iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: VII Month of publication: July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73282

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Self-Supervised Learning for Small Data Environments

Thota Sharon[1], Terli Lavanya[2], Ganeshwar Sai[3], Vasudev Sharma[4], Vempadapu Sai Charan[5], M. Rupasri[6]
[1, 2, 3, 4, 5] *Student, MCA, Dr. Lankapalli Bullayya College, Andhra Pradesh, India*
[6] *Assistant Professor, Dept. of Computer Science, Dr. Lankapalli Bullayya College, Andhra Pradesh, India*

Abstract: Self-Supervised Learning (SSL) presents itself as a dominant learning paradigm that allows models to extract valuable information from unlabeled data collections. The general success of SSL in handling large-scale datasets does not address its potential application in limited data settings. This study analyzes SSL approaches designed for small data systems while focusing on their implementation methods in addition to their challenges and developments. This paper investigates three methods to improve learning efficiency through minimal supervision by focusing on data augmentation and contrastive learning and pre training strategies. Experimental studies show that SSL produces superior performance than standard supervised learning approaches when used in limited data circumstances.

I. INTRODUCTION

Machine learning systems usually need millions of labeled training data to become highly efficient. However, ground truths are often limited in real scenarios because of large annotation cost, privacy issues, or limited data availability. SSL provides a good answer through the use of unidentified data to learn valuable representation before fine tunings on little labelled datasets. This paper describes SSL techniques tailored for small data domains and examines their effectiveness. Self-Supervised Learning represents a thriving technique for resolving vision and imaging data problems because it handles restrictions from limited labeled data across various fields including medicine. Self-Supervised Learning uses inherent data structure to help models extract valuable representations that need only minimal labeled data to achieve effective downstream performance. The technique offers significant advantages to address situations with expensive or impractical or time-consuming annotation needs.

II. SELF-SUPERVISED LEARNING: AN OVERVIEW

Self-Supervised- Learning (SSL) enables models to learn useful data representations without relying on manual annotations. In principle, the learning process is based as a result of automatically generated labels, which arise out of the data itself. This method works well in different areas, such as computer vision, natural language processing (NLP), speech based applications etc. In this section we examine three well-known SSL methods

A. Contrastive Learning

Contrastive Learning is centered on learning disruptive characteristics by seeing that similar information examples are mapped along a similar line in the capacity space while disparate examples are scattered separately. This is most often accomplished with positive and negative sample pairing.

Key Techniques

- SimCLR (Simple Contrastive Learning Representation) : Uses augmented views of the same image as positive pairs and different images as negative pairs. A contrastive loss (e.g., InfoNCE) ensures meaningful feature learning.
- MoCo (Momentum Contrast) : Maintains a queue of past embeddings and applies a moving-average encoder to stabilize contrastive training.
- BYOL (Bootstrap Your Own Latent) : Eliminates the need for negative samples by employing two networks—one learning representations and the other acting as a momentum-updated target.

Applications

- Image recognition (e.g., using pre-training models as a starting point for classifying images easily with few labelled copies).
- Video understanding (learning temporal relationships in videos).
- Speaker recognition (distinguishing voices whereas few labels are there).

B. Predictive Coding

Predictive coding training methods learn models to predict the missing or missing parts of the input data. By learning to complete incomplete information, the model participates in abstract representations.

Key Techniques

- Masked Language Modeling (MLM): Employed in NLP (e.g. in BERT) where a few words of a sentence are randomly masked and the word has to be predicted.
- Masked Image Modeling (MIM): Visualization models such as MAE (Masked Autoencoders) mask areas of image as well as train a transformer to recapture absent pixels.
- Wave2Vec (for Speech Transcription): Factors speech representations from predicting masked audio chunks.

Applications

- NLP tasks such as text classification and question answering, machine translation.
- Image inpainting (reclaiming of parts of images).
- Speech recognition with a small amount of transcribed audio.

C. Clustering-Based Ssl

Clustering-based SSL is a strategy that assigns pseudo-labeling to unlabeled data by cluster similar sample. A model is then trained using this pseudo-labels on these pseudo-labels so that it can learn structured representations.

Key Techniques

- DeepCluster : Applies k-means clustering to assign labels for image features and trains a network to classification.
- SwAV (Swapping Assignments between Views) : Semi-supervised contrastive learning with pseudo-labels dynamically generated in the process of reconstructions, without explicit contrastive losses.
- SEER (Self-Supervised Embeddings for Efficient Recognition) : Facebook AI’s SSL technique employing clustering-based representation learning for large-scale vision work.

Applications

- Unsupervised learning for image classification.
- Anomaly detection (e.g., fraud detection where labeled anomalies.
- Recommendation systems (that categorize user similarities).

III. THE CHALLENGE OF SMALL DATASETS

Conventional models of deep learning require big data with annotations for getting the best performances. Nevertheless, getting such datasets may not be easy because of the following factors:

1) Privacy Issues: Health information is considered one of the most sensitive forms of information that is shared in the health care sector and this must not neglect the privacy laws set in the society.
2) Absence of Data: At-times, there may be very less amount of data available in a particular domain.
3) High Cost: Cost of the annotation especially by qualified personnel like Radiologists in the case of health image annotation remains high.

Such limitations contribute to the rather limited use of supervised deep learning methods especially for clinics, small hospitals, or specific research niches.

IV. OVERCOMING LIMITATIONS WITH SELF-SUPERVISED LEARNING

This occurs when the amount of data is relatively smaller compared to the number of input variables, which is not the case with Self-Supervised Learning, and it offers an effective solution to this problem by training models through the labeled data while using the unlabeled data to boost the results.

The strategy can be formulated as follows: to introduce “pretext tasks” that make the model learn useful representations of the data. Such representations can then be used in other tasks, which include classification or segmentation or any other task that requires labeled data with little retraining.

V. COMMON SELF-SUPERVISED LEARNING TECHNIQUES

There are several successful Self-Supervised Learning techniques that may be applied in small data :

1) Contrastive Learning: Contrastive learning is designed to learn representations with similar samples' locations in the identical embedding space while dissimilar samples are located on the opposite ends of the space. Contrastive learning has been applied in various fields such as medical image analysis and photoplethysmogram (PPG) signal artifact removal.
2) Masked Image Modeling (MIM): In MIM, like in VIM, certain areas of the image are masked and the model needs to learn to regenerate the masked regions. Specifically it makes the model learn contextual relations and dependencies within the inputs so important for the task. As stated above, masked autoencoders are popular examples of MIMs and also demonstrate transferability across various tasks.
3) Self-Distillation: In this method, a student network is trained to replicate the output of a teacher network and the latter can be another instance of a student network or a few different models. DINO is self-distillation using no labels, and is classified under this category of methods.

VI. ADVANTAGES OF SELF-SUPERVISED LEARNING

There are several benefits associated with using Self-Supervised Learning especially when working with a small data set.

1) Improved Generalization: As a result of the amount of patterns the Self-Supervised Learning learns on unlabeled data, its ability to generalize on confined labeled data is enhanced; overfitting is also minimized.
2) Less Supervised Training Data Requirement: Self-Supervised Learning diminishes the demand for large amounts of annotated data meaning it is plausible to develop excellent models even in places where there is little dataset.
3) Enhanced Robustness: Self-Supervised Learning helps to reduce the effect of noise and variations in data hence its enhanced robustness in deployment of models in real world situations.

VII. APPLICATIONS OF SELF-SUPERVISED LEARNING IN SMALL DATA ENVIRONMENTS

Self-Supervised Learning has been applied in many domains for which data is scarce in many ways:

1) Medical Imaging: This is because medical imaging has one of the challenges of limited data availability caused by data privacy, expensive annotation, or due to the RARE nature of the illness. But Self-Supervised Learning has aimed to become an effective solution in this field.
2) Image Segmentation: Jin Kim, Matthew Brown and Dan Ruan discussed their work on self-supervised learning using the DINO approach for segmentation of chest X-ray (CXR) images without using annotations. Their contribution did show that the Self-Supervised Learning method aids in improving CXR segmentation with the use of unsupervised data as opposed to extensive annotated data.
3) Polyp Diagnostics: Several authors, including Heba El-Shimy, Hind Zantout, M. Lones, and N. E. Gayar, studied pre-training the capsule network using self-supervision methods in the diagnosis of colon cancer polyps. The study showed that the contrastive learning and in-painting are appropriate techniques for the auxiliary task of self-supervised learning in the medical field having better performance by 5.26% over weight initialization methods.
4) Classification of Shoulder Implants: In the present study, Laith Alzubaidi et al. addressing the problem of small dataset while classifying the shoulder implants in X-ray image proposed a new transfer learning, namely self-supervised pertaining (SSP). The self-supervision learning technique involves producing deep learning models that use a large collection of greyscale image datasets that are not labeled as shoulder implants, and updating the features, and then train on a small set of labeled X-ray images used in implants.
5) Gastrointestinal Lesions Classification : In order to reduce the impact of small data size, Z. M. Lonseko, N. Rao, Cheng-Si Luo, P, E. Adjei and Tao Gan introduced the GI lesions classification based on supervised contrastive representative learning. It outperforms the other related state-of-the-art classification methods to get better the lesion classification accuracy of 96.4%.
6) Liver Fibrosis and NAS Scoring : Self-supervised approach was introduced by Ananya Jana, Hui Qu, Carlos D. Minacapelli, Carolyn Catalano, Vinod K. Rustgi, and Dimitris Metaxas to predict NAS scores on noninvasive CT images that were not addressed earlier due to the requirement of a huge annotated database and the issue of domain shift.
7) Crohn's Disease Detection: For the detection of Crohn's disease, as there is typically little labeled data, Jing Xing and Harold Mouchere proposed a contrastive self-supervised method to improve the problem of CAD for WCE images.

VIII. STRUCTURAL HEALTH MONITORING

Mingyuan Zhou, Xudong Jian, Ye Xia, and Zhilu Lai investigated the employment of self-supervised learning-based techniques for anomaly detection in data of bridge structural health monitoring (SHM). Still, considering the findings of this study, the mainstream Self-Supervised Learning methods were compared and tested on the SHM data of two typical in-service bridges and it is established that the employment of the Self-Supervised Learning methods enhances the performance of data anomaly detection with higher F1 scores compared to the conventional supervised training with a minute amount of labeled data.

IX. FOOD FRAUD DETECTION

In order to solve the problem of imbalanced learning in food adulteration detection based on HSI, Kunkun Pang, Yisen Liu, Songbin Zhou, Yixiao Liao, Zexuan Yin, Lulu Zhao, and Hong Chen put forward a new nondestructive detection method called Dice Loss Improved Self-Supervised Learning-Based Prototypical Network (Proto-DS). In the experiments, they were able to show that Proto-DS consistently outperforms the other methods with the highest average balanced accuracy of 88.18% in various training sample sizes.

X. REMOTE SENSING

There are several difficulties in the remote sensing scene classification: remoteness of the scene means that characterizing them precisely is difficult; the classes often overlap; there is a lack of enough labeled scenes for training.

1) Scene Classification: Najd Alosaimi, H. Alhichri, Y. Bazi, Belgacem Ben Youssef, and N. Alajlan designed a deep Self-Supervised Learning method known as RS-FewShotSelf-Supervised Learning for RS scene classification only under the scenario where having fewer than twenty labelled scenes per class is possible. This proposed RS-FewShotSelf-Supervised Learning is made of an online network and a target network, both of which utilize EfficientNet-B3 CNN model as feature encoder backbone.

2) Land-Use and Land-Cover Fraction Estimation: The application of Self-Supervised Learning for LULC fraction estimation for the RGB satellite patches yielding to in-domain knowledge is done by Andres J. Sanchez-Fernandez, S. Moreno-lvarez, Juan A. Rico-Gallego, and S. Tabik. Their experiments proved that the accuracy of Self-Supervised Learning is as good or even better when trained on a comparatively smaller in-domain high-quality set rather than the supervised model trained on ImageNet-1k.

XI. OTHER APPLICATIONS

1) PPG Signal Artifact Detection: Self-supervised learning has been used to learn the features from the abundant unlabeled PPG signals and then transfer the learned knowledge to the labeled data set to solve the problem of detecting artifact from PPG signal: Thanh-Dung Le, Clara Macabiau, Kevin Albert, P. Jouvet, and R. Noumeir. This work showed that Self-Supervised Learning improved the Transformer models ability to learn representations as it provides robustness in the artifact classification tasks.

2) Molecular Property Prediction: Molecular property prediction was learned using Self-Supervised Learning by Yuankai Luo, Lei Shi, and Veronika Thost; they used persistent homology (PH), which is a mathematical tool for analyzing topological structures of data at different scales.

3) Human Activity Recognition: Masked reconstruction is proposed to be an effective self-supervised pre-training objective for human activity recognition, H. Haresamudram, Apoorva Beedu, Varun Agrawal, Patrick Grady, Irfan Essa, Judy Hoffman and T. Pltz presented an investigation of masked reconstruction relative to other self-supervision approaches that are nowadays popular.

4) Few-Shot Learning: Several approaches to few-shot learning were examined by Puneet Mangla, M. Singh, Abhishek Sinha, Nupur Kumari, V. Balasubramanian, and Balaji Krishnamurthy such as learning relevant feature manifold using self-supervision and various regularizations. They pointed out that the feature manifold was regularized so that few-shot learning performance could be boosted with the use of Manifold Mixup, a self-supervised technique that enriches the feature manifold.

5) Object Discovery and Detection: Etienne Pot, Alexander Toshev and J. Kosecka used self-supervisory signals that come from a robot navigating in the forward darkness environment in order to learn representations of the objects in the environment. They also showed that this representation is useful for finding the objects and for learning detectors of such labels .

XII. CHALLENGES AND CONSIDERATIONS

However there are few issues and concerns associated with Self-Supervised Learning that should be recognised:

1) Pretext Task Design: This paper also exhibits that the selection of pretext task is very important in learning the Good representations. Thus, it is important to choose the pretext task that goes well with the downstream task and also retain the characteristics that would be helpful.
2) Software and Computational Resources: As mentioned earlier, the Self-Supervised Learning models can be costly, specifically, when large datasets are processed and complicated models are adopted.
3) Transferability: It has been found that the transferability of the representations learned through Self-Supervised Learning in different domains can be limited, and this depends on the similarity of the specific domain of the pre-training and that of the fine-tuning dataset.
4) Evaluation Metric: To compare the efficacy of the Self-Supervised Learning methods being proposed, there is the need to choose criteria that best fit the task and the downstream application.
5) Data Imbalance: In the current implementations of the hyperspectral imaging (HSI) for food fraud including our preliminary work, there was an assumption of balanced classes while in a real application, class distributions are far from balanced.

XIII. STRATEGIES FOR OPTIMIZING SELF-SUPERVISED LEARNING IN SMALL DATA ENVIRONMENTS

Therefore, there are various approaches that can be used to optimize Self-Supervised Learning in small data environments:

1) Selecting Proper Self-Supervised Learning Techniques: There are various techniques for Self-Supervised Learning and some may be more appropriate for certain operations and with specific type of data. For instance, contrastive learning has been recently proposed in two different contexts: medical imaging and PPG signal artifact detection whereas masked image modeling might be more beneficial in the context, which requires extensive contextual information.
2) Data Augmentation : Depending on the type of Self-Supervised Learning problem, the field can use proper data augmentation methods that would help in diversifying the training data and enhance the Self-Supervised Learning model's generalization capability.
3) Transfer Learning from Related Domains : The use of a model from a related domain can be very useful when the data availability is limited since it will reduce the number of epochs required to train the model.
4) Regularization : Analyzing the performance of Self-Supervised Learning models, it is possible to use such regularization methods as weight decay and dropout to get rid of overfitting.
5) Joint Optimization : The combination of optimizing for the primary function as well as a Self-Supervised Auxiliary Task (SSAT) is helpful when the amount of data on hand is limited as stated in.
6) New Contrastive Loss Functions: It has been suggested that optimizing of contrastive loss functions can lead to less oscillation when training whereas increasing the conception of contrastive loss function improves artifact classification.

XIV. FUTURE DIRECTIONS

The area of Self-Supervised Learning is still vibrant and growing and several directions for further enhancing its effectiveness have been identified for small data regions:

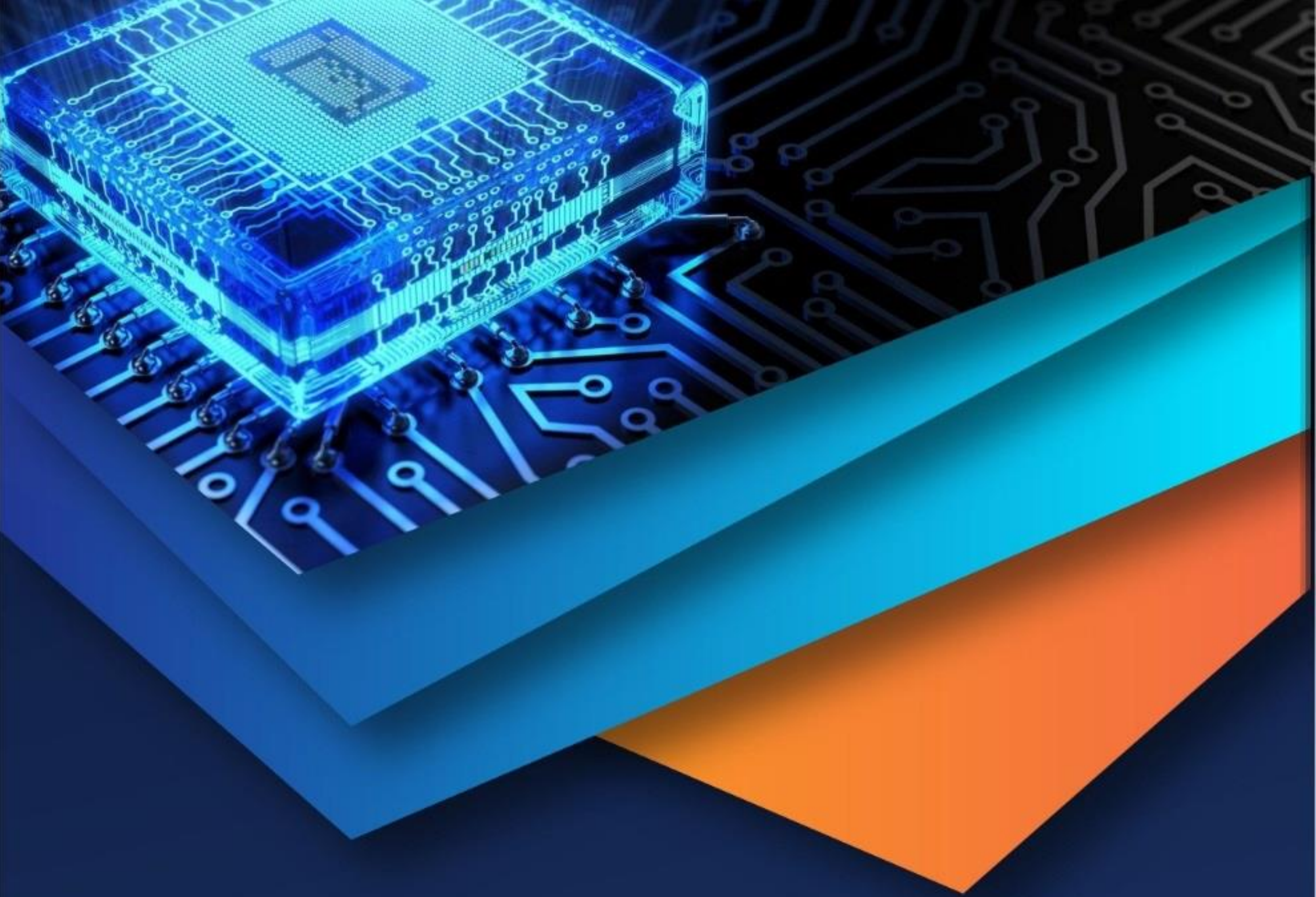
1) New ideas for Self-Supervised Learning approach: It is also considered better research to come up with Self-Supervised Learning algorithms that demand less amount of resources such as time and space and are easy to train with the available limited data.
2) Exploration of Novel Pretext Tasks: There is current research on extending pretext tasks that obtain more of the characteristics of the data to enhance transferability of learned representations.
3) Integration of Domain Knowledge: Domain Knowledge Advantages in Self-Supervised Learning can help to enhance the learning process as well as the resultant representations when incorporated in the frameworks.
4) Development of Automated Self-Supervised Learning Pipelines: Presently, work is in progress on designing automated Self-Supervised Learning pipelines that will enable it to independently decide on the right Self-Supervised Learning techniques, data augmentations, or hyperparameters to apply on a particular task or dataset.
5) Application of Self-Supervised Learning to New Domains: Self-Supervised Learning is being used in various new domains where data scarcity incorporates as one of the major problem statements such as robotics, material science and drug discovery

XV. CONCLUSION

Self-Supervision provides an effective way to address limitations associated with the availability of samples of small data in different areas. Self-Supervised Learning allows models to learn representations from the raw structures of the materials and render maximal utilizations of these representations for relearning by using limited labeled data. It is for this reason that as Self-Supervised Learning techniques are enhanced and the efficiency increases, they will become more useful in boosting machine learning in scenarios that lack ample data. There is also a look being made to optimize Self-Supervised Learning more in terms of computational complexity. Research is also being conducted about the more specific issue of selecting a better pretext task that captures data characteristics and is more transferable. Another branch of research is the incorporation of domain knowledge to the Self-Supervised Learning frameworks in a way to control the learning process and enhance representation. More advanced approaches could be applied when creating Automated Self-Supervised Learning pipelines to make selection of the best techniques and hyperparameters for particular tasks easier. Last but not least, the Self-Supervised Learning is being applied in the new domains where the data are scarce, for instance, the robotics applications & drug discovery.

REFERENCES

[1] Fdo, E Maria Joseph Saron, et al.. "Self-supervised learning for small-scale medical imaging dataset" World Journal of Advanced Engineering Technology and Sciences, 2024,. https://doi.org/10.30574/wjaets.2024.13.2.0526
[2] El-Shimy, Heba, et al.. "Self-Supervised Learning for Pre-training Capsule Networks: Overcoming Medical Imaging Dataset Challenges" arXiv.org, 2025,. https://doi.org/10.48550/arXiv.2502.04748
[3] Le, Thanh-Dung, et al.. "A Novel Transformer-Based Self-Supervised Learning Method to Enhance Photoplethysmogram Signal Artifact Detection" IEEE Access, NaN,. https://doi.org/10.1109/ACCESS.2024.3488595
[4] Nguyen, H., et al.. "Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis" None, 2024,. https://doi.org/10.1109/IJCB62174.2024.10744497
[5] Kim, Jin, et al.. "Self-supervised learning without annotations to improve lung chest x-ray segmentation" None, 2024,. https://doi.org/10.1117/12.3008582
[6] Alzubaidi, Laith, et al.. "SSP: self-supervised pertaining technique for classification of shoulder implants in x-ray medical images: a broad experimental study" Artificial Intelligence Review, 2024,. https://doi.org/10.1007/s10462-024-10878-0
[7] Lonseko, Z. M., et al.. "Supervised contrastive learning for gastrointestinal lesions classification in endoscopic images" None, 2022,. https://doi.org/10.1117/12.2662633
[8] Jana, Ananya, et al.. "Liver Fibrosis And NAS Scoring From CT Images Using Self-Supervised Learning And Texture Encoding" None, 2021,. https://doi.org/10.1109/isbi48211.2021.9433920
[9] Pang, Kunkun, et al.. "Proto-DS: A Self-Supervised Learning-Based Nondestructive Testing Approach for Food Adulteration with Imbalanced Hyperspectral Data" Foods, 2024,. https://doi.org/10.3390/foods13223598
[10] Alosaimi, Najd, et al.. "Self-supervised learning for remote sensing scene classification under the few shot scenario" Scientific Reports, 2023,. https://doi.org/10.1038/s41598-022-27313-5
[11] Sanchez-Fernandez, Andres J., et al.. "Self-Supervised Learning on Small In-Domain Datasets Can Overcome Supervised Learning in Remote Sensing" IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, NaN,. https://doi.org/10.1109/JSTARS.2024.3421622
[12] Luo, Yuankai, et al.. "Improving Self-supervised Molecular Representation Learning using Persistent Homology" Neural Information Processing Systems, 2023,. https://doi.org/10.48550/arXiv.2311.1732
[13] Haresamudram, H., et al.. "Masked reconstruction based self-supervision for human activity recognition" International Workshop on the Semantic Web, 2020,. https://doi.org/10.1145/3410531.3414306
[14] Mangla, Puneet, et al.. "Charting the Right Manifold: Manifold Mixup for Few-shot Learning" IEEE Workshop/Winter Conference on Applications of Computer Vision, 2019,. https://doi.org/10.1109/WACV45572.2020.9093338
[15] Pot, Etienne, et al.. "Self-supervisory Signals for Object Discovery and Detection" arXiv.org, 2018,. https://doi.org/None
[16] Kinakh, Vitaliy, et al.. "ScatSimCLR: self-supervised contrastive learning with pretext task regularization for small-scale datasets" None, 2021,. https://doi.org/10.1109/ICCVW54120.2021.00129
[17] Liu, Andy T., et al.. "Efficient Training of Self-Supervised Speech Foundation Models on a Compute Budget" Spoken Language Technology Workshop, 2024,. https://doi.org/10.1109/SLT61566.2024.10832361
[18] Wang, Shanshan, https://doi.org/10.1109/ICASSPW62465.2024.10626141



10.22214/IJRASET

INDEX COPERNICUS
45.98

IMPACT FACTOR:
7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR:
7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY