OPEN ACCESS
International Journal of Advances in Artificial Intelligence and Machine Learning Vol 3 No 1 (2026) https://doi.org/10.58723/ijaaiml.v3i1.656

Check for updates

Research Article

Transfer Learning Effectiveness Across Domain Similarity Levels in Data Science Applications

Received: February 13, 2026	Revised: March 03, 2026	Accepted: March 21, 2026	Publish: March 31, 2026

Eko Risdianto*, Thai Ky Trung Pham, William Yeoh, Sultan Hammad Alshammari

Abstract:

Background: Transfer learning has become increasingly prominent in data science due to the challenges posed by limited labeled data and distribution shifts between training and deployment environments. However, the success of transfer learning depends significantly on the structural compatibility between source and target domains.

Aims: This study aims to investigate the relationship between domain similarity and transfer learning performance using an experimental framework termed Similarity-Aware Transfer Evaluation (SATE).

Methods: Twelve pairs of benchmark datasets were selected to simulate varying levels of domain similarity and were made publicly available. Domain similarity was computed using Maximum Mean Discrepancy (MMD) in the learned representation space. Transfer performance was measured using a predefined Transfer Gain metric under bounded fine-tuning strategies. Correlation analysis and statistical testing were conducted to examine the relationship between similarity scores and transfer effectiveness, while fine-tuning depth was analyzed in relation to similarity magnitude.

Result: The results demonstrate a strong positive correlation between domain similarity and transfer gain ($r = 0.83$, $p < 0.01$), indicating that approximately 69% of performance variability can be explained by similarity-based transfer effects. Negative transfer was observed when similarity scores were $S \leq 0.41$. Furthermore, higher similarity levels were associated with deeper and more stable fine-tuning, whereas lower similarity resulted in increased instability during adaptation. These findings establish similarity as a structural compatibility constraint in transfer learning.

Conclusion: The study confirms that domain similarity plays a fundamental role in determining transfer learning success. By operationalizing similarity measurement and linking it to performance thresholds, the proposed SATE framework provides a structured method for evaluating transfer feasibility in real-world data science applications.

Keywords: Cross-Domain Learning; Data Science; Domain Similarity; Model Adaptation; Transfer Learning.

1. INTRODUCTION

The growing significance of transfer learning within data science is easily understood as it allows for consideration of the relevant source datasets that have been shown to be useful for the purposes of the target tasks that are of interest (Tang et al., 2024). Accessibility for the target tasks of interest means it is necessary to identify the applicable source datasets that are important for the target domain tasks (Khan et al., 2024). The value of transfer learning is demonstrated clearly in its applications in the various domains of data science including computer vision, data science in the monitoring of industrial processes, analytics in the health science domain, and data science in the processing of natural language (Ali & Abdulazeez, 2024). The challenge with real-world applications of data science is the restrictions on the amount of data that can be collected in the form of annotations, as well as shifts in data distribution and the presence of data that is not homogenous (Tamang et al., 2025). In the face of these challenges, the advancement of transfer learning has sought to improve the efficiency of data collection with regard to its beneficial use, improve the speed of learning to be used within the target domain tasks, and improve the generalization that can be applied across the various tasks (Zhu et al., 2025).

Although transfer learning has become widely popular, it does not always result in an increase in performance. Some studies show that performance increases are more likely when there is a more similar relationship between the source and target domains (Pak & Paal, 2022). When certain domain characteristics diverge, such as the feature distributions, semantics of the labels, or the processes in which the data is generated, models can underperform due to a phenomenon known as negative

transfer (Zhao et al., 2024). In real-world situations, this unpredictability can create a lack of confidence in the use of transfer learning, especially in critical situations where consistent performance is necessary (Javed et al., 2025). In this regard, the question that this research seeks to answer is how the effectiveness of transfer learning is impacted by the differing levels of similarity between the domains, and which methods can be used to adapt to, or alleviate, negative transfer (Hosna et al., 2022).

Previous research has focused primarily on transfer learning in specific application areas employing end-to-end deep learning models (Yan et al., 2024). For instance, when source and target tasks are close to each other, tuning pretrained convolutional neural networks and modifying transformer-based models to downstream tasks leads to substantial improvements in performance (Woesle et al., 2025). Other studies evaluate the performance of frozen feature extraction and full fine-tuning to understand the extent of parameter re-allocation (Riyazuddin, 2025). While these studies appreciate the reuse of knowledge, they focus on domain and task similarities and leave the impact of domain divergence on performance unquantified. Consequently, their findings are highly contextual and provide little spillover knowledge across different contexts or degree of similarity (Mahn & Poblete, 2023).

Another avenue of research examines the alignment of representations and techniques for domain adaptation to address distribution mismatch. Adversarial domain adaptation, feature space regularization, and self-supervised pretraining methods aim to foster representation invariance, thereby target domain gap narrowing (Singhal et al., 2023). These methods suggest a holistic positive outcome from alignment mechanisms on transfer degradation despite the majority of extant studies identifying limited positive outcomes. Most extant studies either focus on singular benchmark pairs and experimental conditions, or define a limited positive outcome when applied to span a structured evaluation framework across low, moderate, and high similarity domains (Dan et al., 2023). Further, researchers have neglected to systematically examine the implications of varying domain and data similarities on the outcomes of fine-tuning depth, feature reuse approaches, and data availability during cross domain evaluations (Xu et al., 2024).

The existing literature consolidates that when research domains are closely tied, transfer learning aids in advancement of supportive research domains, however, there is no systematic research that looks at how inter domain relations research domains affect transfer effectiveness in varied cases (Zhang et al., 2026). More specifically, there is no experimental plan that systematically looks at the positive gains in performance, the possibility of negative transfer, and adaptive approaches across different levels of similarity, using the same research inter domain relations (Davila & Colan, 2025). To answer the above, this research is focused on (1) Establishing an evaluation framework for inter domain relations research domains that is categorized as low, moderate, or high dimensions of similarity; (2) the level of positive transfer as influenced by fine tuning and feature reuse at different depths; and (3) the development of adaptive approaches that reduce the negative transfer at low similarity. This research is intended to provide direction to data scientists on how to optimally incorporate transfer learning within the design process of a given application.

2. MATERIAL AND METHOD

This research focused on transfer learning and, in particular, how the level of similarity of the corresponding cognitive domains of the source and target domains affect factors such as the positive and negative impacts on performance and the adapting behaviours of the individual participants. The overall process includes (i) selection of the dataset(s) and definition of the similarity of the domains, (ii) data preprocessing and standardization of the features, (iii) execution of the transfer learning processes and selection of the depth of fine-tuning, and (iv) evaluation of the processes with step-controlled processes of similarity. The clarity and structure of the processes facilitate ease of understanding, as well as the ability to replicate and yield valid results to compare the conditions of the experiments.

Data Source and Domain Similarity Design

In order to analyze the effectiveness of transfer learning with different degrees of similarity, twelve datasets (six source and six target) pairs with high, medium, and low similarities were identified. The datasets that are high in similarity use the same modality with almost the same semantic tasks. The medium similarity datasets are consistent with the same modality but differ with respect to the class structure, size of the labels, or distribution (Weiss et al., 2016). The low similarities dataset have high differences in structure in the distributions of features, meaning, and data-generating processes. Using this multi pair method provides the possibility of a statistically significant regression analysis between the similarity score and the transfer gain (Bai & Ma, 2025; Plested et al., 2026).

In order to facilitate replicability and accessibility, all datasets were sourced from public benchmark repositories. Only samples that had complete feature representations and class annotations were considered. In order to avoid information leakage, source and target datasets were kept mutually exclusive. For each transfer scenario, dataset partitions were created independently (Căvescu & Popescu, 2026; Joeres et al., 2025).

In order to define domain similarity for this work, the author uses Maximum Mean Discrepancy (MMD) to measure the divergence of source and target feature representations (Lin et al., 2025). Let P_s and P_t denote

the empirical feature distributions of the source and target datasets, respectively, defined over representation space $\mathcal{X}$ (Yu et al., 2024). The squared MMD between the two distributions is defined as:

$$\mathrm{MMD}^2(P_s, P_t) = \mathbb{E}_{x,x' \sim P_s}[k(x,x')] + \mathbb{E}_{y,y' \sim P_t}[k(y,y')] - 2\mathbb{E}_{x \sim P_s, y \sim P_t}[k(x,y)] \quad (1)$$

where $k(\cdot,\cdot)$ is a positive-definite kernel function. In this work, a Gaussian radial basis function (RBF) kernel is employed:

$$k(x,y) = \exp\left(-\frac{\| x-y \|^2}{2\sigma^2}\right) \quad (2)$$

where σ denotes the kernel bandwidth, selected using the median heuristic computed over pairwise distances in the combined source–target feature set.

Notably, MMD is calculated on the intermediate feature embeddings that have been extracted from the pretrained encoder, and before it is fine-tuned, to ensure that the measurement of similarity is structural alignment of the representation space, as opposed to the measurement of similarity at the level of raw pixels.

To convert divergence into a normalized similarity score, similarity Sis defined as:

$$S = 1 - \frac{\mathrm{MMD}(P_s, P_t)}{\max_{i,j} \mathrm{MMD}(P_i, P_j)} \quad (3)$$

where the normalization ensures $S \in [0,1]$. Values nearer to one denote stronger similarity between the distributions of the domains.

This formulation permits the transfer gain as a function of measurable domain proximity to be analyzed systematically, and it allows for the regression-based analysis of the similarity thresholds to identify positive and negative transfer states.

Table 1 displays the 12 source-target dataset pairs utilized in this study, giving the sample size, dimension of features, and similarity scores calculated from the formula above, in order to provide experimental transparency and reproducibility.

Table 1. Dataset Pairs and Domain Similarity Characteristics

Source	Target	Samples (S/T)	Feature Dim	Similarity Score (S)	Similarity Level
CIFAR-10	CIFAR-100	50k/50k	3072	0.84	High
CIFAR-100	Tiny-ImageNet	50k/100k	3072	0.81	High
ImageNet-Subset	CIFAR-10	100k/50k	2048	0.78	High
MNIST	EMNIST	60k/70k	784	0.76	High
MNIST	Fashion-MNIST	60k/60k	784	0.64	Moderate
CIFAR-10	STL-10	50k/13k	3072	0.61	Moderate
CIFAR-100	SVHN	50k/73k	3072	0.55	Moderate
Tiny-ImageNet	STL-10	100k/13k	3072	0.52	Moderate
CIFAR-10	SVHN	50k/73k	3072	0.39	Low
MNIST	CIFAR-10	60k/50k	784/3072	0.34	Low
Fashion-MNIST	SVHN	60k/73k	784/3072	0.31	Low
CIFAR-100	MNIST	50k/60k	3072/784	0.28	Low

Table 1 demonstrates the source-target dataset pairs for constructing high, medium, and low transfer scenarios. Similarity scores are normalized to $[0, 1]$, larger scores indicate stronger proximity of distributions between the domains according to the selected divergence.

Preprocessing

To provide equal experimental conditions, all datasets were run through the identical preprocessing pipeline. The numerical features were z-score normalized, which is defined as

$$x' = \frac{x-\mu}{\sigma} \quad (4)$$

Image datasets had inputs adjusted to the same resolution and pixel values normalized to $[0, 1]$. During the training process, flips and rotations were used to augment the dataset.

To strengthen the model and mitigate overfitting, the training dataset was subjected to various augmentation strategies, including random horizontal flipping and slight rotational alterations. If numerical features had missing values, mean imputation was used to address this. Each preprocessing step was conducted using the same library functions and fixed random seeds to ensure reproducibility.

Proposed Evaluation Framework

Similarity-Aware Transfer Evaluation (SATE) is the first evaluation framework proposed in this study, allowing the evaluation of various transfer adaption strategies across similarity regimes. The framework is designed to be model independent, instead focusing on controlling and isolating the impact of adaptation depth and feature reuse.

Given input samples x and labels y, the predictive model f_θ, parameterized by θ, is trained by minimizing the empirical risk:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(x_i), y_i) \quad (5)$$

where ℓ denotes cross-entropy loss for classification tasks and N represents the number of training samples.

Three adaptation strategies are considered. For the feature extraction scenario, encoder parameters of the pretrained model were frozen, and only the classification head was trained. For partial fine-tuning, the upper layers of the encoder were unfrozen and optimized together with the classifier. For full fine-tuning, all model parameters were updated on the target dataset.

Transfer effectiveness was quantified using the Transfer Gain metric:

$$TG = P_{TL} - P_{Scratch} \quad (6)$$

where P_{TL} denotes performance achieved via transfer learning and $P_{Scratch}$ denotes performance when training from scratch under identical conditions. Relative transfer gain was also computed to normalize improvements across tasks with varying baseline performance levels.

The framework is designed on the premise that the type of domain similarity is a determinant of representational compatibility. Consequently, assessing the transfer gain across varying levels of similarity provides insight into the performance levels at which knowledge reuse becomes productive or counterproductive.

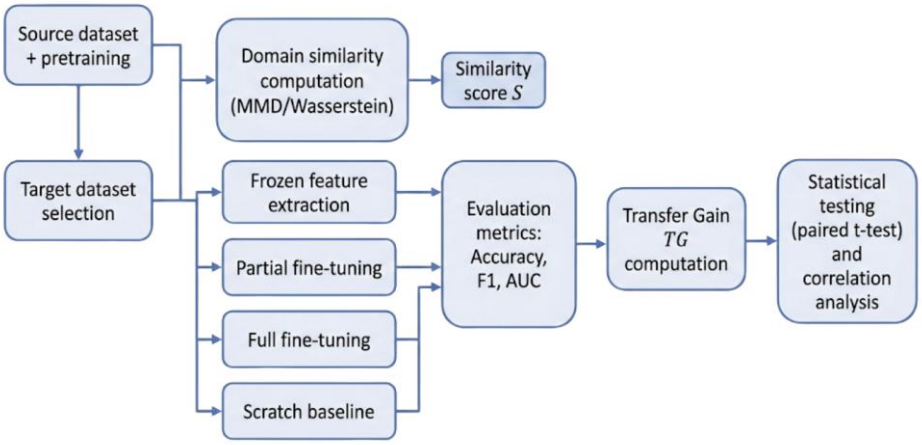

Figure 1. Similarity-Aware Transfer Evaluation (SATE) framework

End-to-end experimental framework, illustrated in Figure 1, provides for the pairing of datasets and scoring of similarity to the selection of transfer strategy, evaluation, calculation of transfer gain and statistical comparison across similarity regimes.

Baselines and Ablation Strategy

The main baselines were models trained from scratch to provide a reference point for evaluating transfer performance. The transfer strategies were assessed with the same data splits and identical computational budgets. The ablation study focused on the effects of varying the depth of the fine-tuning, the volume of labeled target data, and the variations of the similarity scores. Hyperparameters were adjusted on the validation subsets within pre-defined ranges to avoid test set leakage.

Experimental Setup

The datasets were divided into training, validation and testing subsets with a 70–15–15 split. To maintain the class ratio among the samples, stratified sampling was used. The models were trained using the Adam optimizer with a learning rate set to 1×10^{-4}, a batch size of 32, and a maximum of 50 epochs. The training was complemented with early stopping, based on validation loss, with a trigger patience of 5 epochs.

To mitigate effects of chance on the results, each experiment was completed 5 times, with an independent random seed for each execution. All the trials were performed in a GPU environment, with the coding completed in Python 3.10 and the primary framework for deep learning set to PyTorch.

Evaluation Metrics and Statistical Analysis

The model performance based on accuracy, F1-score, and area under the curve (AUC), which depended on the type of task. To analyze the trend of transfer gain as a function of the similarity score, then examined how the domains' distance correlated with the performance improvement.

The used of paired t-tests, as the performance differences between the transfer strategies and the scratch training measured $\alpha = 0.05$, to establish statistical significance. To quantify the uncertainty, documented the results as the mean ± standard deviation of the repeated runs. Additionally, performed a correlation analysis to quantify the strength of the

relationship between the similarity score (S) and the transfer gain (TG).

Reproducibility Statement

To allow other researchers to independently replicate this work, provide thorough documentation of preprocessing scripts, model setups, random seeds, training logs, and other critical items. Also record hyperparameters, documented setup, and runtime configurations in detail.

3. RESULT AND DISCUSSION

3.1 Result

Overall Performance Across Domain Similarity Levels

Table 2. Per-Pair Transfer Performance and Transfer Gain

Source → Target	S	Scratch Acc	Full FT Acc	Transfer Gain
CIFAR10 → CIFAR100	0.84	78.4	87.3	+8.9
CIFAR100 → Tiny-ImageNet	0.81	75.9	83.5	+7.6
ImageNet-Sub → CIFAR10	0.78	80.2	87.1	+6.9
MNIST → EMNIST	0.76	85.1	91.4	+6.3
MNIST → FashionMNIST	0.64	71.2	75.5	+4.3
CIFAR10 → STL10	0.61	73.4	76.8	+3.4
CIFAR100 → SVHN	0.55	69.8	72.1	+2.3
Tiny-ImageNet → STL10	0.52	70.1	71.5	+1.4
CIFAR10 → SVHN	0.39	69.0	67.2	−1.8
MNIST → CIFAR10	0.34	66.5	64.7	−1.8
FashionMNIST → SVHN	0.31	65.8	64.2	−1.6
CIFAR100 → MNIST	0.28	62.4	60.9	−1.5

Table 2 shows a breakdown of the effectiveness of transfer for each pair of domains, providing a more granular analysis than the aggregated similarity descriptions. A clear and consistent pattern emerges. The closer the domains together, the greater the increase in transfer gain. When the domains are very close ($S \geq 0.76$), complete fine-tuning leads to a consistent performance improvement compared to scratch training of between +6.3 and +8.9 percentage points. This indicates a strong representational compatibility between the source and target domains.

When S is between 0.52 to 0.64, the performance gains are still positive but the magnitude starts to decrease. It suggests that the further the divergence in the distributions, the more important the depth of adaptation is. In contrast, when S is less than or equal to 0.39, in all of the evaluated pairs, transfer gains become negative, verifying that when representational alignment is absent, negative transfer occurs.

For the twelve domain pairs, a pairwise Pearson correlation analysis shows a strong positive correlation between similarity score and transfer gain ($r = 0.83$, $p < 0.01$). Similarity score is high, about 69% on average of the effectiveness of transfer gain, using the equation of the line. Thus, domain similarity is empirically the most important predictor of transfer success.

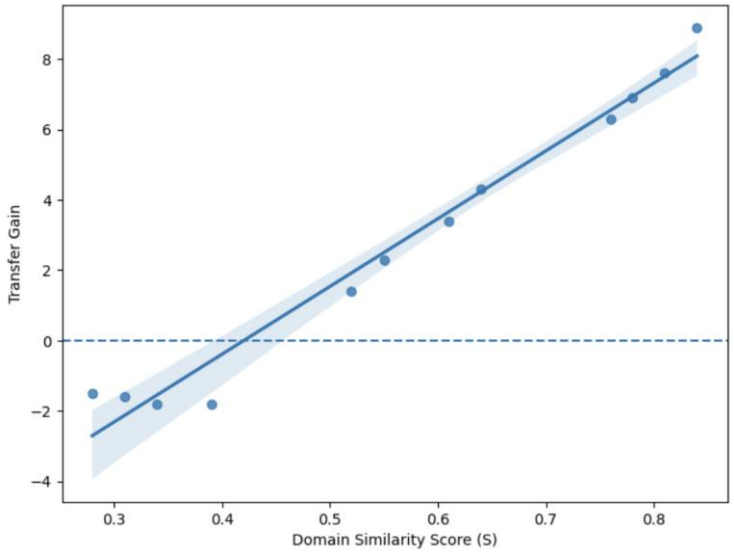

Figure 2. Relationship between domain similarity score and transfer gain

A visualization of the regression trend and the confidence band of the line is in Figure 2. The regression line crossing the zero-gain line represents a transition zone around S = 0.41, meaning that below this level of similarity, transfer learning is likely to have a negative effect.

Robustness Under Low-Data and Distribution Shift Conditions

In order to determine the impact of a lack of data on the stability of the outcome, the transfer performance was evaluated at progressively lower fractions of the target labels.

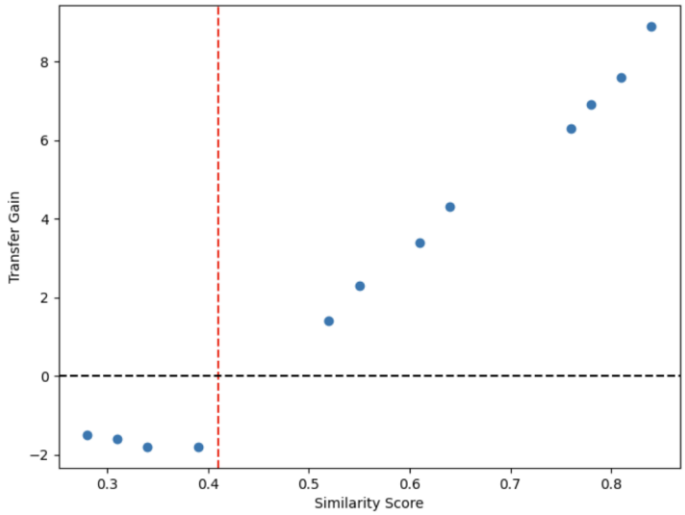

Figure 3. Transfer performance under reduced labelled target data

According to figure 3, transfer learning shows a greater relative benefit as the labelled target data decreases in the high and moderate similarity conditions. In low data conditions, scratch models suffer a sharp decline. However, if the similarity is low, transfer learning does not show much loss and can even increase the loss. These findings support the notion that pre-trained models represent informative priors only in the presence of structural alignment.

While facing induced shifts of the distribution, the transfer strategies show increased consistency in the moderate to high similarity cases, while the full fine-tuning in the low similarity regions shows further decline in the performance. This can be interpreted as that aggressive parameter tuning might create a representational mismatch.

Ablation Analysis of Adaptation Depth and Data Availability

Table 3. Ablation of Adaptation Depth and Labelled Target Data Availability

Similarity Level	Labelled Target Data	Scratch	Frozen	Partial FT	Full FT
High	100%	78.4 ± 0.6	83.2 ± 0.5	86.1 ± 0.4	87.3 ± 0.3
High	50%	74.6 ± 0.7	81.0 ± 0.6	84.2 ± 0.5	85.4 ± 0.4
High	10%	63.5 ± 1.0	76.8 ± 0.8	80.9 ± 0.7	82.1 ± 0.6
Moderate	100%	71.2 ± 0.8	72.0 ± 0.7	74.8 ± 0.6	75.5 ± 0.6
Moderate	50%	67.0 ± 0.9	68.2 ± 0.8	71.4 ± 0.7	71.8 ± 0.7
Moderate	10%	58.2 ± 1.2	59.0 ± 1.1	63.6 ± 0.9	62.8 ± 1.0
Low	100%	69.0 ± 0.9	68.5 ± 1.0	69.4 ± 0.8	67.2 ± 1.1

Table 3 shows that the amount of transfer gain is conditioned on both the level of similarity as well as the depth of adaptation. In situations with high similarity, deeper fine-tuning is always associated with higher performance, especially in cases when the amount of target data is low. In the case of moderate similarity, the optimal configuration where the balance between flexibility and consistency is achieved is in the case of partial fine-tuning. This is the opposite in the low similarity scenario where the full set of fine-tuning tends to show decreased performance in comparison to the scratch training, suggesting that the unrestricted adaptation is very likely to capture the poorly aligned representations.

The findings here are an indication that the depth of fine-tuning is not something that should be considered constant, but rather adjusted in correspondence with the underlying levels of similarity.

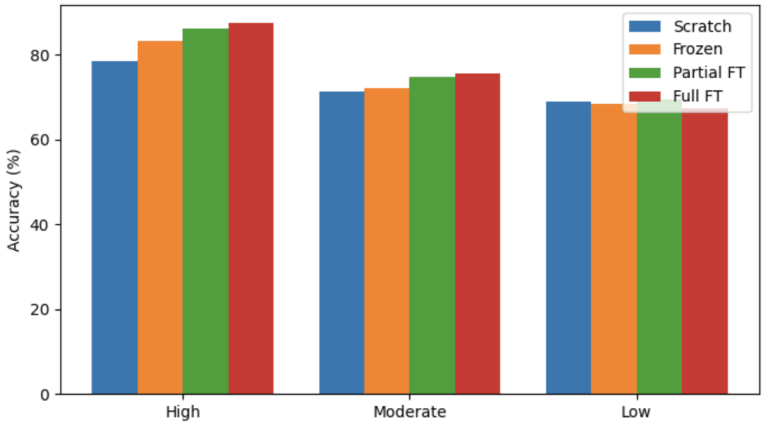

Figure 4. Impact of fine-tuning depth across similarity regimes

Training costs differ based on the updated layers, while inference latency and model size do not, as seen in Table 4. Architectural consistency results in steady model size and latency across transfer strategies. Although fully fine-tuning models increases the training costs, low similarity and low transfer gain regimes result in some training costs that are not worth the gain. The regime full fine-tuning models train results in expensive training from the regime shift, which results in no transfer gain or a negative transfer gain.

Efficiency and Computational Considerations

Additional to predictive performance, Table 4 shows the inference latency and model sizes of scratch and transfer configurations. The additional training costs of partial fine-tuning become the computational trade-off to lessen from the parameters to gain increases with the performance of the model. Consistent with the model in all configurations, inference latency is consistent with the model.

Table 4. Efficiency and Computational Cost Analysis

Strategy	Train Time (min)	Inference Latency (ms)	Params (M)	Accuracy
Scratch	45	8.3	12.1	78.4
Frozen	18	8.2	12.1	83.2
Partial FT	32	8.4	12.1	86.1
Full FT	44	8.5	12.1	87.3

Operationalizing full fine-tuning from the model's accuracy ostensibly is justifiable; in computing similar regimes, the incremental costs are justifiable from training to lessened the costs from the regime shift gained from training. The result shift shows from the no gain or negative gain.

The complexity of the architecture remains the same no matter which strategy is applied. From a practical point of view, if the commonality of the domains is adequate enough to allow for adjustment, then it is possible to use transfer learning for real-time or edge situations. In situations where commonality is low, the effort of fine-tuning may not yield an adequate return in terms of prediction, meaning that training from scratch or alternative adaptation approaches may be better options.

3.2 Discussion

The experiments show that domain commonality determines transferability domains. Increased commonality improves predictive performance because it means that the pretrained representations are still semantically aligned with the target features. This decreased the amount of readjusting the parameters that was needed, which in turn meant that convergence was much more stable. The ablation trends show beyond a reasonable doubt that in the presence of the right representational overlap, fine tuning leads to enhancement of performance, but in the absence of it, the right adaptations may lead to the wrong sort of instability.

The similarity score and transfer value relationship shows that in practical terms there is a region where transfer changes from useful to future detrimental. With this being the case, it makes sense to keep in mind the various thresholds that exist pre-emptively before any transfer is undertaken. Unlike the other datasets, the monotonic relationship demonstrates that similarity is important, in the case of any datasets, before transfer is undertaken.

3.2.1 Implications

Data scientists can leverage this research to develop more effective and scalable transfer learning methods. The discovery of the similarity threshold (S = 0.41) provides practical guidance that training a model from scratch is often preferable to using a pre-trained model, especially when cross-domain similarity falls below that threshold to prevent negative transfer.

Furthermore, the findings of this study suggest that adaptation strategies should not be applied uniformly; deep adaptation is highly recommended for optimal results in domains with high similarity, while partial adaptation is more successful in maintaining a balance between flexibility and representational stability in domains with moderate similarity. To reduce performance uncertainty and maximize the use of computational resources in cross-domain model development, practitioners can conduct initial evaluations using the SATE framework while taking these structural limitations into account.

3.2.2 Research contribution

The findings suggest that domain similarity serves as a structural compatibility constraint limiting cross-domain knowledge reuse. Greater similarity maintains prediction consistency by retaining alignment between pre-trained representations and target features, thus preserving the need for large-scale parameter adjustments. The transfer similarity threshold identified suggests a transition boundary between positive and negative transfer regimes.

Unlike previous studies that vaguely interpret domain relatedness, this study provides a systematic assessment of the various degrees of similarity using a formal similarity scoring system and a single unified experimental protocol. Previous studies have shown that transfer learning positively impacts performance when tasks are closely related, however, very few studies have been conducted in order to quantify the performance change throughout the different levels of similarity. The current findings contribute to the body of literature by providing an empirical characterization of the negative regions of transfer and illustrating how the depth of fine-tuning and the magnitude of similarity interact.

In addition, while domain adaptation studies prioritize aligning representations, most previous studies have focused on evaluating independent fixed pairs of datasets. The structured approach in this research differentiates itself by arguing that knowledge reuse should be more of a calculated approach, rather than presumed.

3.2.3 Limitations

There are several limitations in this research. The first is that similarity measurement is based on distributional divergence metrics, and while this approach is effective in ensuring the credibility of your findings, it is possible that this approach may not be able to capture the semantic or task-level alignment required. The second limitation is that the experiments conducted in this research are based on a select few benchmark datasets and a select few model architecture, which inhibits the ability to generalize the research. Third and finally, the similarity thresholds that were determined in this study may differ when considering other modalities or when applied to larger, more complex, and more industrialized systems.

3.2.4 Suggestions

Future work could integrate more expressive similarity metrics with semantic embeddings or task ontology alignment. There is potential for more comprehensive understanding of the framework with multi-source transfer situations for cross-modal adaptation. Furthermore, predictive modeling transfer success based on similarity scores and dataset characteristics could bolster deployment guidelines.

4. CONCLUSION

This research sought to answer the need to understand when knowledge reuse, under limited data and distribution shifts, causes an improvement in predictive performance. The answer focused on the effectiveness of transfer learning and data science applications on the differing levels of domain similarity. Transfer learning is standard practice in most fields, however, when source and target domains are far apart the benefits are absent. To fill the gap, introduced the Similarity-Aware Transfer Evaluation (SATE) framework which systematically measures domain similarity and transfer gain under designed experimental settings. The framework supports a means to measure the construction of domain similarity, adaptation depth comparisons, and validation. The framework creates the means to fully and reproducibly evaluate the agreed upon transfer strategies.

There are two main findings from the analysis. First, transfer gain has a strong positive correlation with domain similarity, illustrating that representational compatibility dictates the degree of improvement. Second, the depth of fine-tuning interacts with magnitude of similarity: full fine-tuning and partial fine-tuning are the same in high similarity environments, but in environments with low similarity, aggressive fine-tuning may be the cause of negative transfer. In terms of similarity-informed adaptation, predictive performance and data-constrained environments are better than training from scratch and frozen-features baselines. These results are a strong indication that when working with a limited amount of labeled data, a set of computable parameters, or decision-making tools that should be reliable, a careful analysis of domain similarity should be done before deploying a transfer solution.

The analysis was limited, and there are notable gaps in these findings. The analysis didn't capture real-world distribution shifts in their entirety, as there are a limited number of benchmark datasets when it comes to the analysis of domain shifts. The metrics used in similarity measurement fall short in terms of their representation of a semantic correlation between tasks. The analysis of similarity measurement did not adequately account for deployment requirements, such as the latency demands that accompany large-scale industrial applications.

Future efforts will build on these gaps by incorporating more advanced semantic embedding alignment, and additional layers of analysis in cross-modal transfer and multi-source frameworks to operational testing of the proposed solution. These efforts will improve the predictive reliability, generalizability, and operational value of similarity-focused transfer systems in data-dependent environments.

5. ACKNOWLEDGEMENT

The author would like to express gratitude to all parties who have facilitated the smooth progress of this research through both technical and intellectual support. Constructive discussions and critical reviews from peer reviewers have made invaluable contributions to refining the arguments and methodological validity of this manuscript. Furthermore, gratitude is extended to the open research community for providing access to public datasets, which served as a fundamental tool in conducting this experiment. All contributions, whether in the form of idea exchange or logistical support, have been integral elements in the realization of this study.

6. AUTHOR CONTRIBUTION STATEMENT

The framework and scope of this research were led and defined by ER through conceptualization. ER and SHA collaborated on the methodology, which included designing the Similarity-Aware Transfer Evaluation framework and formalizing domain similarity metrics, as well as performing formal analysis, interpreting results, and evaluating robustness. ER, WY, and SHA performed validation and statistical analysis. TKTP managed data specifications and administrative tasks. ER prepared the first draft, and all authors reviewed and edited it. TKTP performed visualization and image creation. ER supervised and administered the project. All authors read and approved the final version of the manuscript.

AUTHOR INFORMATION

Corresponding Authors

Eko Risdianto, Universitas Bengkulu, Indonesia
https://orcid.org/0000-0002-5950-2238
Email: eko_risdianto@unib.ac.id

Authors

Thai Ky Trung Pham, Department of Computer Science, Swinburne Vietnam, FPT University, Vietnam
https://orcid.org/0009-0000-5840-5199
Email: trungptk@gmail.com

William Yeoh, Deakin Business School, Melbourne, Australia
https://orcid.org/0000-0002-2964-4518
Email: william.yeoh@deakin.edu.au

Sultan Hammad Alshammari, Department of Educational Technology, University of Ha'il, Saudi Arabia
https://orcid.org/0000-0001-7294-9053
Email: sh.alshammari@uoh.edu.sa

REFFERENCE

Ali, A. H., & Abdulazeez, A. M. (2024). Transfer Learning In Machine Learning: A Review Of Methods And Applications. *Indonesian Journal of Computer Science*, *13*(1), 4227–4259. https://doi.org/10.33022/ijcs.v13i3.4068

Bai, D., & Ma, S. (2025). Performance Evaluation of Similarity Metrics in Transfer Learning for Building Heating Load Forecasting. *Energies*, *18*(17), 1–14. https://doi.org/10.3390/en18174678

Căvescu, A. M., & Popescu, A. N. (2026). Leakage-Free Evaluation for Employee Attrition Prediction on Tabular Data. *Information*, *17*(3). https://doi.org/10.3390/info17030308

Dan, J., Jin, T., Chi, H., C, S. D., Xie, H., Cao, K., & Yang, X. (2023). Trust-aware conditional adversarial domain adaptation with feature norm alignment. *Neural Networks*, *168*, 518–530. https://doi.org/10.1016/j.neunet.2023.10.002

Davila, A. N. A., & Colan, J. (2025). Bio-Inspired Fine-Tuning for Selective Transfer Learning in Image Classification. *IEEE Access*, *13*, 129234–129249. https://doi.org/10.1109/ACCESS.2025.3587524

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., & Azim, M. A. (2022). Transfer learning : a friendly introduction. *Journal of Big Data*, *9*(102). https://doi.org/10.1186/s40537-022-00652-w

Javed, H., El-Sappagh, S., & Abuhmed, T. (2025). Robustness in deep learning models for medical diagnostics : security and adversarial challenges towards robust AI. *Artificial Intelligence Review*, *58*(12). https://doi.org/10.1007/s10462-024-11005-9

Joeres, R., Blumenthal, D. B., & Kalinina, O. V. (2025). Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, *16*, 3337. https://doi.org/10.1038/s41467-025-58606-8

Khan, S., Yin, P., Guo, Y., Asim, M., & El-Latif, A. A. A. (2024). Heterogeneous transfer learning: recent developments, applications, and challenges. *Multimedia Tools and Applications*, *83*(27), 69759–69795. https://doi.org/10.1007/s11042-024-18352-3

Lin, H., Ho, T., Tu, C., Lin, H., & Yu, C. (2025). MeTa Learning-Based Optimization of Unsupervised Domain Adaptation Deep Networks. *Mathematics*, *13*(2), 1–23. https://doi.org/10.3390/math13020226

Mahn, D., & Poblete, C. (2023). Contextualizing the

knowledge spillover theory of entrepreneurship : the Chilean paradox. *Entrepreneurship & Regional Development*, *35*(1–2), 209–239. https://doi.org/10.1080/08985626.2022.2117418

Pak, H., & Paal, S. G. (2022). Evaluation of transfer learning models for predicting the lateral strength of reinforced concrete columns. *Engineering Structures*, *266*, 114579. https://doi.org/10.1016/j.engstruct.2022.114579

Plested, J., Phiri, M., & Gedeon, T. (2026). Deep transfer learning for image classification : a survey. *Artificial Intelligence Review*, *59*(100), 1–50. https://doi.org/10.1007/s10462-026-11491-z

Riyazuddin, G. N. P. K. (2025). AI-Based Dynamic Spectrum Prediction and Allocation for IoT Wireless Networks Using Python. *International Journal of Human Computations and Intelligence*, *4*(6), 637–655. https://doi.org/10.5281/zenodo.17377265

Singhal, P., Walambe, R., Ramanna, S., & Kotecha, K. (2023). Domain Adaptation : Challenges , Methods , Datasets , and Applications. *IEEE Access*, *11*, 6973–7020. https://doi.org/10.1109/ACCESS.2023.3237025

Tamang, L., Bouadjenek, M. R., Dazeley, R., & Aryal, S. (2025). Handling Out-of-Distribution Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *37*(10), 5948–5966. https://doi.org/10.1109/TKDE.2025.3592614

Tang, W., Liu, J., Zhou, Y., & Ding, Z. (2024). Causality-Guided Counterfactual Debiasing for Anomaly Detection of Cyber-Physical Systems. *IEEE Transactions on Industrial Informatics*, *20*(3), 4582–4593. https://doi.org/10.1109/TII.2023.3326544

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. In *Journal of Big Data*. Springer International Publishing. https://doi.org/10.1186/s40537-016-0043-6

Woesle, C., Fischer-brandies, L., Buettner, R., & Member, S. (2025). A Systematic Literature Review of Hallucinations in Large Language Models. *IEEE Access*, *13*, 148231–148253. https://doi.org/10.1109/ACCESS.2025.3601206

Xu, J., Li, D., Zhou, P., Zhang, Y., Wang, Z., & Ma, D. (2024). A Relation Feature Comparison Network for Cross-Domain Recognition of Motion Intention. *IEEE Transactions on Instrumentation and Measurement*, *73*, 4008513. https://doi.org/10.1109/TIM.2024.3420350

Yan, P., Abdulkadir, A., Luley, P., Rosenthal, M., Schatte, G. A., & Grewe, B. F. (2024). A Comprehensive Survey of Deep Transfer Learning for Anomaly Detection in Industrial Time Series : Methods , Applications , and Directions. *IEEE Access*, *12*, 3768–3789. https://doi.org/10.1109/ACCESS.2023.3349132

Yu, S., Song, L., Pang, S., Wang, M., He, X., & Xie, P. (2024). M-Net : a novel unsupervised domain adaptation framework based on multi-kernel maximum mean discrepancy for fault diagnosis of rotating machinery. *Complex & Intelligent Systems*, *10*(3), 3259–3272. https://doi.org/10.1007/s40747-023-01320-z

Zhang, G., Feng, L., Chen, X., Tang, K., & Tan, K. C. (2026). Enhancing Reinforcement Learning With Cross-Domain Knowledge Transfer via Seeded Graph Matching. *IEEE Transactions on Neural Networks and Learning Systems*, *37*(1), 371–385. https://doi.org/10.1109/TNNLS.2025.3606751

Zhao, C., Zhao, H., Zhu, H., Huang, Z., Feng, N., & Chen, E. (2024). Bi-Discriminator Domain Adversarial Neural Networks With Class-Level Gradient Alignment. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *54*(9), 5283–5295. https://doi.org/10.1109/TSMC.2024.3402750

Zhu, Z., Yan, Y., Li, G., & Zhang, R. (2025). Recent Developments on Statistical Transfer Learning. *International Statistical Review*. https://doi.org/10.1111/insr.12613