *Active learning and human-in-the-loop deep learning for medical image analysis: A survey*. Medical Image Analysis, 71, 102062.

[10] Wang, X., Chen, W., & Zhu, R. (2020). *A survey on curriculum learning and its applications to deep learning*. arXiv preprint arXiv:2010.13166

[11] Zhao, B., Mopuri, K. R., & Bilen, H. (2021). *Dataset condensation with gradient matching*. In International Conference on Learning Representations (ICLR).

[12] Jiang, K., Lee, A., & Gupta, S. (2023). *OpenDataVal: A unified benchmark and toolkit for data valuation*. In NeurIPS Datasets and Benchmarks Track.

[13] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021). *GLISTER: Generalization-based data subset selection for efficient and robust learning*. In Proceedings of the AAAI Conference on Artificial Intelligence.

[14] Liu, B., Park, S., & Chen, L. (2023). *FIRE: Fast interpretable rule extraction for deep models*. Journal of Machine Learning Research, 24, 1–28.

[15] Subbaswamy, A., Saria, S., & Schulam, P. (2021). *Evaluating model robustness and stability under dataset shifts*. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS).

[16] Contreras, V., Moreno, J., & Silva, R. (2022). *DEXiRE: Deep explanations and rule extraction for neural models*. Electronics (MDPI), 11(5), 800.

[17] Nguyen, V. L., Tran, H., & Li, Q. (2022). *Measuring uncertainty in uncertainty sampling for active learning: Methods and comparisons*. Machine Learning, 111(9), 3117–3140.

[18] Shao, Q., Patel, D., & Fernandez, A. (2024). *Representative sampling for semi-supervised learning: Methods to enhance label efficiency*. In Advances in Neural Information Processing Systems (NeurIPS).

[19] Majidi, F., Openja, M., Khomh, F., & Li, H. (2022). *An empirical study on the usage and impact of automated machine learning tools*. ACM Transactions on Software Engineering and Methodology.

[20] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., & Hein, M. (2021). *RobustBench: A standardized adversarial robustness benchmark*. In NeurIPS Datasets and Benchmarks Track.

[21] Cazenavette, G., Wang, T., Torralba, A., Efros, A. A., & Zhu, J.-Y. (2022). *Dataset distillation by matching training trajectories*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Дата надходження статті до редакції: 21.08.2025
Дата затвердження статті до друку: 07.09.2025
Опубліковано: 24.11.2025