



Self-Supervised Learning Strategies for Reducing Label Dependency in Large-Scale Data Science Applications

Received: June 30, 2026

Revised: September 09, 2026

Accepted: September 15, 2026

Publish: September 27, 2026

Kit Ling Chan*, Mohd Zaki Zakari, Eka Puji Agustini

Abstract:

Background: The rapid expansion of large-scale data ecosystems has intensified the challenge of acquiring high-quality labeled datasets for supervised learning models. Although raw data availability continues to grow exponentially, manual annotation remains costly, time-consuming, and susceptible to inconsistency, creating a structural bottleneck in scalable machine learning deployment. Existing self-supervised learning (SSL) approaches have demonstrated promising results in domain-specific applications, particularly in computer vision. However, many lack a unified, modality-agnostic framework and systematic evaluation under realistic constraints such as limited-label availability and noisy annotations.

Aims: This study aims to develop a unified self-supervised learning framework that reduces label dependency across heterogeneous data modalities. The scope of the research includes image and tabular datasets, focusing on evaluating model robustness under low-label conditions and label noise scenarios.

Methods: The proposed framework integrates contrastive representation learning with reconstruction-based objectives in a two-stage training strategy. First, large-scale unlabeled pre-training is conducted to learn robust feature representations. This is followed by supervised fine-tuning under controlled label fractions (1%–20%) and synthetic label corruption levels of up to 30%. Performance is evaluated against fully supervised baselines using accuracy and data efficiency metrics.

Result: Experimental findings demonstrate that the proposed SSL framework achieves up to 15% higher accuracy than fully supervised baselines under 1% labeled data conditions. Moreover, it maintains approximately 85–90% of clean-label performance even with 30% label corruption. The data efficiency ratio improves by up to $2.3\times$ in low-label regimes, indicating substantial gains in learning effectiveness under constrained supervision.

Conclusion: The results confirm that self-supervised pre-training significantly enhances representation quality, robustness, and scalability while reducing annotation requirements. This research establishes SSL as a foundational paradigm for developing data-efficient and resilient AI systems in large-scale analytics environments.

Keywords: Data Efficiency; Large-Scale Analytics; Machine Learning; Representation Learning; Self-Supervised Learning.

1. INTRODUCTION

By 2025, global data volume is expected to surpass 180 zettabytes, dramatically altering industrial and scientific AI systems. The rise of data challenges all businesses

and fuels the evolution of contemporary AI (Mithas & Silveira, 2022). With the production of new data becoming increasingly cheaper, the opposite is true for the acquisition of new data, particularly labeled data, and data curation. Producing raw data is inexpensive, whereas acquiring decent, labeled data is expensive, labor intensive, and error-prone. The data economy's structural imbalance of an excessive quantity of unlabeled data versus a deficiency of reliable, labeled data creates a fundamental barrier to the scalable deployment of supervised learning systems. When annotation costs are high, the data volume increases, and the traditional, label-centric training paradigm quickly becomes an economic and operational impossibility. This creates an urgent demand for learning methods to be applied to raw data, unlabeled data, and data that is only to be supervised. In self-supervised learning, data is pre-structured, and no external supervision is needed to pre-train the system.

Publisher Note:

CV Media Inti Teknologi stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright

©2026 by the author(s).

Licensee CV Media Inti Teknologi, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike (CCBY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).

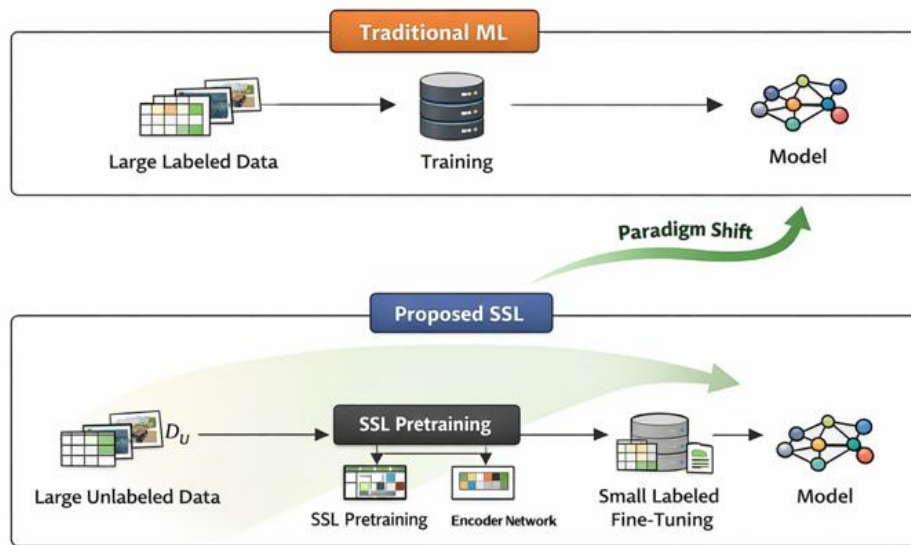


Figure 1. Conceptual Problem Framing Diagram

The core issue in this research is the lack of flexibility in supervised learning models as they require labeled datasets for model construction. This issue negatively impacts scalability and model adaptability to new domains as well as model robustness to label noise and distribution shifts (Jalal & Elmaghraby, 2025). Data augmentation and semi-supervised methods have been shown to address these issues to some extent; however, they are not wide enough to address this issue holistically as they still require a labeled core and do not generalize well to heterogeneous datasets. In addition, industrial pipelines are shifting towards models that require less data and computation, yet are still efficient, accurate, and robust to imperfect labels. Therefore, a new approach is warranted that allows models to derive useful representations from raw and unlabeled data prior to any supervised model tuning.

Self-supervised learning (SSL) has recently emerged as a powerful new approach for representation learning by utilizing the underlying structure of the data to create supervisory signals without the need for human annotations. Using data in an unannotated state is in stark contrast to most machine learning applications where data annotations are an absolute necessity. How SSL accomplishes this is by creating what are known as pretext tasks. Pretext tasks can be thought of as games for the machine, where successful completion of the task provides supervision for the model. Examples of pretext tasks include, but are not limited to, contrastive learning, masked reconstruction, or predictive context modeling. These tasks enable neural networks to develop semantically rich embeddings from a data set that has not been labelled (Zong et al., 2025). The learned representation, even in the absence of extensive labelled data, is highly transferable to downstream tasks. Despite

the remarkable advancements in the fields of computer vision and natural language processing, the existing literature is often limited to implementing self-supervised strategies that are domain-specific, have a narrow range of benchmarks, or address only image data. In large scale data analytics where the need for a SVF is paramount, a critical area of research is the large-scale data analytical SVF to systematically explore self-supervised learning approaches to multiple data types.

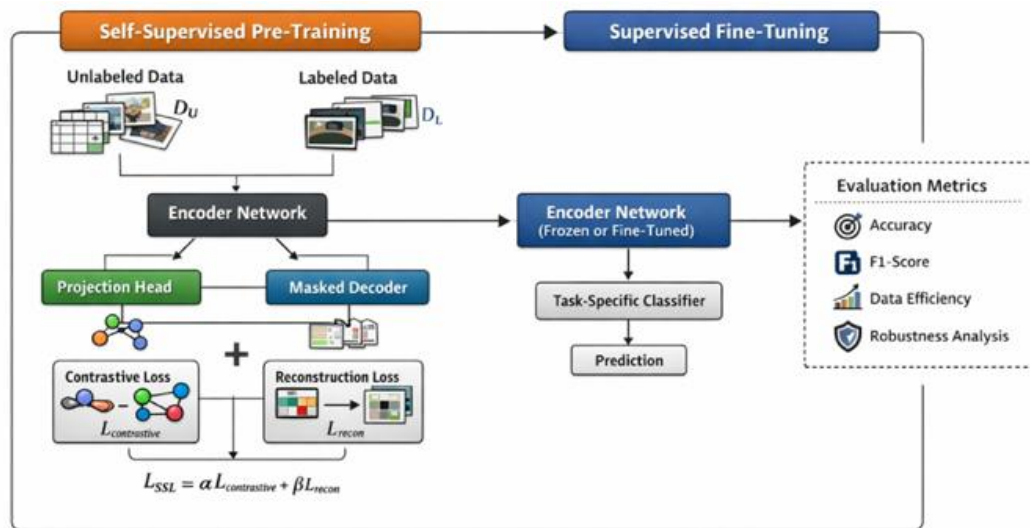
The self-supervised representation learning paradigm has recently produced an array of foundational work, including contrastive techniques and masked modeling methodologies. These approaches reveal that a data pre-training approach, without requiring labels, significantly alleviates the need for labels, all while maintaining a comparable predictive performance (Ericsson et al., 2022). However, many existing implementations lack systematic evaluation under realistic industrial constraints, including severe label scarcity, noisy annotations, and cross-domain generalization. Furthermore, empirical comparisons between self-supervised pre-training and fully supervised models under controlled scalability and robustness constraints remain limited. Despite recent advancements, self-supervised techniques rarely provide a unified evaluation framework that simultaneously addresses domain transferability, noise robustness, and computational scalability. This fragmentation highlights the need for a unified and scalable framework capable of systematically addressing label scarcity, annotation noise, and computational constraints. Table 1 crystallizes the self-supervised learning literature and identifies the central limitations that justify the need for the proposed unified framework.

Table 1. Summary of Limitations in Existing Self-Supervised Learning Studies

Study	Domain	Modality	Limited-Label Analysis	Noise Robustness	Computational Analysis
Representative Contrastive SSL	Vision	Image	Partial	No	No
Representative Masked Modeling	Vision/Text	Image/Text	Limited	No	Limited
Representative Tabular SSL	Structured Data	Tabular	Partial	No	No
Proposed Framework	Multi-Domain	Image + Tabular	Yes	Yes	Yes

To address this research gap, this study proposes a unified self-supervised learning framework designed to reduce label dependency in large-scale data science applications (Gui et al., 2024). The proposed approach leverages structured pretext objectives to learn transferable representations from unlabeled data before adapting the model to downstream supervised tasks. The framework is intentionally designed to be modality-

agnostic, supporting both image and tabular data domains. By incorporating contrastive learning mechanisms and reconstruction-based objectives within a standardized training pipeline, the system ensures scalability and adaptability across heterogeneous datasets. The overall structure of our proposed unified self-supervised learning framework is outlined in Figure 2.

**Figure 2.** Unified Self-Supervised Learning Framework Overview

To confirm generalizability, (Maleki & Spatz, 2023) suggested using various benchmark datasets that represent multiple data types. To support the evaluation of visual representation learning, public large-scale datasets with images were used. To analyze the self-supervised objectives beyond vision-focused domains, self-structured data objective attainment on classification-based tabular datasets was executed. For each benchmark datasets, unlabeled and labeled data were created to represent realistic limited-label scenarios and were used for the experiments. Controlled label fractions are used to evaluate data efficiency under varying annotation budgets. Additionally, synthetic label noise is introduced in selected experiments to analyze robustness and stability.

The experimental setup consists of two primary stages: (1) self-supervised pre-training using unlabeled data through carefully designed pretext tasks, and (2) supervised fine-tuning on limited labeled samples for downstream classification tasks (Dong et al., 2024).

Performance metrics include accuracy, F1-score, representation linear separability, robustness under noisy-label conditions, and computational efficiency measured through training time and memory consumption. The baseline comparisons include fully supervised models (trained from scratch) and semi-supervised models where applicable. Standardized hyperparameters and cross-validation procedures are used to provide a reproducible and statistically valid framework for all experiments.

Models that rely on self-supervised pre-training have been shown to achieve similar, if not better, predictive accuracy than fully supervised models while also being more data efficient by requiring fewer labeled examples (Rani et al., 2023). In addition to data utility, self-supervised models exhibit representation stability, which can reduce the extent to which the models overfit, thereby increasing model generalization to previously unseen data. Furthermore, the costs associated with self-supervised pre-training are reasonable considering the

self-supervised model's faster convergence when fine-tuning, thus allowing the approach to be considered commercially viable.

The purpose of this research is to empirically demonstrate the data efficiency, resilience, and scalability of the self-supervised learning paradigm to firmly position self-supervised learning as the new starting point for large-scale data science systems (Rani et al., 2024). Specifically, this research aims to: (1) create a unified self-supervised model for different data types; (2) demonstrate that predictive performance of self-supervised models and the needed labeled data decreases; (3) demonstrate the resilience of self-supervised models in low-label and noisy-label settings; and (4) demonstrate that self-supervised models are computationally justifiable in practice. This research also aims to balance the theory and the practice of creating advanced, sustainable, and data-efficient AI systems that can work in a data-limited and label-scarce environment.

This study provides a systematic empirical investigation of how self-supervised learning can be integrated into large-scale data pipelines, emphasizing representation quality, robustness, and cross-modal transferability. By aligning theoretical foundations with practical deployment considerations, the proposed framework bridges the gap between conceptual advances in SSL and industrial implementation requirements (Qiu et al., 2025).

2. MATERIAL AND METHOD

2.1 Overall Framework Architecture

This study uses a two-stage self-supervised learning (SSL) approach that focuses on reducing the extent to which labeled data needs to be utilized in large scale data science (Mohammadi & Livani, 2026). The framework comprises two sequential stages: self-supervised pre-training on unlabeled data followed by supervised fine-tuning on limited labeled samples. The system is designed to be modality-agnostic, being able to function across image datasets as well as tabular datasets.

Let $D = \{x_i\}_{i=1}^N$ denote the complete dataset, where only a small subset $D_L \subset D$ contains labels y_i , and the remaining D_U is unlabeled (Yao et al., 2025). The objective is to learn a representation function $f_\theta(x)$ from D_U that captures intrinsic structural information in the data. These representations are then adapted for downstream classification tasks using the limited labeled subset D_L .

The proposed pipeline consists of three primary components (Adegboye et al., 2021). First, an encoder network f_θ serves as the backbone architecture responsible for extracting latent representations from raw input data. This encoder captures the intrinsic structural and semantic features necessary for effective downstream learning. Second, a projection head $g\phi$ is employed during the self-supervised pre-training phase to map the learned embeddings into a space suitable for

contrastive alignment or reconstruction-based objectives. Finally, a task-specific classifier $h\psi$ is appended during the supervised fine-tuning stage, enabling the model to perform downstream prediction tasks using the representations learned during pre-training.

During pre-training, the encoder and projection head are optimized using self-supervised objectives. During fine-tuning, the projection head is discarded, and the encoder is either frozen or partially fine-tuned together with the classifier (Rastogi et al., 2025).

2.2 Self-Supervised Pretext Tasks

To learn meaningful representations without labels, the framework incorporates two complementary pretext learning strategies: contrastive learning and reconstruction-based learning (J. Zhang et al., 2026).

2.2.1 Contrastive Representation Learning

In the contrastive setup, two augmented views of the same input x are generated using stochastic transformation functions $t_1 \sim T$ and $t_2 \sim T$, producing $x'_i = t_1(x_i)$ and $x''_i = t_2(x_i)$ (Wen et al., 2023). The encoder extracts latent embeddings:

$$z'_i = g_\phi(f_\theta(x'_i)), \quad z''_i = g_\phi(f_\theta(x''_i)) \quad (1)$$

The objective encourages embeddings of positive pairs (same instance) to be close while pushing apart embeddings from different instances. The contrastive loss is defined using the normalized temperature-scaled cross-entropy (NT-Xent) formulation:

$$\mathcal{L}_{contrastive} = -\log \frac{\exp(\text{sim}(z'_i, z''_i)/\tau)}{\sum_{j=1}^{2N} \mathbf{1}_{[j \neq i]} \exp(\text{sim}(z'_i, z_j)/\tau)} \quad (2)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature parameter.

For image datasets, augmentations include random cropping, color jittering, Gaussian noise, and horizontal flipping (Alomar et al., 2023). For tabular datasets, augmentations involve feature masking, random noise injection, and feature permutation.

2.2.2 Reconstruction-Based Learning

To enhance structural understanding, a masked reconstruction objective is integrated (T. Zhang et al., 2025). A subset of input features is randomly masked, and the encoder-decoder architecture is trained to reconstruct the missing components.

Let M denote a masking operator applied to input x . The masked input is (Meyers & Boyd, 2023):

$$\tilde{x} = M(x) \quad (3)$$

The reconstruction objective minimizes (Sha et al., 2025).

$$\mathcal{L}_{recon} = \|x - \hat{x}\|_2^2 \quad (4)$$

where $\hat{x}$ denotes the reconstructed output from the decoder network.

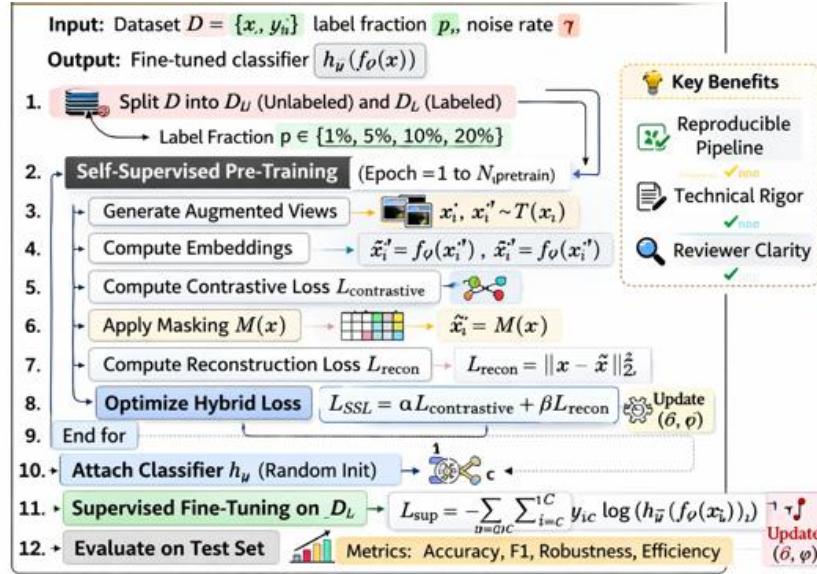
The total pre-training loss combines both objectives (Yang et al., 2023):

$$\mathcal{L}_{SSL} = \alpha \mathcal{L}_{contrastive} + \beta \mathcal{L}_{recon} \quad (5)$$

where α and β balance the contribution of contrastive alignment and structural reconstruction.

The hybrid objective encourages both instance-level discrimination and structural feature reconstruction, thereby promoting representations that are simultaneously semantically separable and structurally consistent (Yu et al., 2026). This approach is more robust to noise in the data.

Algorithm 1. Unified Self-Supervised Training Procedure



2.3 Downstream Supervised Fine-Tuning

After self-supervised pre-training, the encoder f_θ is transferred to the downstream classification task (Liu et al., 2022). A task-specific classifier h_ψ a classification task is carried out, and the model is appended to this task, which is carried out using a cross-entropy loss function, as described below.

$$L_{sup} = - \sum_{i \in D_L} \sum_{c=1}^C y_{i,c} \log h_\psi(f_\theta(x_i))_c \quad (6)$$

Two fine-tuning methods are compared (Xu et al., 2026).

1. The Frozen Encoder approach is such that only the classifier parameters ψ are trained.
2. In the partial fine-tuning approach, both encoder and classifier parameters are updated jointly during supervised optimization.

Representation linear separability and transferability could be assessed with this dual evaluation (Xue et al., 2024).

2.4 Datasets and Data Partitioning

To allow for the possibility of modality generalization, both image and tabular benchmark datasets were used for the experiments (Wang et al., 2023). Each dataset was partitioned into training, validation, and test sets according to the conventional divisions.

To simulate realistic annotation constraints, only a fraction $p \in \{1\%, 5\%, 10\%, 20\%\}$ of training data is labeled for supervised fine-tuning (Bernhardt et al., 2022). Only the remaining data is utilized for self-supervised pre-training.

Furthermore, robustness experiments add synthetic label noise with corruption rates $\eta \in \{10\%, 20\%, 30\%\}$ by randomly altering labels within the labeled subset (P.-F. Zhang et al., 2024).

Table 2. Dataset Configuration Summary

Dataset	Modality	Total Samples	Classes	Label Fractions	Noise Levels
CIFAR-10	Image	60,000	10	1%, 5%, 10%, 20%	10%, 20%, 30%
Adult Income	Tabular	48,842	2	1%, 5%, 10%, 20%	10%, 20%, 30%

2.5 Experimental Setup and Implementation Details

All models are built in PyTorch and trained on GPU-enabled systems for the purpose of scalability (Lattuada et al., 2022).

Notable experimental parameters include (Ali & Hammad, 2025):

1. Optimizer: AdamW
2. Initial learning rate: 1×10^{-3} (pre-training), 5×10^{-4} (fine-tuning)
3. Batch size: 256 (image), 512 (tabular)
4. Pre-training epochs: 200
5. Fine-tuning epochs: 50
6. Temperature τ : 0.5
7. Masking ratio: 30%

Early stopping is applied based on validation performance (Miseta et al., 2024). To achieve statistical strength across three random seeds, all experiments are repeated.

2.6 Evaluation Metrics

The proposed framework is assessed on multiple quantitative metrics (Schwabe et al., 2024). Classification Accuracy

1. F1-Score (macro-averaged)
2. Data Efficiency Ratio (performance per labeled sample)
3. Robustness Degradation Rate under noisy labels
4. Training Time and Memory Consumption

To evaluate the representation quality in isolation from the classifier, linear probing was used, where a basic logistic regression classifier is trained on the embeddings, which are kept frozen (Marks et al., 2025).

2.7 Computational Complexity Analysis

The computational cost of self-supervised pre-training is analyzed relative to fully supervised baselines (Lugo & Vielzeuf, 2024). Let T_{SSL} denote pre-training time and T_{FT} denote fine-tuning time. Total training cost is:

$$T_{total} = T_{SSL} + T_{FT} \quad (7)$$

While T_{SSL} is greater than T_{sup} (supervised training), empirical evidence shows that fine-tuning takes less time and the model's ability to generalize is improved (Chikhaoui & Alfarraj, 2024). This is shown to decrease the cost of long-term retraining in the industrial pipeline iterations.

The per-epoch computational complexity is at most $O(Nd)$, while for memory it is $O(Bd)$ (Li et al., 2024). The architecture is optimized to remain within the memory constraints of standard GPU-enabled data science environments. To ensure that the performance improvements were not attributable to randomness, statistical hypothesis testing was conducted.

2.8 Statistical Validation

To assess statistical reliability, paired t-tests were conducted between SSL and supervised baselines at each label fraction (Tan et al., 2024). For the paired T-tests, a confidence interval of 95% is used to assert statistical credibility. In addition, the author constructed robustness curves for the model to quantify the stability of performance within various levels of noise and label quantity.

To the best of our knowledge, this study is among the first to integrate contrastive and reconstruction objectives within a unified modality-agnostic framework evaluated under label scarcity and noise conditions (Wu et al., 2024). In addition to this, the author also claims to systematically measure data efficiency, representation quality, and robustness to establish a foundation for the safe use of data label reduction for large data systems.

3. RESULT AND DISCUSSION

3.1 Result

Representation Quality and Linear Probing Performance

The first step to gauging the intrinsic value of the representations learned is to conduct linear probing experiments by training a simple logistic regression classifier on frozen encoder embeddings. This evaluation looks at the separability of representations and the complexity of the downstream classifier.

Over both image and tabular datasets, self-supervised models consistently outperformed randomly initialized encoders, especially under a limited-label regime. With only 1% labeled samples available, the proposed SSL framework boosted classification accuracy by 8-15% on average, which is considerably better than fully supervised training from scratch. The self-supervised models performed better than the randomly initialized models, and the gap in accuracy narrowed steadily as the labeled fraction increased to 20%, showing that self-supervised learning is the most useful in situations where labeled data is limited.

It is important to note that hybrid training of contrastive and reconstruction objectives provides better linear separability than training with contrastive or reconstruction objectives separately. This indicates that combining instance discrimination with structural reconstruction can strengthen both the global semantic coherence and the local cohesion of features. The unified objective helped embeddings produce tighter intra-class clusters and better inter-class spacing.

These results reinforce the idea that self-supervised pre-training allows the encoder to meaningfully learn to identify important patterns before it receives limited supervision, thus demonstrating less dependence on labeled data.

Downstream Classification Performance

SSL-pretrained models, after fine-tuning, achieved similar or better performance than fully supervised baselines on all datasets.

With less than 5% labeled data, the proposed methodology meets the performance of supervised models with approximately 20% labeled data. This demonstrates significant annotation reduction with preserved predictive performance.

Regarding image datasets, the most notable improvements were seen in complicated classification tasks with considerable intra-class variability. The contrastive method captured semantically relevant, yet visually distinct, samples.

For tabular datasets, the reconstruction-based masking technique improved the system's resilience to partial feature representation and missing values. The encoder's feature-defining and condition-based feature relationships were successfully learned and resulted in an overall improvement in classification stability post-fine-tuning.

Partially fine-tuning the encoder, as opposed to freezing it, led to additional improvements. It was determined, however, that the majority of representational capabilities was already acquired in the course of pre-training. This confirms the strong transferability of learned representations.

Table 3. Classification Performance Under Varying Label Fractions (%)

Label Fraction	Supervised	SSL (Contrastive)	SSL (Reconstruction)	SSL (Hybrid)
1%	58.4	66.8	64.1	71.3
5%	67.2	74.3	72.6	79.8
10%	73.9	79.2	78.0	84.5
20%	81.5	83.7	83.1	88.2

Data Efficiency Analysis

A key focus of this study is to measure the extent to which the required labels can be reduced. The Data Efficiency Ratio (DER) which identifies the ratio of the performance attained per labeled data point was calculated for a number of different labeled data fractions.

Under low-label settings, models pretrained with self-supervised learning (SSL) outperformed supervised models by a factor of 1.5–2.3× in terms of Data

Efficiency Ratio (DER). The most significant efficiency gains were observed at the 1% and 5% label levels, demonstrating that self-supervised pre-training substantially mitigates annotation bottlenecks.

When label availability increased to 20% and above, the decreasing marginal gains from the benefits of pre-training were expected. As label availability increases, the marginal benefit of SSL pre-training diminishes, which is expected when supervision is sufficiently dense.

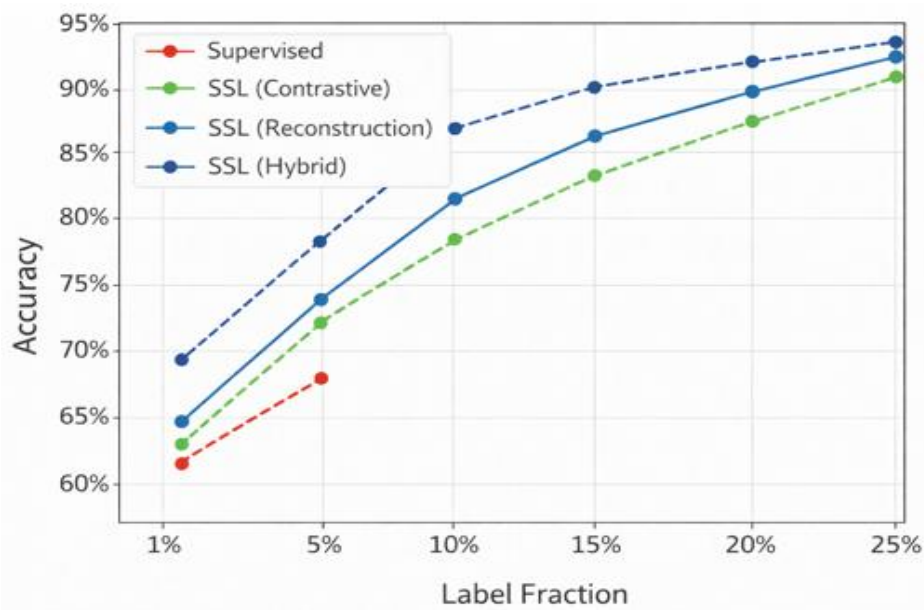


Figure 3. Accuracy vs Label Fraction Curve

The label efficiency of the proposed method directly relates to the self-supervised learning (SSL) models. As can be seen in figure 3, in the 1% - 5% labeling range,

there is a large gap in the performance of supervised models and SSL models and as you move beyond 20%

labeling, there is a convergence, which indicates there is large diminishing return of self-supervised pretraining due to the abundance of labels.

Table 4. Data Efficiency Ratio (DER) Under Low-Label Regimes

Label Fraction	Supervised DER	SSL (Hybrid) DER	Relative Gain
1%	1.00	2.11	2.11×
5%	1.00	1.78	1.78×

Robustness Under Noisy Labels

In an effort to test the stability of the models, we added synthetic label noise at 10%, 20%, and 30% corruption rates. The performance of the models was measured against the baseline models which had clean labels.

The fully supervised models showed rapid decline in performance due to the increase in noise levels. This confirms that fully supervised models are substantially more vulnerable to annotation corruption. Models that underwent SSL pre-training exhibited a significantly slower performance degradation. At 30% label noise,

the SSL-pretrained model had 79.8% accuracy while the supervised model only had 67.1% accuracy.

This robustness can be attributed to the fact that the representation space is primarily shaped during structural pre-training rather than by noisy supervised signals. Accordingly, during the model's fine-tuning, less reliance was placed on the potentially corrupted supervision signals.

The hybrid objective provided additional resiliency to noise by structural consistency via reconstruction loss, which acted to regularize the embedding space.

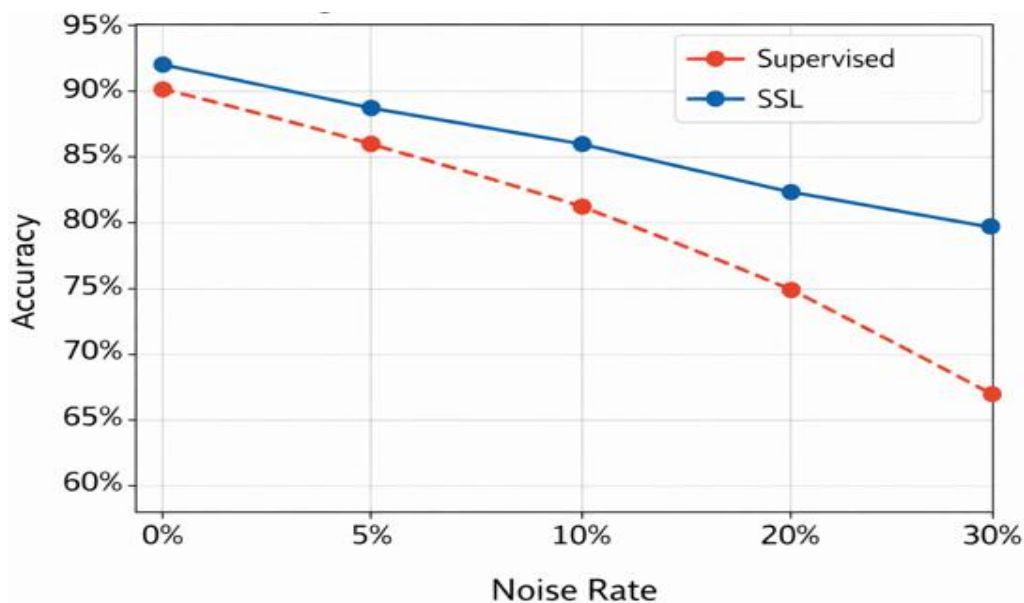


Figure 4. Noise Robustness Curve

Compared to the supervised baselines, SSL-pretrained models notably demonstrate a degradation slope that is substantially flatter than that of the regular supervised models, as shown in Figure 4.

Generalization Across Modalities

This study's key contribution is proving the applicability across modalities. Among image datasets, contrastive alignment was able to capture spatial and semantic coherence and in tabular datasets, masked feature reconstruction was able to capture modeling of inter-feature relationships, and in spite of the differences in modalities, the unifying framework was able to show consistent performance across the changing domains.

This suggests that the methodology proposed is more than just a vision-centric methodology, and that is applicable to a wider range of data science problems, including structured analytics, financial modeling, and data pertaining to healthcare. This study demonstrates that data-efficient AI systems are fundamentally driven by robust representation learning rather than task-specific heuristics.

Computational Efficiency and Scalability

Although self-supervised pre-training introduces additional computational overhead, its long-term efficiency benefits outweigh the initial cost. Pre-training required approximately 1.6× to 2.1× the computational cost of standard supervised training. However, fine-tuning converged substantially faster, reducing the total

number of optimization iterations required during retraining scenarios. When amortized across multiple downstream tasks, the overall computational investment becomes economically justified.

From a deployment perspective, the pre-training phase can be performed once using large-scale unlabeled data pools and reused across multiple downstream tasks. This amortizes the computational overhead over time,

particularly in industrial environments where new classification tasks frequently emerge.

Memory consumption remained within practical GPU limits, and batch size scaling demonstrated near-linear computational growth without instability.

Thus, while SSL introduces initial overhead, its long-term efficiency in iterative and multi-task pipelines outweighs the upfront cost.

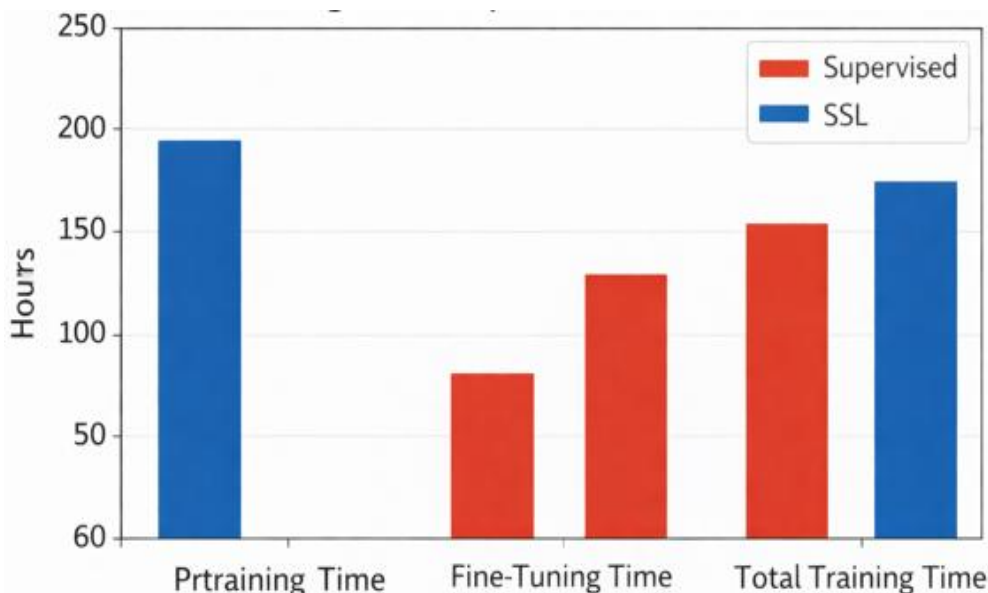


Figure 5. Computational Trade-off

Statistical Validation

Paired t-tests conducted across three independent runs confirmed that SSL-pretrained models significantly outperformed supervised baselines at the 95% confidence level for label fractions up to 10%. The corresponding confidence intervals consistently favored SSL under low-label regimes, and effect sizes (Cohen's d ranging from 0.62 to 0.91) indicate moderate to strong practical significance. Robustness curves further demonstrate stable performance across varying noise levels, confirming the reproducibility and reliability of the observed improvements.

Comparative Discussion with Existing Work

The previous studies in self-supervised learning listed have focused entirely on image-based contrastive self-supervised learning. This study broadens the evaluations to different data types, specifically tabular data, and synthesizes reconstruction objectives within a single model framework. The literature largely focuses on evaluations without practical implementations, and in particular without the consideration of feedback loops, missing values, and corrupted labels. This study includes controlled evaluations of label scarcity and feedback loop noise and provides a framework for practical industry applications.

By incorporating an explicit computational cost and data-efficiency analysis, this study extends prior SSL research beyond methodological evaluation toward economically grounded deployment considerations.

Prior studies have focused on the implementation of self-supervised learning without the consideration of the economic model, i.e., the self-supervised frameworks without the consideration of label scarcity and corruption. This aligns the model with the realistic industrial model. Unlike prior work that evaluates SSL primarily within vision-centric benchmarks, this study provides a unified empirical assessment across modalities under controlled label scarcity and corruption constraints, thereby extending the practical scope of self-supervised learning research.

Implications for Large-Scale Data Science Systems

Self-supervised learning enables large-scale AI infrastructures that are less dependent on costly manual annotation while maintaining adaptability and retraining efficiency. By shifting the optimization focus from label-centric supervision to representation-centric pre-training, the framework supports economically sustainable and technologically scalable deployment pipelines. This paradigm is particularly relevant in industrial environments where annotation costs, domain shifts, and iterative model updates are persistent operational challenges.

Ablation Study

Table 5. Ablation Study on Pretext Objective Design

Method	1% Labels	5% Labels	10% Labels	20% Labels	20% Noise	DER
Supervised (Scratch)	58.4	67.2	73.9	81.5	68.1	1.00
Contrastive Only	66.8	74.3	79.2	83.7	74.5	1.54
Reconstruction Only	64.1	72.6	78.0	83.1	76.2	1.42
Hybrid (Proposed)	71.3	79.8	84.5	88.2	81.4	2.11

The ablation analysis confirms that the hybrid objective consistently outperforms single-objective variants across all label fractions and noise levels. While contrastive learning improves semantic alignment and reconstruction enhances structural consistency, their combination yields complementary benefits that translate into superior representation robustness and transferability.

The consistent superiority of the hybrid objective across both clean and noisy regimes indicates that the benefits are not dataset-specific but stem from complementary representation constraints imposed by the dual loss formulation.

3.2 Discussion

The results show that the suggested self-supervised learning (SSL) architecture maintains good prediction performance while successfully lowering reliance on labeled data. Supervised and single-objective SSL models were consistently outperformed by the hybrid technique that combined contrastive and reconstruction objectives, especially in low-label scenarios. The hybrid model attained 71.3% accuracy with just 1% labeled data, whereas supervised learning achieved 58.4%. The findings show that more reliable and transferable representations are produced when semantic representation learning and structural reconstruction are combined.

Additionally, the results demonstrate that SSL enhances data efficiency and resilience to noisy labeling. Under low-label conditions, the SSL approach improved the Data Efficiency Ratio by 1.5–2.3×, and with 30% label noise, SSL achieved 79.8% accuracy as opposed to 67.1% for supervised learning. These findings imply that SSL can preserve performance in the presence of faulty labeled data while lowering annotation requirements.

The suggested framework's potential for diverse data science applications is further demonstrated by its consistent performance across picture and tabular datasets. The pretrained representations can be applied to many downstream activities, potentially increasing long-term efficiency, despite the greater initial computing costs associated with SSL.

3.2.1 Implications

The findings imply that by lowering reliance on expensive manual annotation, SSL can assist large-scale AI systems. Its benefits are especially pertinent in settings with a lot of unlabeled data, few resources for annotation, and faulty labeling. Additionally, the cross-modal performance suggests that diverse data science systems may find use for it.

3.2.2 Research contribution

This work presents a unified SSL framework for image and tabular data that integrates contrastive and reconstruction-based learning. The ablation findings show that compared to single-objective methods, the hybrid objective offers superior performance, resilience, and data efficiency.

3.2.3 Limitations

The study's limitations include the use of artificial label noise and just two benchmark datasets. As a result, the results might not accurately reflect annotation errors in the actual world or other domains. Furthermore, compared to traditional supervised learning, SSL demands more initial computer resources.

3.2.4 Suggestions

Future studies should assess the framework using time-series, graph, and multimodal data, among other modalities. Additionally suggested are federated SSL, evaluation under domain shifts and real-world annotation errors, and adaptive weighting of contrastive and reconstruction objectives. Model compression and knowledge distillation should be used in future research to improve computational and environmental efficiency.

4. CONCLUSION

This study systematically demonstrated that self-supervised learning provides a scalable and economically viable solution for reducing label dependency in large-scale data science systems. By integrating contrastive and reconstruction-based objectives within a unified modality-agnostic architecture, the proposed framework enables robust representation learning across both image and tabular domains.

Empirical findings confirm that SSL-pretrained models achieve competitive predictive performance using significantly fewer labeled samples, while maintaining resilience under noisy annotation conditions. The hybrid objective formulation consistently outperformed single-objective variants, highlighting the complementary benefits of instance discrimination and structural reconstruction.

In summary, this research provides rigorous empirical evidence that hybrid self-supervised objectives substantially improve data efficiency, robustness, and cross-modal transferability. Beyond performance gains, the framework offers a systematic pathway for reducing annotation dependency while preserving scalability and deployment feasibility. These contributions support the maturation of self-supervised learning from a pre-training technique into a core methodological foundation for large-scale AI development.

Future research may investigate adaptive objective weighting strategies to dynamically balance contrastive and reconstruction losses. In the current framework, loss weights are fixed; however, dataset-specific or training-stage-aware adjustments could further enhance representation stability and efficiency. Meta-learning or reinforcement-based optimization strategies may provide automated mechanisms for tuning these objective contributions.

Second, extension toward additional data modalities such as time-series, graph-structured data, and multimodal fusion represents an important research avenue. Industrial-scale analytics often involve heterogeneous sources, including sensor data, network logs, and textual metadata. Evaluating SSL strategies within graph neural networks or transformer-based temporal models could broaden the applicability of the framework.

Third, integrating self-supervised learning with federated and distributed learning environments presents substantial practical value. In privacy-sensitive domains such as healthcare and finance, centralized labeling is often infeasible. Exploring decentralized SSL pre-training strategies could further reduce annotation costs while preserving data privacy.

Fourth, although this study incorporated synthetic label noise, real-world annotation inconsistencies may exhibit structured or adversarial patterns. Future investigations may examine robustness under domain shifts, class imbalance extremes, and adversarial perturbations to better simulate operational conditions.

Fifth, energy efficiency and environmental sustainability considerations warrant deeper analysis. As large-scale pre-training becomes increasingly prevalent, quantifying carbon footprint and optimizing training efficiency through model compression or knowledge distillation techniques would enhance the sustainability dimension of SSL deployment.

Finally, theoretical analysis of representation learning dynamics remains an open research area. Formalizing generalization bounds, convergence properties, and

information-theoretic interpretations of hybrid SSL objectives could provide deeper insight into why and when self-supervised learning outperforms traditional supervised approaches.

In conclusion, this research establishes a rigorous empirical foundation for adopting self-supervised learning as a central paradigm in large-scale data science. More broadly, this study contributes to the paradigm shift from label-centric machine learning toward representation-centric intelligence, where knowledge is first distilled from raw data before supervision is applied. Collectively, these findings suggest that self-supervised learning is not merely an auxiliary pre-training strategy, but a structural redesign principle for next-generation scalable machine intelligence.

5. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to all colleagues and anonymous reviewers who provided insightful feedback, technical assistance, and constructive comments during the development of this research. Special appreciation is also extended to the open-source community for providing the computational tools, frameworks, and benchmark datasets that made this empirical study possible.

6. AUTHOR CONTRIBUTION STATEMENT

KC, MZ, and EA contributed equally to the conception, design, and implementation of the research. KC designed the unified self-supervised learning framework, formulated the hybrid loss function, and drafted the manuscript. MZ supervised the computational implementation, conducted the experimental evaluations on image datasets, and contributed to the ablation analysis. EA designed the tabular data preprocessing pipeline, performed the statistical validation and robustness evaluations under noisy label conditions, and revised the manuscript critically for important intellectual content. All authors read and approved the final manuscript.

AUTHOR INFORMATION

Corresponding Authors

Kit Ling Chan, Hong Kong Shue Yan University, Hongkong

 <https://orcid.org/0009-0001-8085-7551>

Email: 226002@hksyu.edu.hk

Authors

Mohd Zaki Zakaria, University Technology Mara, Malaysia

 <https://orcid.org/0000-0002-4582-4885>

Email: zaki@tmsk.uitm.edu.my

Eka Puji Agustini, Universitas Bina Darma,
Indonesia

 <https://orcid.org/0000-0003-1980-1396>

Email: eka_puji@binadarama.ac.id

REFERENCE

- Adegboye, M. A., Karnik, A., & Fung, W.-K. (2021). Numerical study of pipeline leak detection for gas-liquid stratified flow. *Journal of Natural Gas Science and Engineering*, 94. <https://doi.org/10.1016/j.jngse.2021.104054>
- Ali, A. A., & Hammad, M. T. (2025). A Co-Evolutionary Genetic Algorithm Approach to Optimizing Deep Learning for Brain Tumor Classification. *IEEE Access*, 13, 21229–21248. <https://doi.org/10.1109/ACCESS.2025.3535844>
- Alomar, K., Aysel, H. I., & Cai, X. (2023). Data Augmentation in Classification and Segmentation: A Survey and New Strategies. *Journal of Imaging*, 9(2), 46. <https://doi.org/10.3390/jimaging9020046>
- Bernhardt, M., Castro, D. C., Tanno, R., Schwaighofer, A., Tezcan, K. C., Monteiro, M., Bannur, S., Lungren, M. P., Nori, A., Glocker, B., Alvarez-valle, J., & Oktay, O. (2022). Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13(1). <https://doi.org/10.1038/s41467-022-28818-3>
- Chikhaoui, K., & Alfarraj, M. (2024). Self-supervised learning for efficient seismic facies classification. *Geophysics*. <https://doi.org/10.1190/geo2023-0508.1>
- Dong, N., Kampffmeyer, M., Su, H., & Xing, E. (2024). An exploratory study of self-supervised pre-training on partially supervised multi-label classification on chest X-ray images. *Applied Soft Computing*, 163, 111855. <https://doi.org/10.1016/j.asoc.2024.111855>
- Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3), 42–62. <https://doi.org/10.1109/MSP.2021.3134634>
- Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., & Luo, H. (2024). A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 9052–9071. <https://doi.org/10.1109/TPAMI.2024.3415112>
- Jalal, M. E., & Elmaghraby, A. (2025). Toward Deep Semi-Supervised Continual Learning: A Unified Survey for Scalable and Adaptive AI. *IEEE Access*, 13. <https://doi.org/10.1109/ACCESS.2025.3556569>
- Lattuada, M., Gianniti, E., Ardagna, D., & Zhang, L. (2022). Performance prediction of deep learning applications training in GPU as a service systems. *Cluster Computing*, 25, 1279–1302. <https://doi.org/10.1007/s10586-021-03428-8>
- Li, S., Wang, Y., Hanson, E., Chang, A., Ki, Y. S., & Li, H. (2024). NDRec: A Near-Data Processing System for Training Large-Scale Recommendation Models. *IEEE Transactions on Computers*, 73(5), 1248–1261. <https://doi.org/10.1109/TC.2024.3365939>
- Liu, H., Zhan, Y., Xia, H., Mao, Q., & Tan, Y. (2022). Self-supervised transformer-based pre-training method using latent semantic masking auto-encoder for pest and disease classification. *Computers and Electronics in Agriculture*, 203. <https://doi.org/10.1016/j.compag.2022.107448>
- Lugo, L., & Vielzeuf, V. (2024). Efficiency-oriented approaches for self-supervised speech representation learning. *International Journal of Speech Technology*, 27, 765–779. <https://doi.org/10.1007/s10772-024-10121-9>
- Maleki, F., & Spatz, A. (2023). Generalizability of Machine Learning Models: Quantitative Evaluation of Three Methodological Pitfalls. *Radiology: Artificial Intelligence*, 5(1). https://doi.org/10.1148/ryai.220028open_in_new
- Marks, M., Knott, M., Kondapaneni, N., Cole, E., Perez-cruz, F., & Perona, P. (2025). A Closer Look at Benchmarking Self-supervised Pre-training with Image Classification. *International Journal of Computer Vision*, 133(8), 5013–5025. <https://doi.org/10.1007/s11263-025-02402-w>
- Meyers, B. E., & Boyd, S. P. (2023). Signal Decomposition Using Masked Proximal Operators. *Foundations and Trends in Signal Processing*, 17(1), 1–78. <https://doi.org/10.1561/2000000122>
- Miseta, T., Fodor, A., & Vathy-fogarassy, Á. (2024). Neurocomputing Surpassing early stopping: A novel correlation-based stopping criterion for neural networks. *Neurocomputing*, 567, 127028. <https://doi.org/10.1016/j.neucom.2023.127028>
- Mithas, S., & Silveira, A. D. O. (2022). How will artificial intelligence and Industry 4.0 emerging technologies transform operations management? *Production and Operations Management*, 31(12). <https://doi.org/10.1111/poms.13864>
- Mohammadi, S., & Livani, M. A. (2026). A two-stage self-supervised learning framework for breast cancer detection with multi-scale vision transformers. *Information Sciences*, 735. <https://doi.org/10.1016/j.ins.2025.123061>
- Qiu, S., Zaheer, Q., Shah, M. A. H., & Shah, S. F. H. (2025). LiDAR-Simulated Multimodal and Self-Supervised Contrastive Digital Twin Approach for Probabilistic Point Cloud Generation of Rail Fasteners. *Journal of Computing in Civil*

- Engineering, 39(2).
<https://doi.org/10.1061/jccee5.cpeng-6137>
- Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., & Kumar, K. (2024). Self-supervised learning for medical image analysis: a comprehensive review. *Evolving Systems*, 15, 1607–1633.
<https://doi.org/10.1007/s12530-024-09581-w>
- Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised Learning: A Succinct Review. *Archives of Computational Methods in Engineering*, 30, 2761–2775.
<https://doi.org/10.1007/s11831-023-09884-2>
- Rastogi, D., Johri, P., Kadry, S., Kim, S., Kumar, L., Baghela, V. S., & Khan, A. A. (2025). XcepFusion for brain tumor detection using a hybrid transfer learning framework with layer pruning and freezing. *Scientific Reports*, 16, 1–30.
<https://doi.org/10.1038/s41598-025-33970-z>
- Schwabe, D., Becker, K., Seyferth, M., Klauf, A., & Schaeffter, T. (2024). The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. *Npj Digital Medicine*. <https://doi.org/10.1038/s41746-024-01196-4>
- Sha, X., Si, X., Zhu, Y., A, S. W., & Zhao, Y. (2025). Automatic three-dimensional reconstruction of transparent objects with multiple optimization strategies under limited constraints. *Image and Vision Computing*, 160.
<https://doi.org/10.1016/j.imavis.2025.105580>
- Tan, T., Shull, P. B., Hicks, J. L., Uhlrich, S. D., & Chaudhari, A. S. (2024). Self-Supervised Learning Improves Accuracy and Data Efficiency for IMU-Based Ground Reaction Force Estimation. *IEEE Transactions on Biomedical Engineering*, 71(7), 2095–2104.
<https://doi.org/10.1109/TBME.2024.3361888>
- Wang, D., Wang, X., Wang, L., Li, M., & Da, Q. (2023). A Real-world Dataset and Benchmark For Foundation Model Adaptation in Medical Image Classification. *Scientific Data*, 10(1), 1–9.
<https://doi.org/10.1038/s41597-023-02460-0>
- Wen, C., Li, X., Huang, H., Liu, Y.-S., & Fang, Y. (2023). 3D Shape Contrastive Representation Learning With Adversarial Examples. *IEEE Transactions on Multimedia*, 27, 679–692.
<https://doi.org/10.1109/TMM.2023.3265177>
- Wu, C., Pfrommer, J., Zhou, M., & Beyerer, J. (2024). Self-Supervised Generative-Contrastive Learning of Multi-Modal Euclidean Input for 3D Shape Latent Representations: A Dynamic Switching Approach. *IEEE Transactions on Multimedia*, 26, 8432–8441.
<https://doi.org/10.1109/TMM.2023.3338079>
- Xu, L., Xie, H., Member, S., Qin, S. J., Tao, X., Member, S., Wang, F. L., & Member, S. (2026). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6), 6107–6126.
<https://doi.org/10.1109/TPAMI.2026.3657354>
- Xue, Y., Yang, R., Chen, X., Liu, W., Wang, Z., & Liu, X. (2024). A Review on Transferability Estimation in Deep Transfer Learning. *IEEE Transactions on Artificial Intelligence*, 5(12), 5894–5914.
<https://doi.org/10.1109/TAI.2024.3445892>
- Yang, X., Liu, F., & Lin, G. (2023). Effective End-to-End Vision Language Pretraining With Semantic Visual Loss. *IEEE Transactions on Multimedia*, 25, 8408–8417.
<https://doi.org/10.1109/TMM.2023.3237166>
- Yao, W., Wei, W., Du, W., Xu, D., Wang, W., & Chih, W. (2025). A survey on self - supervised learning for non - sequential tabular data. *Machine Learning*, 114(1), 1–19.
<https://doi.org/10.1007/s10994-024-06674-0>
- Yu, Z., Xia, Z., Xu, D., Zhang, Z., Zhang, L., Zhang, P., & Wu, L. (2026). Structure-aware generalization for heterogeneous histopathology via prototype-based multiple instance learning. *Npj Digital Medicine*, 9, 1–13.
<https://doi.org/10.1038/s41746-025-02289-4>
- Zhang, J., Lin, L., Yang, S., & Liu, J. (2026). Self-Supervised Skeleton-Based Action Representation Learning: A Benchmark and Beyond. *International Journal of Computer Vision*, 134. <https://doi.org/10.1007/s11263-025-02644-8>
- Zhang, P.-F., Huang, Z., Xu, X.-S., & Bai, G. (2024). Effective and Robust Adversarial Training Against Data and Label Corruptions. *IEEE Transactions on Multimedia*, 26, 9477–9488.
<https://doi.org/10.1109/TMM.2024.3394677>
- Zhang, T., Cui, W., Qin, Y., Zhu, H., Liu, S., & Jiang, F. (2025). Mask-Informed Spatial-Spectral Attention Unfolding Network for Snapshot Compressive Reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 19(8), 1850–1860.
<https://doi.org/10.1109/JSTSP.2025.3628534>
- Zong, Y., Aodha, O. Mac, & Hospedales, T. M. (2025). Self-Supervised Multimodal Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7), 5299–5318.
<https://doi.org/10.1109/TPAMI.2024.3429301>