Rethinking Model Selection Beyond ImageNet Accuracy for Waste Classification

Nermeen Abou Baker[a] and Uwe Handmann[b]
Computer Science Institute, Ruhr West University of Applied Science, Luetzowstr. 5, Bottrop, Germany

Keywords: Transfer Learning, Pretrained Model Selection, Transferability Metrics, Waste Classification.

Abstract: Waste streams are growing rapidly due to higher consumption rates, and they present repeating patterns that can be classified with high accuracy due to advances in computer vision. However, collecting and annotating large datasets is time-consuming, but transfer learning can overcome this problem. Selecting the most appropriate pretrained model is critical to maximizing the benefits of transfer learning. Transferability metrics provide an efficient way to evaluate pretrained models without extensive retraining or brute-force methods. This study evaluates six transferability metrics for model selection in waste classification: Negative Conditional Entropy (NCE), Log Expected Empirical Prediction (LEEP), Logarithm of Maximum Evidence (LogME), TransRate, Gaussian Bhattacharyya Coefficient (GBC), and ImageNet accuracy. We evaluate these metrics on five waste classification datasets using 11 pretrained ImageNet models, comparing their performance for finetuning and head-training approaches. Results show that LogME correlates best with transfer accuracy for larger datasets, while ImageNet accuracy and TransRate are more effective for smaller datasets. Our method achieves up to 364x speed-up over brute-force selection, which demonstrates significant efficiency in practical applications.

1 INTRODUCTION

It is estimated that by 2050, waste generation will increase by 70% due to the increasing consumption of consumers(Statista, 2023). Automating waste classification using a combination of AI and robotics will be critical to keep up with this growth. Waste patterns are difficult to sort because they can come in different shapes, colors, and states, and the scarcity of this data can limit the accuracy of the classification. Therefore, this study introduces transfer learning to overcome this challenge. Moreover, the growing need to save computational complexity and energy costs in the training phase is necessary for industrial applications.

Transfer learning leverages knowledge from a source domain/task and applies it to a related target domain/task (Thrun and Pratt, 1998). Pretrained models are deep learning architectures trained on large datasets, such as ImageNet (Deng et al., 2009). Task adaptation depends on the characteristics of both the pretrained model and the target task. Since different tasks require different pretrained models, this study limits the target datasets to a set of waste classification datasets to reduce domain shift. Specifically, five datasets from Kaggle and GitHub are used, with images crawled from search engines. Transfer learning from ImageNet is appropriate, as these datasets consist of natural images from real-world applications.

There are two ways to implement transfer learning:

- Retrain head (or feature extractor): This approach preserves the weights of the source features by freezing the feature extractor layer, which is a task-related layer, and retraining it using the target dataset.
- Finetuning: This technique involves replacing the task-related layer with a new one, and then finetuning the whole model.

Recently, several pretrained models have been studied, such as model hubs, model zoos, and model pools. This variety raises the following research question:

Which pretrained model should be selected without prior training on a classification task for a waste dataset?

[a] https://orcid.org/0000-0002-9683-5920
[b] https://orcid.org/0000-0003-1230-9446

Baker, N. A. and Handmann, U.
Rethinking Model Selection Beyond ImageNet Accuracy for Waste Classification.
DOI: 10.5220/0013111200003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 223-234
ISBN: 978-989-758-730-6; ISSN: 2184-4313

Although ImageNet accuracy is commonly used as a transferability metric, the performance of a model that excels on ImageNet does not necessarily indicate that it will perform best on other datasets. The effectiveness of a pretrained model can vary depending on the specific characteristics of the target task and dataset. In domain-specific applications, such as waste classification in this study, other transferability metrics may provide different insights. This work aims to answer the previous research question by adding the following contributions:

1. This study provides a thorough comparative analysis of six transferability metrics, including ImageNet accuracy correlation, NCE (Tran et al., 2019), LEEP (Nguyen et al., 2020), GBC (P'andy et al., 2021), TransRate (Huang et al., 2021), and LogME (You et al., 2021), specifically applied to five waste classification datasets, demonstrating their utility for this task.
2. This work shows that the effectiveness of these metrics varies with dataset size and compares their performance in feature extraction versus finetuning scenarios, emphasizing the importance of model selection over brute-force methods in transfer learning.
3. Performing a quantitative evaluation of the computational efficiency of these metrics, particularly the significant speed-ups compared to brute-force methods, while also providing insights into why certain metrics perform better in specific contexts.

The transferability scores should be taken without training on the target task. The best score must be effective, easily applicable to most pretrained models, and computationally efficient without training on the target data. Figure 1 shows the evaluation of the transferability metrics method in the selection of pretrained models for a target dataset.

This paper is structured to provide an analysis of transferability metrics in waste classification. Following this introduction, Section 2 reviews the existing literature on model selection strategies to provide the context for our research. Section 3 details our methodological approach, including dataset selection criteria, pre-processing techniques, and experimental design. In Section 4, we present our results, critically analyzing the performance of six transferability metrics in different waste classification datasets. Section 5 provides insights derived from our results, and the conclusion summarizes our main contributions and suggests directions for future research. By systematically evaluating these metrics, we aim to provide both theoretical insights and practical guidance for transfer learning researchers and practitioners on waste classification.

2 RELATED WORK

Previous work has attempted to evaluate the selection of pretrained models for supervised classification tasks in two approaches (Renggli et al., 2020):

- **Task Agnostic Model Search Strategies:** it ranks pretrained models before observing the target datasets. However, they used brute-force, which is expensive, and trained these models extensively on benchmark datasets to provide some guidelines for selecting the best models. The work of (Kornblith et al., 2018) compared 16 pretrained models on 12 datasets, and the authors found that there is a strong correlation between ImageNet accuracy and transfer accuracy in general, but not on fine-grained datasets. In addition, our previous work presented guidelines and evaluated how to select the most appropriate pretrained model that matches the target domain for image classification tasks based on application requirements by measuring accuracy, accuracy density, training time, and model size (Abou Baker et al., 2022).
- **Task-Aware Model Search Strategies:** Taskonomy used the loss (Zamir et al., 2018) and Task2Vec used the target dataset with additional computation by extracting learned representations from the pretrained model, then training a linear or K-Nearest Neighbour (KNN) classifier on these representations, and selecting the model with the highest accuracy using the Fisher information matrix after fully finetuning the pretrained model on the target dataset (Achille et al., 2019).

While these methods can provide some guidance in selecting the appropriate model source, they are computationally expensive. In addition, with a large number of pretrained models available on open-source frameworks such as PyTorch, TensorFlow, Hugging Face, Caffe, MATLAB, etc., it is becoming increasingly difficult to select the best pretrained model to meet the application requirements. These requirements vary in accuracy, energy, and computational cost in terms of memory (FLOPS) and training time. Brute-force is therefore not an efficient method.

Overall, this suggests the need for a better understanding of the pretrained model selection to evaluate the model pool. To determine source-task learning representations, a few scores have been introduced to assess the transferability measure that eliminates the need for training models and is therefore computationally efficient. Therefore, a fast, accurate,

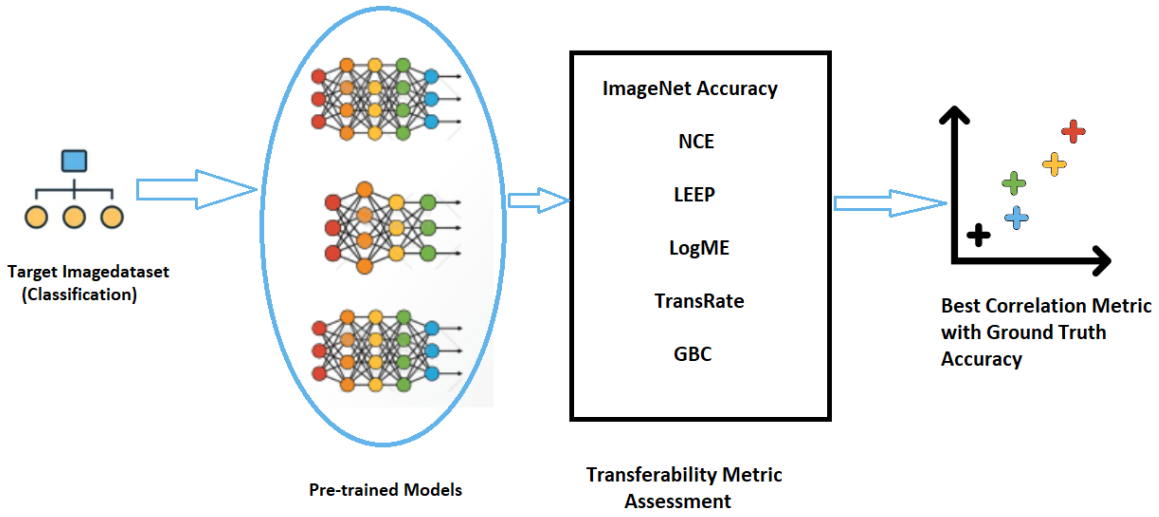

Figure 1: Evaluating pretrained model selection for a target dataset.

and generic assessment method is needed to solve the problem. The following transferability measures can be considered as a starting point for selecting the model among several others to achieve the best performance on a target task. Related works assess model selection (You et al., 2022), (Agostinelli et al., 2022), and (Renggli et al., 2020). However, all of them evaluate model ranking on fine-grained datasets or datasets with different label representations.

To estimate the transferability score of the tested candidates and to select the one with the maximum transferable score based on the available methods, there are two types of quantifying transferability measures (Bolya et al., 2021):

- **Label comparison-based (or probability-based) methods** that compute the dependence of the source and target label spaces. These methods assume equivalence between labels in the source and target domains or compute pseudo-labels by passing the source model to the target domain once, such as NCE and LEEP.
- **Source embedding-based methods** rely only on the feature extractor to embed labels from the target domain. Scores are then computed using these embeddings and their corresponding labels, such as LogME, TransRate, and GBC.

In addition, recent work has standardized the evaluation of transferability scores for pretrained model selection across 11 general vision datasets and evaluated 14 transferability scores using CNN and ViT models. The study evaluates both accuracy and computational complexity, using the weighted Kendall Tau score to efficiently rank models (Abou Baker and Handmann, 2024). While focused on general vision datasets, our waste classification study provides a more focused, empirical validation of transferability metrics for a specific domain.

2.1 Negative Conditional Entropy (NCE)

This method quantifies the amount of information from the source to the target domain, based on an information-theoretic quantity to assess transferability between tasks. The NCE score is shown to be related to the loss of the transferred model. It assumes that the training labels are random variables and investigates their statistics as follows: NCE estimates the joint distribution $P(y_t, y_s)$ with one-hot labels and predictions, then computes NCE as $-H(y_t|y_s)$ which represents the negative conditional entropy of the target labels y_t given the predictions y_s as ground truth source (Tran et al., 2019). The authors assume cross-entropy as the loss function and then show that the conditional entropy between the label sequences of their training sets for two tasks can define how well (or the likelihood of success) the representation learned from one task will perform on another task. This avoids training models and is therefore computationally efficient.

2.2 Log Expected Empirical Prediction (LEEP)

The idea behind the LEEP score is to measure the resonance between a pretrained model and a target dataset. The log-likelihood of the empirical conditional distribution is measured by calculating the av-

erage log-likelihood of the source and target labels (Nguyen et al., 2020). LEEP scores are calculated in three steps:

- Compute the dummy label distributions of the inputs by making a single forward pass of the pretrained model through the target dataset.
- Compute the empirical conditional distribution of the target label given the source label. This step estimates the joint distribution of the predicted and the true labels to compute an empirical predictor.
- The LEEP score is calculated by estimating the likelihood of an empirical predictor that maps the target labels to the predictions of the source model.

LEEP uses indirect representations of distributions, where the output label distribution is a linear transformation of the features, and the dummy labels contain information about the input features. The authors show that LEEP can also predict the convergence speed when finetuning the model. The scores are obtained without training on the target task, thus avoiding parameter optimization. LEEP uses the softmax output layer, which limits this score to classification tasks only.

2.3 The Logarithm of Maximum Evidence (LogME)

It is introduced to estimate the compatibility between source models and target datasets. LogME score estimates the accuracy of the target dataset using the following steps:

- The target images are embedded using the source feature extractor.
- The LogME score computes the probability condition (which is the evidence) of the target labels over these embeddings.
- To compute this evidence, the authors set up a graphical model that assumes the samples are independent.

LogME ranges in $[-1,1]$, where the closest value to -1 indicates the worst transferability values, and the value closest to $+1$ indicates the best. LogME doesn't require a softmax output layer, which makes it a candidate score for regression and unsupervised learning. Since LogME is generic, it can be used for classification and regression. However, this study focuses only on classification tasks. The original paper reports that compared to brute-force finetuning, computing the LogME provides at most a 3700 speed-up in wall-clock time and requires 1% of the memory.

2.4 TransRate

The TransRate score is designed to measure transferability by using the mutual information between target labels and features extracted by a pretrained model. Unlike many existing approaches, TransRate computes transferability in a single pass across all instances of the target dataset. Its key advantages include eliminating the need for computationally intensive modeling or training, significantly reducing computational costs by using coding rate as a proxy for entropy, and maintaining effectiveness even with finite datasets.

TransRate could also be used to compare transferability between source tasks, source models, and layers. Furthermore, this comparison is applied to supervised and self-supervised trained models for classification and regression tasks.

2.5 Gaussian Bhattacharyya Coefficient (GBC)

The GBC score measures the overlap between target classes in the source feature space. It measures how well a pretrained model transfers to the target dataset. According to the GBC score, the more classes overlap in the feature space, the more difficult it is to finetune the pretrained model for high accuracy. The GBC score is measured as follows: All target images are embedded in the feature space defined by the source model and represented with a per-class Gaussian, then the pairwise separability is estimated by the Bhattacharyya coefficient. According to the GBC score, the more classes overlap in the feature space, the more difficult it is to finetune the pretrained model for high accuracy. The authors applied the GBC score to semantic segmentation, where GBC outperformed state-of-the-art metrics.

3 METHODS

This study aims to evaluate the effectiveness of different transferability metrics in selecting optimal source models for transfer learning without the need for extensive training. We evaluated five transferability metrics, as well as the ImageNet accuracy correlation proposed by (Kornblith et al., 2018), on five different waste classification datasets. Our evaluation uses 11 models pretrained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 classification task, providing a robust foundation for comparison.

Our experimental framework is designed to systematically evaluate the performance of each transferability metric under different conditions. We consider two primary transfer learning scenarios: The retrain head and finetuning. For each scenario, we compute the correlation between the values of the transferability metrics and the ground truths of the target datasets. This approach allows us to assess not only the predictive performance of each metric but also its consistency in different transfer learning experiments.

3.1 Datasets

3.1.1 Selection Criteria

The selection of appropriate datasets is important for a comprehensive assessment of transferability metrics. We used the following criteria to ensure the relevance and diversity of our benchmark:

- **Domain Relevance:** Waste classification is a critical challenge that intersects environmental sustainability and computer vision. The selected datasets include diverse waste shapes, color variations, and spatial origins. This experimental framework goes beyond the traditional brute-force method. It provides a controlled representative domain for testing transferability metrics.
- **Label Diversity:** The datasets span a wide range of classifications. They include broad categories like glass, plastic, and metal, as well as fine-grained material identification of specific packaging types. This variety supports a thorough evaluation of transfer learning methods. It highlights how knowledge representations adapt to different levels of semantic granularity and contextual specificity.
- **Vision Complexity:** The datasets vary greatly in size. Smaller collections like Manon include 320 images, while large repositories like GarbageFine have 23,715 images. This scale diversity offers a robust platform for benchmarking. It demonstrates how transferability metrics perform under different data constraints and computational challenges.
- **Replicability and Accessibility:** The datasets are publicly available on platforms like Kaggle and Github, and were obtained through systematic web crawling. This ensures a transparent and replicable research pipeline. These web-derived image collections reflect real-world computational environments where transfer learning technologies will be deployed. This approach ensures scientific validity and practical relevance.

Table 1: The tested datasets of waste classification that come from web crawling.

Dataset	# of classes	Train size	Test size
Manon str (Yacharki, 2013)	5	320	83
Trashnet (Thung, 2018)	6	2,019	508
Trahbox (TrashBox, 2024)	7	16,060	1,793
WasteFine (WasteFine, 2023)	34	17,873	5,756
GarbageFine (GarbageFine, 2023)	58	23,715	5,958

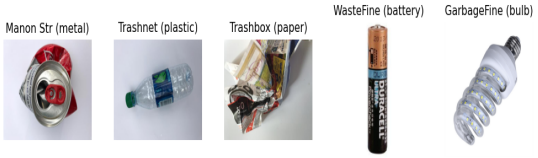

Figure 2: Sample images (with their corresponding label) for each dataset.

3.1.2 Dataset Overview

Table 1 provides a summary of the selected datasets, indicating the number of classes and sample sizes for training and testing. Figure 2 illustrates sample images with their corresponding labels from each dataset, to provide visual context for the classification tasks.

3.1.3 Data Pre-Processing

To improve the generalization capabilities of our models and to ensure consistency across experiments, we applied several data pre-processing and augmentation techniques:

- Dataset splitting: For datasets without predefined splits, we used an 80:20 ratio for training and test sets. Where original training and validation splits existed, we merged them to form the training set, while retaining the original test set for evaluation.
- Data augmentation: We implemented several augmentations to the training and test sets, including random resized crop, random horizontal flip, and image normalization using the mean and standard deviation of the ImageNet dataset to ensure consistency with the pretraining data distribution.

These pre-processing and augmentation steps are essential to improve model generalization and allow fair comparison between different pretrained models and datasets.

3.1.4 Scope and Limitations

Our focus on waste classification is motivated by the critical global challenge of waste management and recycling. The increasing volume of waste and the need for efficient sorting technologies make waste classification a crucial area of research with significant en-

vironmental and economic implications (Abou Baker et al., 2023).

Although our study provides insight into transferability metrics specific to waste classification datasets, we acknowledge the domain-specific nature of our research. The selected datasets, ranging from 5 to 58 classes and representing different waste sorting scenarios, provide a comprehensive exploration within the waste classification domain. However, the results are not intended to be universally applicable to all image classification tasks.

3.2 Models

In this study, we evaluate 11 Convolutional Neural Networks (CNN) architectures that cover a wide range of model complexities and ImageNet accuracies. These architectures represent the current state of image classification models and can be categorized into four groups based on their architectural design: Resnets as skip connections (ResNet-34, ResNet-50, ResNet-101, ResNet-152 (He et al., 2016)), parallel convolution filters (Inception-V3 (Szegedy et al., 2016), and GoogleNet (Szegedy et al., 2015)), densely connected blocks (DenseNet-121, DenseNet-169, DenseNet-201 (Huang et al., 2016)), or convolutional neural networks designed for mobile and edge devices (MnasNet1 − 0 (Tan et al., 2018), MobileNet-V2(Sandler et al., 2018)).

We use these pretrained models in two transfer learning methods: full model tuning and retrain head. This allows a comprehensive evaluation of the correlation between transferability scores and test accuracy.

For transfer learning experiments, we use a standardized training protocol to ensure fair comparisons. We use Stochastic Gradient Descent (SGD) optimization with a momentum of 0.9, an initial learning rate of $10^{-}3$, initial learning rate with 0.1 step weight decay every 7 epochs. We use a batch size of 16 for all experiments, which were run on NVIDIA RTX8000 GPU.

Although we understand that optimal hyperparameters may vary significantly between models and datasets, we choose this uniform setup to maintain consistency and facilitate direct comparisons. This approach is consistent with common practices in transfer learning research, although we recognize that performance could potentially be improved through extensive hyperparameter tuning and advanced training strategies.

3.3 Model Selection Process

To systematically evaluate the effectiveness of different transferability metrics and to simplify the pretrained model selection process, we present the following workflow:

- Feature extraction: Features are extracted from the penultimate layer of each model to capture high-level representations for transfer learning. These features are inputs to the tested transferability metrics, which are used to calculate scores and to measure computation time.
- Evaluation of the transferability metrics: The approach validates each metric by computing the Pearson correlation coefficient between the values of the transferability scores and the ground truth for each dataset. The Pearson correlation, which ranges from -1 to $+1$, measures the strength and direction of the linear relationship, with values near -1 indicating a strong negative correlation, near 0 indicating no linear correlation, and near $+1$ indicating a strong positive correlation. This analysis ranks the pretrained models and identifies the most appropriate metric for each data set.
- Model selection and correlation analysis: The output includes the selected models, correlation coefficients, and computation times, providing a structured and objective approach to simplify model selection in transfer learning without training.

The rationale for this approach addresses the limitation of using ImageNet accuracy as the only predictor of model transferability. By evaluating multiple transferability metrics and correlating their values with ground truth performance, the method provides a more robust strategy for selecting pretrained models. This approach challenges the assumption that ImageNet performance universally indicates transferability and provides a practical methodology for assessing feature transferability across diverse waste datasets. The algorithm 1 shows the steps in the process.

4 RESULTS AND DISCUSSION

Our analysis focuses on the correlation between 6 transferability metrics and actual transfer learning performance in different datasets and transfer learning strategies. Figures 3 and 4 present a comprehensive view of these correlations for the feature extraction (retrain-head) and full model finetuning approaches, respectively.

In Figure 3, we observe the performance of our transferability metrics when applied to the retrain

Algorithm 1: Model selection and evaluation for transfer learning.

Input:

- Target datasets $\{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4, \mathcal{D}_5\}$
- Pretrained models $\mathcal{M} = \{m_j\}_{j=1}^{11}$
- Model selection metrics $\Phi = \{\phi_i\}_{i=1}^{6}$
- Ground truth performance values $\{\mathcal{A}_d\}_{d=1}^{5}$, where each $\mathcal{A}_d = \{a_{dj}\}_{j=1}^{11}$

Output:

- Selected models $\{m_d^*\}_{d=1}^{5}$ for each dataset
- Pearson correlation coefficients $\{\rho_d\}_{d=1}^{5}$
- Computation times $\{\mathcal{T}_d\}_{d=1}^{5}$

Procedure:

1. **Feature extraction**
   - For each dataset $\mathcal{D}_d$ and model $m_j \in \mathcal{M}$:
     - Extract representations from penultimate layer: $f_{dj} = \ell_p(m_j(\mathcal{D}_d))$
   - Set $\mathcal{F}_d = \{f_{dj}\}_{j=1}^{11}$ for each dataset

2. **Evaluate model selection metrics**
   - For each dataset $\mathcal{D}_d$:
     - For each metric $\phi_i \in \Phi$ and model $m_j \in \mathcal{M}$:
       - Start timer t_{start}
       - Calculate transferability score: $\sigma_{dij} = \phi_i(f_{dj}, \mathcal{D}_d)$
       - Record time: $\tau_{dij} = t_{\text{current}} - t_{\text{start}}$
   - Set $\mathcal{S}_d = \{\sigma_{dij}\}$ and $\mathcal{T}_d = \{\tau_{dij}\}$ for each dataset

3. **Model selection and correlation analysis**
   - For each dataset $\mathcal{D}_d$:
     - For each metric $\phi_i \in \Phi$:
       - Let $\sigma_{di} = [\sigma_{di1}, \ldots, \sigma_{di11}]$
       - Calculate Pearson correlation: $\rho_{di} = \frac{\text{Cov}(\sigma_{di}, \mathcal{A}_d)}{s_{\sigma_{di}} s_{\mathcal{A}_d}}$
     - Set $\rho_d = [\rho_{d1}, \ldots, \rho_{d6}]^T$
     - Select best metric $i_d^* = \arg\max_i \rho_{di}$
     - Select best model $m_d^* = \arg\max_{m_j \in \mathcal{M}} \sigma_{di_d^* j}$

Return:

- Selected models $\{m_d^*\}_{d=1}^{5}$
- Correlation coefficients $\{\rho_d\}_{d=1}^{5}$
- Computation times $\{\mathcal{T}_d\}_{d=1}^{5}$

head method. The columns represent our 5 target waste classification datasets, while the rows correspond to the 6 transferability metrics under evaluation: NCE, LEEP, LogME, TransRate, GBC, and ImageNet accuracy. Each individual subplot illustrates the correlation between a specific transferability metric (y-axis) and the ground truth accuracy (x-axis) achieved by our 11 pretrained models on a particular dataset, with the best correlation in bold.

Following the same visualization structure as in Figure 3, we extend our analysis to the full model finetuning method in Figure 4. As described in the section 3, all scores are based on the training set only.

We find that the datasets do not consistently follow the correlation patterns expected from ImageNet accuracy. This observation challenges the common assumption that performance on ImageNet is a reliable predictor of transferability across visual tasks. The inconsistency demonstrates the task-specific nature of transfer learning and suggests that the features learned on ImageNet may not be equally relevant or transferable to all target tasks. For smaller datasets such as Manon Str, the accuracy correlation of ImageNet proves to be a useful metric, ranking first in the Pearson correlation for finetuning the full model and third for retrain-head. This suggests that for tasks with limited data, the broad feature representations learned on ImageNet can provide a strong starting point. The effectiveness here is due to the diversity and scale of ImageNet, which allows models to learn general visual features, which can be particularly beneficial when target data is scarce.

In contrast, LogME proves to be a strong performer, showing a high correlation with most datasets, except for the small Manon Str dataset. The effectiveness of LogME is attributed to its probabilistic approach to estimating the compatibility between source models and target datasets. By modeling the evidence of target labels given the embeddings of the source model, LogME captures a more detailed representation of transferability.

Interestingly, TransRate demonstrates a good correlation with the Manon Str dataset, especially in the retrain head scenario. This is consistent with TransRate's theoretical foundation, which is well-suited for finite examples as shown in section 2.4. Using the coding rate as an alternative to entropy allows it to efficiently capture essential information for transfer, which makes it particularly effective for smaller datasets.

On the other hand, NCE and LEEP do not perform well across experiments. These label comparison-based methods, which rely on retraining a linear classifier to estimate joint distributions between source and target labels, appear prone to overfitting. In waste classification, where class definitions are often ambiguous or overlapping, the assumption of a direct relationship between label spaces does not work, which

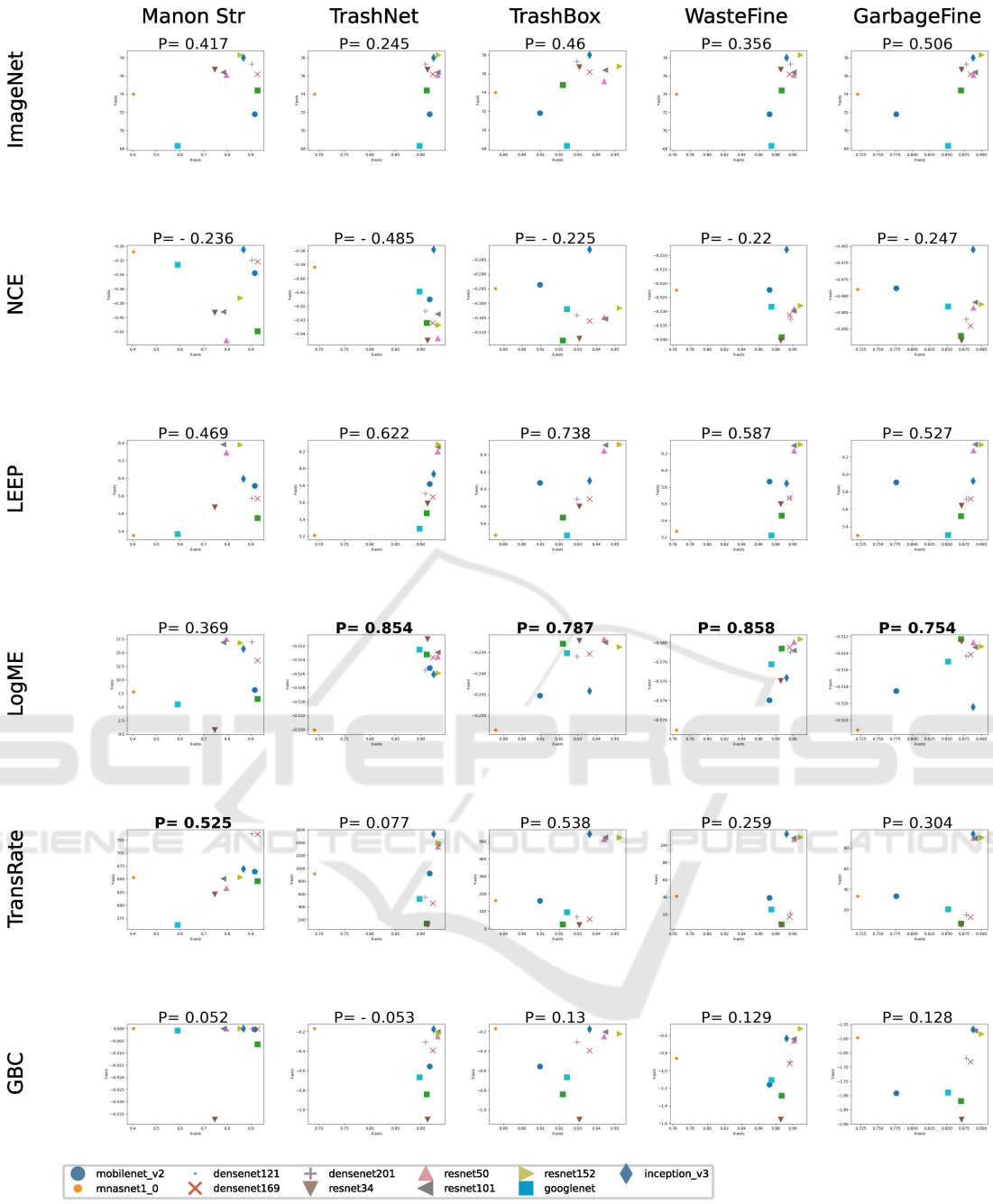

Figure 3: Pearson correlation (P) for **retrain-head** between ground truth accuracy (X-axis) and 6 transferability metrics (Y-axis) with 11 pretrained model selection for 5 target datasets.

leads to unreliable transferability estimates.

Additionally, the GBC score shows mixed results, with negative correlations for smaller datasets such as Manon Str and TrashNet. This behavior shows the limitations of assuming Gaussian distributions and linear separability in feature spaces, especially for complex tasks or limited data scenarios. In waste classification, where object appearance can vary significantly within classes, feature distributions may be far from Gaussian, further invalidating the assumptions of GBC.

Surprisingly, NCE shows a negative correlation across experiments, which challenges the straightforward application of information-theoretic principles

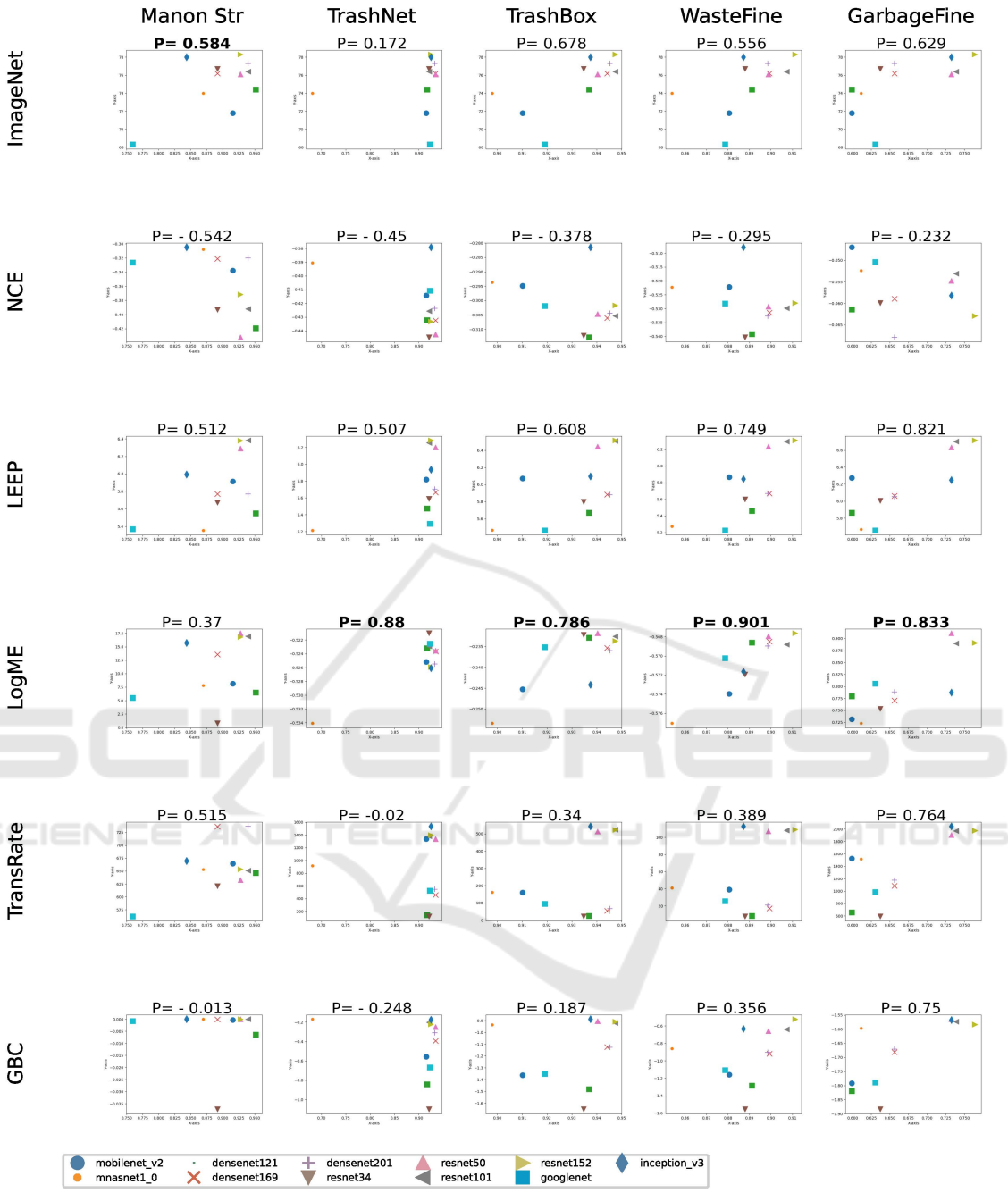

Figure 4: Pearson correlation (P) for **finetuning** between ground truth accuracy (X-axis) and 6 transferability metrics (Y-axis) with 11 pretrained model selection for 5 target datasets.

to transfer learning. This consistent negative correlation suggests that the assumptions underlying NCE, particularly, the relationship between conditional entropy and transferability, may not apply uniformly across different tasks and datasets.

When comparing the computational efficiency of model selection metrics, LogME demonstrates significant advantages over brute-force finetuning. For example, in experiments on the Manon Str dataset, LogME is computed in just 5 minutes for 11 pretrained models, achieving a speed-up of 42.6x over the brute-force method, which takes almost 3.5 hours. For the larger GarbageFine dataset, LogME takes 7 minutes compared to almost 42 hours for brute-force,

resulting in a remarkable speed-up of 363.7x. This efficiency is due to the ability of LogME to filter out redundant information in features, combined with its strong performance. This makes it particularly attractive for fast and efficient model selection in transfer learning scenarios.

These findings demonstrate the need for a model selection metric to assess transferability in image classification. While ImageNet accuracy correlation can provide useful insights, especially for smaller datasets, it should not be relied upon as the only indicator of transferability. The success of metrics such as LogME in certain scenarios suggests the importance of considering feature space structure and target task characteristics when assessing transferability.

5 INSIGHTS FROM METRIC PERFORMANCE ACROSS MODELS AND DATASETS

The following key takeaways summarise the findings from analyzing various transferability metrics, focusing on their performance, stability, and sensitivity across different models and datasets in the feature extraction and the finetuning scenarios.

- *The Consistent Performance of LogME*: The LogME metric performs consistently well in feature extraction and finetuning scenarios (except for the Manon Str dataset). Its robustness in estimating compatibility between models and datasets suggests that it captures essential aspects of transferability, regardless of the transfer learning method. On the other hand, the GBC shows consistently lower correlation coefficients compared to other metrics. This may indicate that class separability in feature space, as measured by the GBC, may not be a useful factor for transfer success in these specific tasks.
- *Improved Metric Correlations After Finetuning*: Finetuning leads to improved correlation coefficients for many metrics, particularly for LogME and LEEP. This indicates that these metrics have improved predictive performance after finetuning the models, demonstrating the value of finetuning for a better understanding of transferability. However, the varying degrees of improvement across metrics and datasets suggest a complex, task-dependent relationship between initial transferability estimates and performance after finetuning.
- *Dataset-Dependent Metric Performance*: Transferability metrics such as LogME and GBC show significant variability across datasets, suggesting that transferability is not just a property of models, but also depends on model-dataset interactions. This variability emphasizes the importance of dataset characteristics in transfer learning outcomes.
- *Influence of Model Architectures*: Certain architectures, such as Inception_v3 and Mobilenet_v2, perform consistently well across metrics and datasets, especially after finetuning. This suggests that some architectures have inherent characteristics that make them more adaptable to transfer learning.
- *Metrics Variability*: Metrics such as GBC and TransRate show variability, with the sensitivity of the TransRate to the mutual information between target labels and features leading to fluctuations between feature extraction and finetuning. While some metrics, like LogME, remain stable, others show higher sensitivity, suggesting the need for a combination of metrics to get a comprehensive evaluation of transferability.

6 CONCLUSIONS

This study investigates the effectiveness of transferability metrics in selecting pretrained models for waste classification, which is a critical challenge in representation learning. Six metrics (NCE, LEEP, LogME, TransRate, GBC, and ImageNet accuracy) are evaluated by transferring knowledge from ImageNet to five datasets of varying size, label density, and diversity. The analysis examines performance for full model finetuning and head-only retraining, providing practical insights into the utility of metrics in different scenarios.

The results challenge the assumption that ImageNet accuracy reliably predicts transferability across datasets and tasks. While ImageNet accuracy remains effective for smaller datasets and overall model tuning, its correlation with transfer performance is inconsistent for larger datasets. In contrast, LogME shows stronger and more stable performance and emerges as a robust metric for model selection. Additionally, TransRate shows particular promise in head-training scenarios. These results demonstrate the need for a detailed approach to model selection, considering the dataset's characteristics and the considered task.

Although the experimental results focus on waste classification, they show the limitations of ImageNet's accuracy and highlight the need for broader validation. Extending this evaluation framework

to other domain-specific classification tasks, cross-domain experiments, and diverse dataset characteristics will be important for generalizing these results.

Future research should extend beyond the current scope by exploring several promising avenues. First, including Vision Transformer (ViT) models, or finetuning large pretrained models (Abou Baker et al., 2024) would provide insight into how newer architectural paradigms perform in transfer learning scenarios. Second, developing more advanced hyperparameter optimization techniques could further refine model selection strategies. Third, expanding the diversity of datasets to include more domain-specific and cross-domain challenges would test the generalizability of our findings. In addition, exploring the interaction between transferability metrics and emerging techniques such as few-shot learning could provide new approaches for efficient machine learning model adaptation.

In conclusion, effective transferability metrics must balance speed and accuracy to identify appropriate pretrained models without extensive finetuning. This research contributes to a deeper understanding of transferability in deep learning, providing a foundation for broader evaluations and practical guidance in waste classification and beyond.

ACKNOWLEDGEMENTS

This work has been funded by the Ministry of Economy, Innovation, Digitization, and Energy of the State of North Rhine-Westphalia, Germany, within the project Digital.Zirkulär.Ruhr.

REFERENCES

Abou Baker, N. and Handmann, U. (2024). One size does not fit all in evaluating model selection scores for image classification. *Scientific Reports*, 14(1):30239.

Abou Baker, N., Rohrschneider, D., and Handmann, U. (2024). Parameter-efficient fine-tuning of large pretrained models for instance segmentation tasks. *Machine Learning and Knowledge Extraction*, 6(4):2783–2807.

Abou Baker, N., Stehr, J., and Handmann, U. (2023). E-waste recycling gets smarter with digitalization. In *2023 IEEE Conference on Technologies for Sustainability (SusTech)*, pages 205–209.

Abou Baker, N., Zengeler, N., and Handmann, U. (2022). A transfer learning evaluation of deep neural networks for image classification. *Machine Learning and Knowledge Extraction*, 4(1):22–41.

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C., Soatto, S., and Perona, P. (2019). Task2vec: Task embedding for meta-learning. In *ICCV 2019*.

Agostinelli, A., Pándy, M., Uijlings, J., Mensink, T., and Ferrari, V. (2022). How stable are transferability metrics evaluations? In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*, page 303–321, Berlin, Heidelberg. Springer-Verlag.

Bolya, D., Mittapalli, R., and Hoffman, J. (2021). Scalable diverse model selection for accessible transfer learning. In *Neural Information Processing Systems*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

GarbageFine (2023). Garbage dataset. `https://www.kaggle.com/datasets/mrk1903/garbage`. Accessed: 2024-09-27.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Huang, G., Liu, Z., and Weinberger, K. Q. (2016). Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269.

Huang, L.-K., Wei, Y., Rong, Y., Yang, Q., and Huang, J. (2021). Frustratingly easy transferability estimation. In *International Conference on Machine Learning*.

Kornblith, S., Shlens, J., and Le, Q. V. (2018). Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666.

Nguyen, C. V., Hassner, T., Archambeau, C., and Seeger, M. W. (2020). Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*.

P'andy, M., Agostinelli, A., Uijlings, J. R. R., Ferrari, V., and Mensink, T. (2021). Transferability estimation using bhattacharyya class separability. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9162–9172.

Renggli, C., Pinto, A. S., Rimanic, L., Puigcerver, J., Riquelme, C., Zhang, C., and Lucic, M. (2020). Which model to transfer? finding the needle in the growing haystack. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9195–9204.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

Statista (2023). Global waste generation: statistics and facts. `https://www.statista.com/topics/4983/waste-generation-worldwide/topicOverview`. Accessed: 2024-09-27.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In

2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.

Tan, M., Chen, B., Pang, R., Vasudevan, V., and Le, Q. V. (2018). Mnasnet: Platform-aware neural architecture search for mobile. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2815–2823.

Thrun, S. and Pratt, L. (1998). *Learning to Learn: Introduction and Overview*, pages 3–17. Springer US, Boston, MA.

Thung, G. (2018). Trashnet: A dataset of images of garbage. `https://github.com/garythung/trashnet`. Accessed: 2024-09-27.

Tran, A., Nguyen, C., and Hassner, T. (2019). Transferability and hardness of supervised classification tasks. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1395–1405.

TrashBox (2024). Trashbox. `https://github.com/nikhilvenkatkumsetty/TrashBox`. Accessed: 2024-09-27.

WasteFine (2023). Waste pictures dataset. `https://www.kaggle.com/datasets/wangziang/waste-pictures`. Accessed: 2024-09-27.

Yacharki (2013). Manon str cleaned dataset. `https://www.kaggle.com/datasets/yacharki/manon-str-cleaned-dataset`. Accessed: 2024-09-27.

You, K., Liu, Y., Long, M., and Wang, J. (2021). Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*.

You, K., Liu, Y., Zhang, Z., Wang, J., Jordan, M. I., and Long, M. (2022). Ranking and tuning pre-trained models: A new paradigm for exploiting model hubs. *JMLR*.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.