Self-Supervised Learning (SSL) in Deep Learning Contexts

Subjects: Computer Science, Artificial Intelligence

Contributor: Mohammed Majid Abdulrazzaq, Nehad T. A. Ramaha, Alaa Ali Hameed, Mohammad Salman, Dong Keon Yon, Norma Latif Fitriyani, Muhammad Syafrudin, Seung Won Lee

Self-supervised learning (SSL) is a potential deep learning (DL) technique that uses massive volumes of unlabeled data to train neural networks. SSL techniques have evolved in response to the poor classification performance of conventional and even modern machine learning (ML) and DL models of enormous unlabeled data produced periodically in different disciplines. However, the literature does not fully address SSL's practicalities and workabilities necessary for industrial engineering and medicine. Accordingly, this thorough review is administered to identify these prominent possibilities for prediction, focusing on industrial and medical fields. This extensive survey, with its pivotal outcomes, could support industrial engineers and medical personnel in efficiently predicting machinery faults and patients' ailments without referring to traditional numerical models that require massive computational budgets, time, storage, and effort for data annotation.

Keywords: deep learning (DL) ; self-supervised learning (SSL) ; machine learning (ML) ; cognition ; classification ; data annotation

1. Introduction

Concepts of AI, convolutional neural networks (CNNs), DL, and ML considered in the last few decades have contributed to multiple valuable impacts and core values to different scientific disciplines and real-life areas because of their amended potency in executing high-efficiency classification tasks of variant complex mathematical problems and difficult-to-handle subjects. However, some of them are more rigorous than others. More specifically, DL, CNN, and artificial neural networks (ANNs) have a more robust capability than conventional ML and AI models in making visual, voice, or textual data classifications [1].

The crucial rationale for these feasible models includes their significant classification potential in circumstances involving therapeutic diagnosis, maintenance, and production line prognostics. As these two processes formulate a prominent activity in medicine and engineering, the adoption of ML and DL models could contribute to numerous advantages and productivity records [2][3][4][5][6].

2. Major Characteristics and Essential Workabilities of SSL

As illustrated above, supervised learning (SL) needs annotated data to train numerical ML models to enable an efficient classification process in various conventional ML models. On the contrary, unsupervised learning (USL) classification procedures do not require labeled data to accomplish a similar classification task. Rather, USL algorithms can rely solely on identifying meaningful patterns in existing unlabeled data without necessary training, testing, or preparation [7].

For the previously illustrated industrial and medical pragmatic practices, SSL can often be referred to as predictive learning (or pretext learning) (PxL). Labels can be generated automatically, transforming the unsupervised problem into a flexible, supervised one that can be solved viably.

Another favorable solution of SSL algorithms is their efficient categorization of data correlated with natural language processing (NLP). SSL can allow researchers to fill in blanks in databases when they are not fully complete or have a high-quality definition. As an illustration, with the application of ML and DL models, existing video data can be utilized to reconstruct previous and future videos. However, without relying on the annotation procedure, SSL takes advantage of patterns linked to the current video data to efficiently complete the categorization procedure of a massive video database [8][9]. Correspondingly, the critical working principles of the SSL approach can be illustrated in the workflow shown in **Figure 1**.

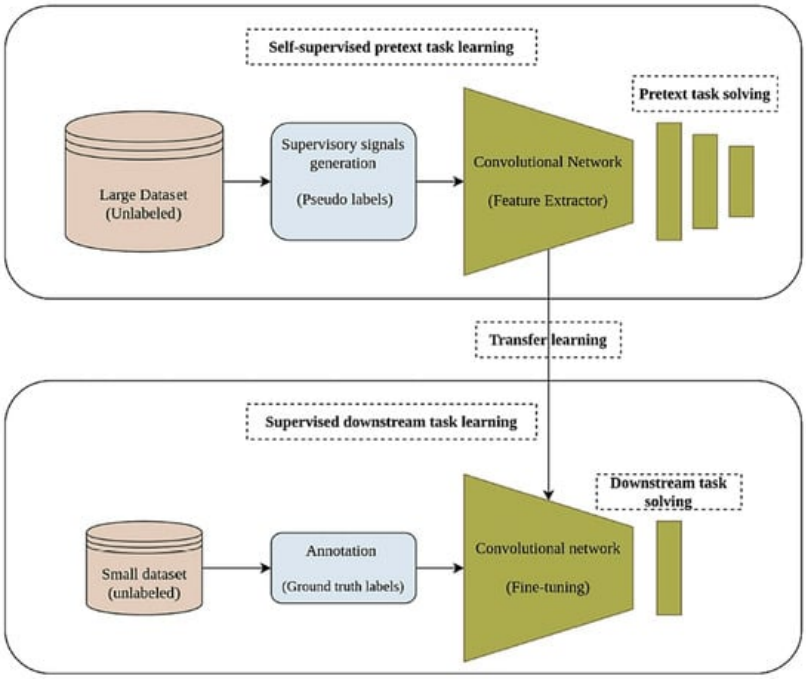

Figure 1. The major workflow related to SSL [10].

From **Figure 1**, during the pre-training stage (pretext task solving), feature extraction is carried out by pseudo-labels to enable an efficient prediction process. After that, transfer learning is implemented to initiate the SSL phase, in which a small dataset is considered to make data annotations (of ground truth labels). Then, fine-tuning is performed to achieve the necessary prediction task.

3. Main SSL Categories

Because it can be laborious to compile an extensively annotated dataset for a given prediction task, USL strategies have been proposed as a means of learning appropriate image identification without human guidance [11][12]. Simultaneously, SSL is an efficient approach through which a training objective can be produced from the data. Theoretically, a deep neural network (DNN) is trained on pretext tasks, in which labels are automatically produced without human annotation. The learned representations can be utilized to complete the pretext tasks. Familiar SSL sorts involve: (A) generative, (B) predictive, (C) contrastive, and (D) non-contrastive models. The multiple contrastive and noncontrastive tactics illustrated here can be recognized as joint-embedded strategies.

However, more types of SSL are considered in some contexts. For example, a graphical illustration in [13] was created, explaining the performance rates that can be achieved when SSL is applied, focusing mainly on further SSL categories, as shown in **Figure 2**.

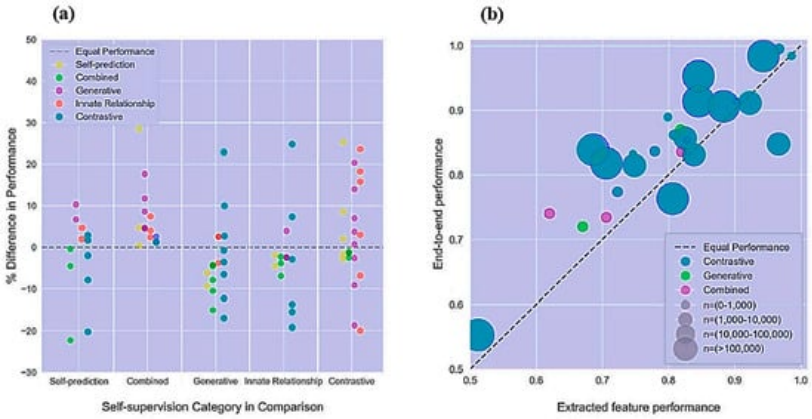

Figure 2. Two profiles of (**a**) some types of SSL classification with their performance level and (**b**) end-to-end performance of extracted features and the corresponding number of each type [13].

It can be realized from the graphical data expressed in **Figure 2**a that the variation in the performance between the self-prediction, combined, generative, innate, and contrastive SSL types fluctuates mostly between 10% and 10%. In **Figure 2**b, it can be noticed that end-to-end performance corresponding to contrastive, generative, and combined SSL algorithms

varies between nearly 0.7 and 1.0, relating to an extracted feature performance that ranges approximately between 0.7 and 1.0.

In the following sections, more explanation is provided for some of these SSL categories.

3.1. Generative SSL Models

Using an autoencoder to recreate an input image following compression is a common pretext operation. Relying on the first component of the network, called an encoder, the model should learn to compress all pertinent data from the image into a latent space with reduced dimensions to minimize the reconstruction loss. The image is then reconstructed by the latent space of a second network component called a decoder.

Researchers in [11][12][14][15][16][17][18] reported that denoising autoencoders could also provide reliable and stable identifications of images by learning to filter out noise. The network cannot learn the identity function owing to extra noise. By encoding the distribution parameters of a latent space, variational autoencoders (VAE) can advance the autoencoder model [19][20][21][22]. Both the reconstruction of error and extra factor, the Kull-Leibler divergence between an established latent distribution (often a unit-centered Gaussian distribution), and the encoder output are minimized during training. The samples from the resulting distribution can be obtained through this regularization of the latent space. To rebuild entire patches with only 25 percent of the visible patches, scholars in [23][24] have recently adopted vision transformers to create large masked autoencoders that work at the patch level rather than pixel-wise. Adding a class token to a sequence of patches or performing global mean pooling on all the patch tokens, as in this reconstruction challenge, yields reliable image representations.

A generative adversarial network (GAN) is another fundamental generative USL paradigm that has been extensively studied [25][26][27]. This architecture and its variants aim to mimic real data's appearance and behavior by generating new data from random noise. To train a GAN, two networks compete in an adversarial minimax game, with one learning to turn the rate of random noise, $\Psi RN \approx RN(0, 1)$

into synthetic data, SD
, which attempts to mimic the distribution of the original data. These aspects can be illustrated in **Figure 3**.

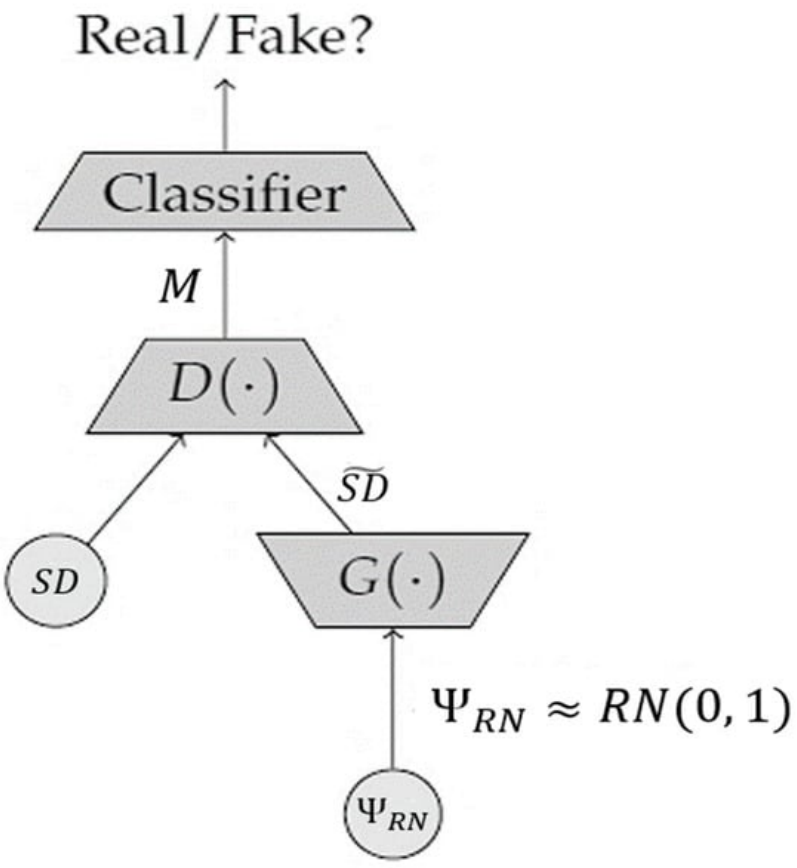

Figure 3. The architecture employed in GAN. Adopted from Ref. [28], used under Creative Commons CC-BY license.

In the adversarial method, a second network, which can be termed discriminator D(.) was trained to distinguish between generated and authentic images from the original dataset. When the discriminator is certain that the input image is from

the true data distribution, it reports a score of 1, whereas for the images produced by the generator, the score is zero. One possible estimation of this adversarial objective function, F_{AO}, can be accomplished by the following mathematical formula:

$$F_{AO} = \min_{G} \max_{D} \frac{1}{N} \sum_{i=1}^{N} \log\left(1 - D\left(G\left(\Psi_{RN_i}\right)\right)\right) + \frac{1}{M} \sum_{i=1}^{M} \log\left(D\left(x_i\right)\right),$$

where:

- $\Psi_{RN} \approx RN(0, 1)$—A group of random noise vectors with an overall amount of N
- $SD \approx Q_{Data}$—A dataset comprising a set of real images having a total number of M.

3.2. Predictive SSL Paradigms

Models trained to estimate the impact of artificial change on the input image express the second type of SSL technique. This strategy is inspired by understanding the semantic items and regions inside an image, which can be essential for accurately predicting the transformation. Scholars in [29] conducted analytical research to improve the performance of the model against random initialization and to approach the effectiveness obtained from the initialization with ImageNet pre-trained weights in benchmark computer vision datasets by pre-training a paradigm to predict the relative positions of two image patches.

Some researchers have confirmed the advantages of colored images [30]. In this method, the input image is first changed to grayscale. Next, a trained autoencoder converts the grayscale image back to its original color form by minimizing the average squared error between the reconstructed and original images. The encoder feature representations are considered in the subsequent downstream processes. The numerical RotNet approach [31] is another well-known predictive SSL approach, which represents a practical training process for mathematical schemes to help predict the rotation that is randomly implemented in the input image, as shown in **Figure 4**.

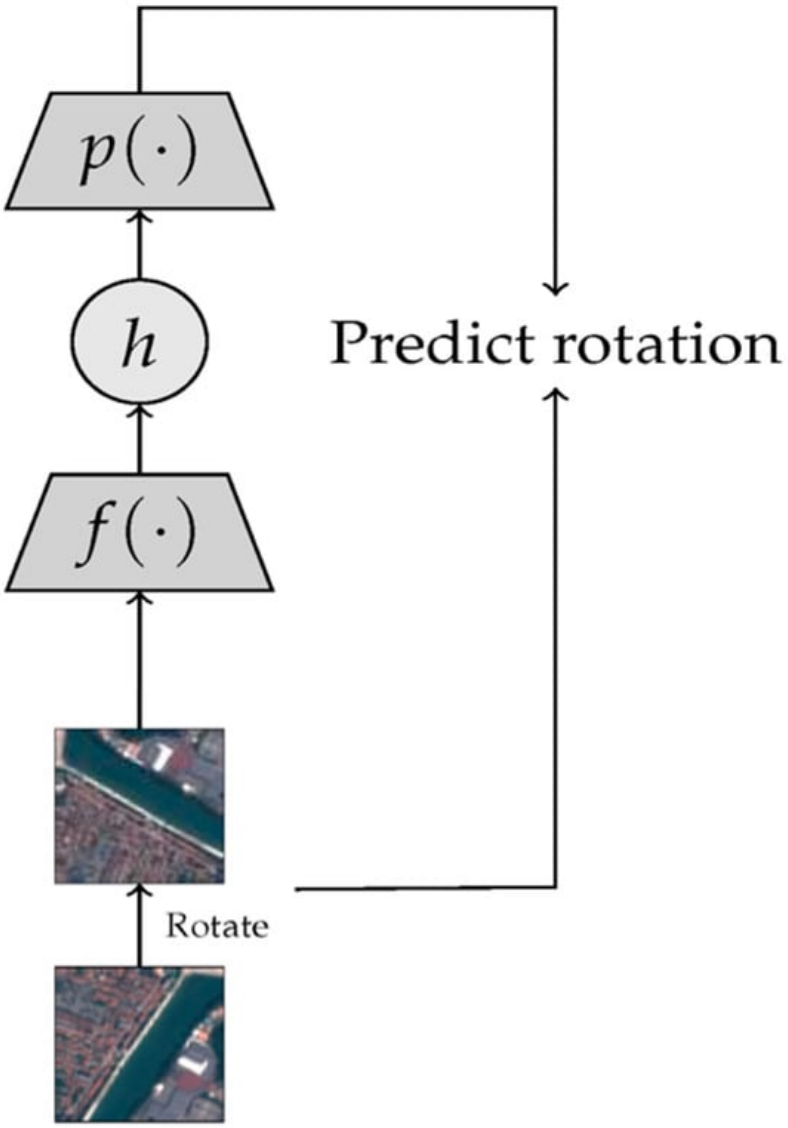

Figure 4. Flowchart configuration of the operating principles related to the intelligent RotNet algorithm relying on the SSL approach for accurate prediction results. From Ref. [28], used under Creative Commons CC-BY license.

To improve the performance of the model in a dynamic rotation prediction task, the relevant characteristics that classify the semantic content of the image should first be extracted. Researchers in [32] considered a jigsaw puzzle to forecast the relative positions of the picture partitions using the shuffled SSL model. The Exemplar CNN was also addressed and trained in [33] to predict augmentations that can be applied to images by considering a wide variety of augmentation types. Cropping, rotation, color jittering, and contrast adjustment are examples linked to the enhancement classes gained by the Exemplar CNN model.

An SSL model can learn rich representations of the visual content by completing one of these tasks. However, the network may not be able to perform effectively on all subsequent tasks contingent on the pretext task and dataset. Because the orientation of objects is not as rigorously practical to handle in remote sensing datasets as in object-centric datasets, the prediction of random rotations of an image would not perform particularly well on such a dataset [34].

3.3. Contrastive SSL Paradigms

Forcing the features of various perspectives in a picture to be comparable is another strategy that can result in accurate representations. The resulting representations are independent of the particular enhancements needed to generate various image perspectives. However, the network can be converged to a stable representation that meets the invariance condition but is unrelated to the input image.

One typical approach to achieving this goal via the acquisition of various representations while avoiding the collapse problem is the contrastive loss. This type of loss function can be utilized to train the model to distinguish between views of the same image (positive) and views of distinct images (negative). Correspondingly, it seeks to obtain homogeneous feature representations for pairs with positive values while isolating features for negative pairs. The triplet loss investigated by researchers in [35] is the simplest form of this family. It requires a model to be trained such that the

distance between the representations of a given anchor and its positive rates is smaller than the distance between the representations of the anchor and the random negative, as illustrated in **Figure 5**.

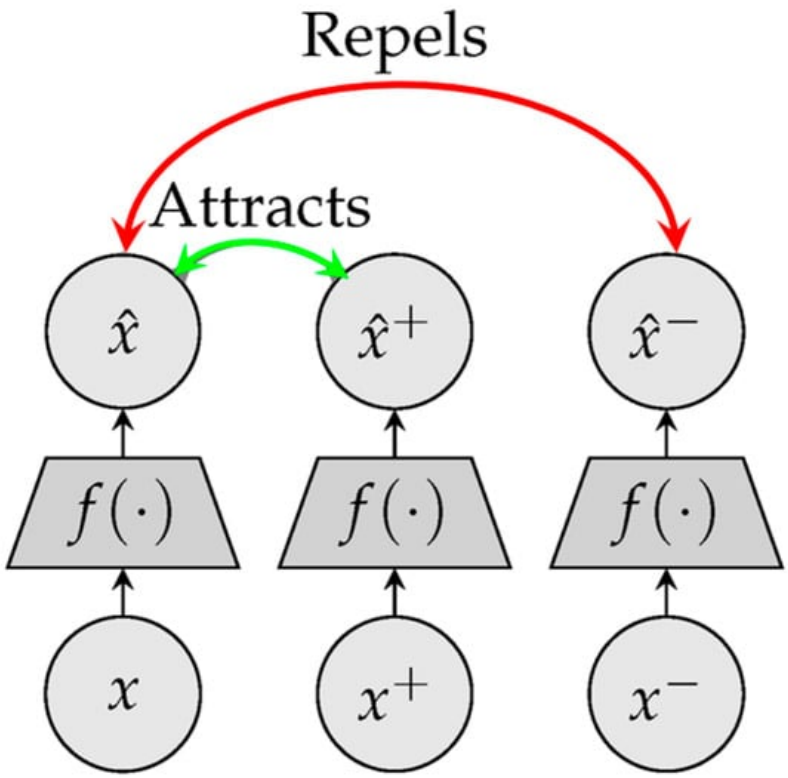

Figure 5. The architecture of the triplet loss function. From Ref. [28], used under Creative Commons CC-BY license.

In **Figure 5**, the triplet loss function is considered helpful in learning discriminative representations by learning an encoder that is able to detect the difference between negative and positive samples. Under this setting, the triplet Loss Function, $\mathfrak{F}_{LossTriplet}$, can be estimated using the following relationship:

$$\mathfrak{F}_{Loss\,Triplet} = \max\left(\left\|f\left(x\right) - f\left(x^+\right)\right\| - \left\|f\left(x\right) - f\left(x^-\right)\right\| + m, 0\right),$$

where:

- x+—The positive vector value of the anchor *x*
- x−—The negative vector value of the anchor *x*
- f(.)—The embedding function
- m—The value of the margin parameter.

In [36], the researchers examined the SimCLR method, which is one of the most well-known SSL strategies. It formulates a type of contrastive representational learning. Two versions of each training batch image can be generated using random-sampling augmentation. After these modified images are fed into the representational method, a prediction network can be utilized to map the representation onto a hypersphere of dimension, D.

The overall mathematical algorithm is trained to elevate the cosine similarity across the representation parameter, z, and its corresponding positive counterpart, z+ (belonging to the same original visual data) and to minimize the similarity between z and all other representations in the batch z−, contributing to the following expression:

$$\mathbb{L}_F\left(z, z^+, z^-\right) = -\log\left(\frac{\exp\left(\langle z, z^+\rangle/\tau\right)}{\sum_{z' \in z^- \cup \{z^+\}} \exp\left(\frac{\langle z, z'\rangle}{\tau}\right)'}\right),$$

where:

- $\langle z,z+\rangle$ —the dot product between z and z^+.
- τ —the temperature variable to scale the levels of similarity, distribution, and sharpness.
- $f(.)$—the embedding function.

At the same time, the algebraic correlation connected with the evaluation process of the complete loss function that assesses the cross-entropy of temperature, which can be dominated as normalized temperature cross-entropy, which is denoted by $N\Theta$

-$\mathfrak{X}$S, is depicted in the following relation:

$$\mathbb{L}\left(N\Theta - \mathfrak{X}\mathrm{S}\right) = \frac{1}{2N}\sum_{z,z^+,z^-}\mathbb{L}_F\left(z,z^+,z^-\right),$$

where N indicates the number of items in the dataset, such as images and textual characters.
Figure 6 shows that the NT-Xent loss [37] acts solely on the direction of the features confined to the D-dimensional hypersphere because the representations are normalized before calculating the function loss value.

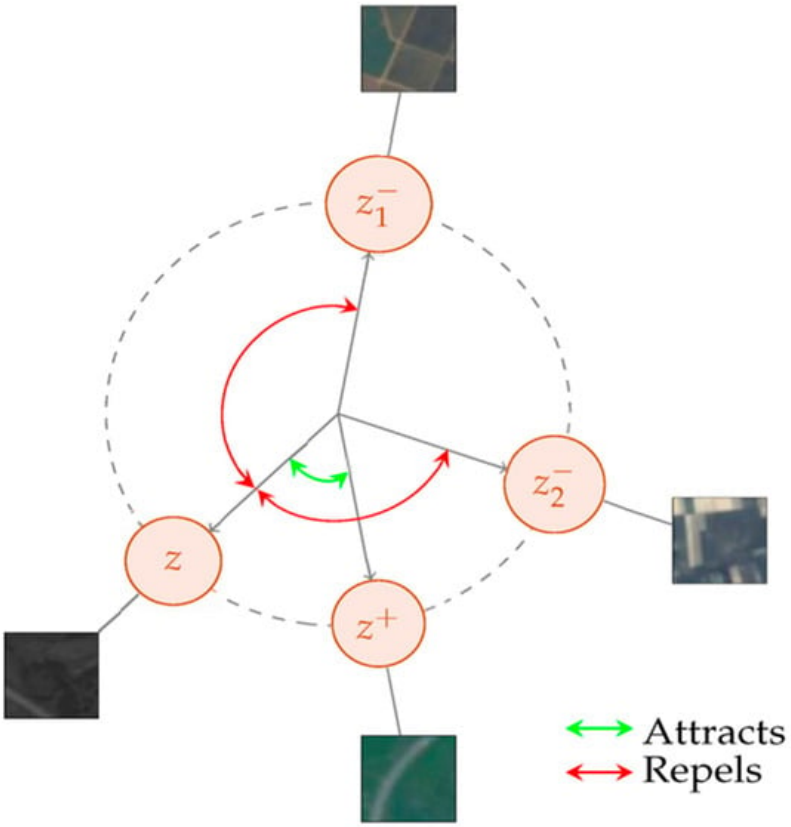

Figure 6. Description of contrastive loss linked to the 2-dimensional unit sphere with two negative parameters (z_1^- and z_2^-) and one positive (z^+) sample from the EuroSAT dataset. From Ref. [28], used under Creative Commons CC-BY license.

By maximizing the mutual data between the two perspectives, this loss ensures that the resulting representations are both style-neutral and content-specific.

In addition to SimCLR, they suggested the momentum contrast (MoCo) technique, which uses a reduced number of batches to calculate the contrastive loss while maintaining the same functional number of negative samples [38]. It employs an exponentially moving average (EMA)-updated momentum encoder whose values are updated by the main encoder's weights and a sample queue to increase the number of negative samples in each batch, as shown in **Figure 7**. To account for the newest positives, the oldest negatives from the previous batch were excluded. Other techniques, such as swapping assignments across views (SwAVs), correlate views to consistent clusters between positive pairs by clustering representations into a shared set of prototypes [37][39][40][41]. The entropy-regularized optimal transport strategy

is also used in the same context to move representations between clusters in a manner that prevents them from collapsing into one another [39][42][43][44][45][46]. Finally, the cross-entropy between the optimal tasks in one branch and the anticipated distribution in the other is minimized by the loss. To feed sufficient negative samples to the loss function and avoid representations from collapsing, contrastive approaches often need large batch sizes.

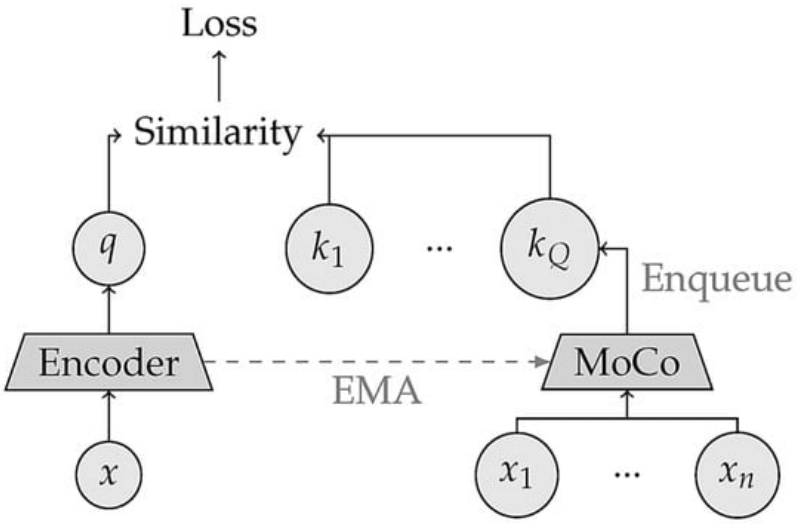

Figure 7. An illustration of Q samples developed utilizing a momentum encoder whose amounts are modified. From Ref. [28], used under Creative Commons CC-BY license.

As shown in **Figure 7**, at each step of the numerical analysis, only the major encoder amounts are updated based on the backpropagation process. The similarity aspects between the queue and encoded batch samples were then employed for contrastive loss.

Compared with traditional prediction methods, joint-embedding approaches tend to generate broader representations. Nonetheless, their effectiveness in downstream activities may vary depending on the augmentation utilized. If a model consistently returns the same representations for differently cropped versions of the same image, it can effectively remove any spatial information about the image and will likely perform poorly in tasks such as semantic segmentation and object detection, which rely on this spatial information. Dense contrastive learning (DCL) has been proposed and considered by various researchers to address this issue [47][48][49][50]. Rather than utilizing contrastive loss on the entire image, it was applied to individual patches. This permitted the contrastive model to acquire representations that are less prone to spatial shifts.

3.4. Non-Contrastive SSL Models

To train self-supervised models, alternative methods within joint-embedded learning frameworks can prevent the loss of contrastive elements. They classified these as approaches that do not rely on contrast. Bootstrap Your Own Latent (BYOL) is a system based on mentor-apprentice pairing [51][52][53]. The student network in a teacher-student setup is taught to mimic the teacher network's output (or characteristics). This method is frequently utilized in knowledge distillation when the instructor and student models possess distinct architectures (e.g., when the student model is substantially smaller than the teacher model) [54]. The weights of the instructor network in BYOL are defined as the EMA of the student network weights. Two projector networks, gA

and gB, are utilized after the encoders, fA and fB, to calculate the training loss. Subsequently, to extract representations at the image level, they retrain only the student encoder fA
. Additional asymmetry is introduced between the two branches by a predictor network superimposed on the student projector, as shown in **Figure 8**.

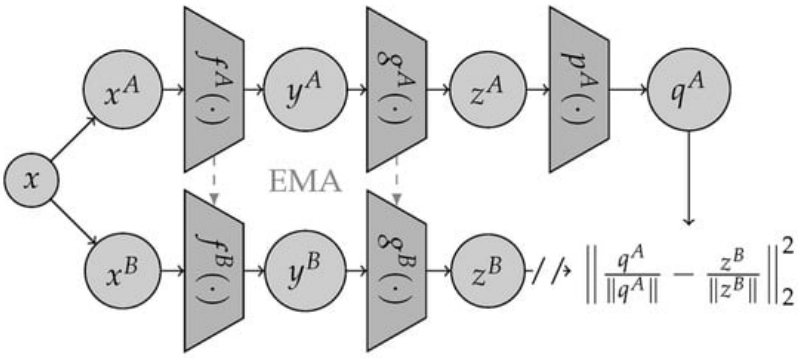

Figure 8. Architecture of the non-contrastive BYOL method, considering student *A* and lecturer *B* pathways to encode the dataset. From Ref. [28], utilized under Creative Commons CC-BY license.

In **Figure 8**, the teacher's values are modified and updated by the EMA technique applied to the student amounts. The online branch is also supported by an additional network, p^A, which is known as the predictor [53].

SimSiam employs a pair of mirror-image networks and a predictor network at the end of a node [55][56][57]. The loss function employs an asymmetric stop gradient to optimize the pairwise alignments between positive pairs because the two branches have identical weights. Relying on a student-teacher transformer design known as self-distillation, DINO (self-distillation with no labels) defines the instructor as an EMA of the weights in the student network [58]. Next, the teacher network's centered and sharpened outputs are utilized to train the student network to make exact predictions for a given positive pair.

Another non-contrastive learning model, known as the Barlow Twins, can be offered according to the information bottleneck theory, which eliminates the need for individual amounts for each branch of the teacher-student model considered in BYOL and SimSiam [59][60]. This technique enhances the mutual information between two perspectives by boosting the cross-correlation of the matching characteristics provided by two identical networks and eliminating superfluous information in these representations. The Barlow twin loss function was evaluated by the following equation:

$$\mathscr{L}_{Barlow.Twins} = \sum_{i=1}^{N} \left(1 - C_{ii}^2\right) + \lambda \sum_{i=1}^{N} \sum_{j \neq i} C_{ij}^2,$$

where C is the cross-correlation matrix calculated by the following formula:

$$C_{ij} = \frac{\sum_b z_{bi}^A z_{bj}^B}{\sqrt{\sum_b \left(z_{bj}^A\right)^2} \sqrt{\sum_b \left(z_{bj}^B\right)^2}},$$

where z^A and z^A express the corresponding outcomes related to the two identical networks provided by the two views of a particular photograph.

Variance, invariance, and covariance regularization (VICReg) approaches have been recently proposed to enhance this framework [61][62][63][64]. In addition to invariance, which implicitly maximizes alignments between positive pairs, the loss terms are independent for every branch, unlike in low twins. Using distinct regularization for each pathway, this method allows for noncontrastive multimodal pre-training between text and photo pairs.

Most of these techniques train a linear classifier on the priority of representations as the primary performance metric. Researchers in [63] analyzed the beneficial impacts of ImageNet, whereas scholars in [62][65] examined CIFAR's advantages, which help accomplish an active analysis of object-centric visual datasets commonly addressed for the pre-training and linear probing phases of DL. Therefore, these techniques may not apply to image classification.

Scholars are invited to examine devoted review articles for further contributory information and essential fundamentals pertaining to SSL types [61][66].

4. Practical Applications of SSL Models

Before introducing the common applications and vital utilizations of SSL models to handle efficacious data classification and identification processes, their critical benefits should be identified as a whole. The commonly-addressed benefits and vital advantages of SSL techniques can be expressed as follows [67][68]:

- Minimizing the massive cost connected with data labeling phases is essential to facilitating a high-quality classification/prediction process.
- Alleviating the corresponding time needed to classify/recognize vital information in a dataset,
- Optimizing the data preparation lifecycle is typically a lengthy procedure in various ML models. It relies on filtering, cleaning, reviewing, annotating, and reconstructing processes through training phases.
- Enhancing the effectiveness of AI models. SSL paradigms can be recognized as functional tools that allow flexible involvement in innovative human thinking and machine cognition.

According to these practical benefits, further workable possibilities and effective prediction and recognition impacts can be explained in the following paragraphs, which focus mainly on medical and engineering contexts.

4.1. SSL Models for Medical Predictions

Krishnan et al. (2022) [69] analyzed SSL models' application in medical data classification, highlighting the critical challenges of manual annotation of vast medical databases. They addressed SSL's potential for enhancing disease diagnosis, particularly in EHR and some other visual clinical datasets. Huang et al. (2023) [13] conducted a systematic review affirming SSL's benefits in supporting medical professionals with precise classification and therapy identification from visual data, reducing the need for extensive manual labeling.

Figure 9 shows the number of DL, ML, and SSL research articles published between 2016 and 2021.

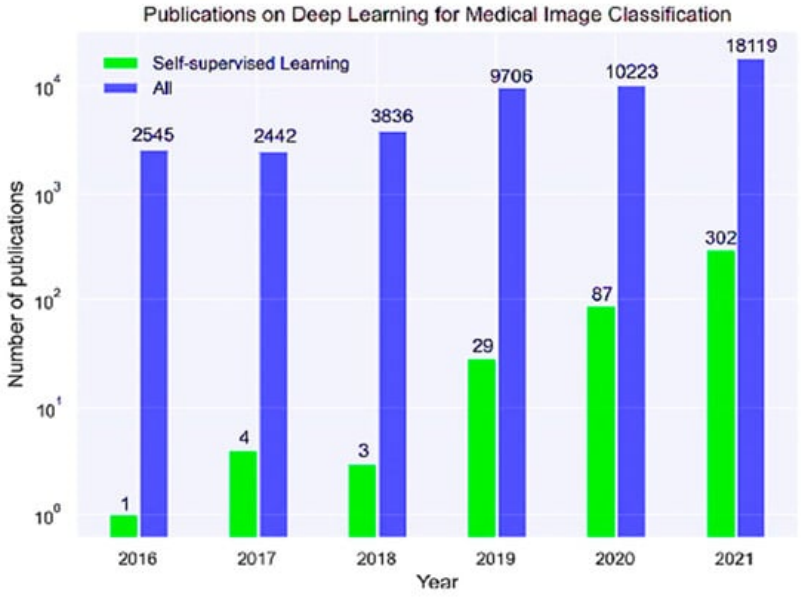

Figure 9. The number of articles on SSL, ML, and DL models utilized for medical data classification [13].

It can be concluded from the statistical data explained in **Figure 9** that the number of research publications addressing ML and DL models' importance and relevance in the medical classification has been increasing per year. Similarly, the increasing trend was for the overall number of academic articles investigating the SSL, ML, and DL algorithms in conducting high-performance identification of problems in images of patients.

Besides these numeric figures, an explanation of the pre-training process of SSL and fine-tuning can be expressed in **Figure 10**.

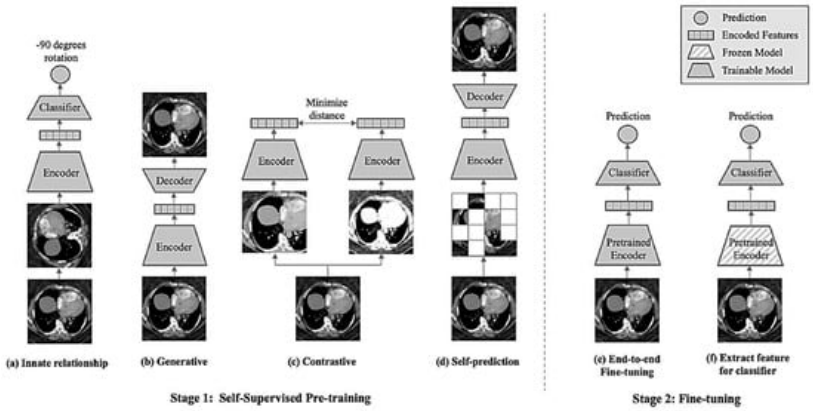

Figure 10. The two stages of pre-training and fine-tuning are considered in the classification of visual data [13].

It can be inferred from the data explained in **Figure 10** that the pre-training SSL process takes into account four critical types to be accomplished, including (**Figure 10**a) innate relationship, (**Figure 10**b) generative, (**Figure 10**c) contrastive, and (**Figure 10**d) self-prediction. At the same time, there are two categories included in the fine-tuning process, which comprise end-to-end and feature extraction procedures.

Before the classification process is done, SSL models are first trained. This step is followed by the encoding of image features. It follows the adoption of the classifier, which is important to enable precise prediction of the medical problem in the image.

In their overview [13], the scholars identified a collection of some medical disciplines in which SSL models can be advantageous in conducting the classification process flexibly, which can be illustrated in **Figure 11**.

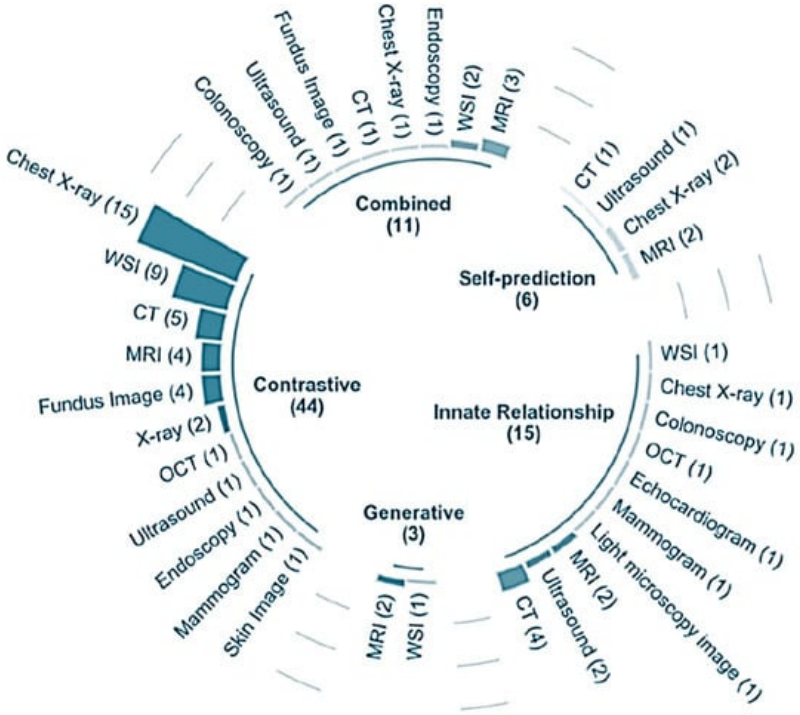

Figure 11. The major categories correlated with medical classification can be done by SSL models [13].

From the data expressed in **Figure 11**, it can be inferred that the possible medical classification types and dataset categories are numerous in SSL models that can be applied reliably for efficient classification. As a result, this aspect makes SSL models more practical and feasible for carrying out robust predictions of problems in clinical datasets.

Various studies have explored the application of SSL models in medical data classification, showcasing their efficacy in improving diagnostic accuracy and efficiency. Azizi et al. (2021) [70] demonstrated the effectiveness of SSL algorithms in classifying medical disorders within visual datasets, particularly highlighting advancements in dermatological and chest X-ray recognition. Zhang et al. (2022) [71] utilized numerical simulations to classify patient illnesses on X-rays, emphasizing the importance of understanding medical images for clinical knowledge. Bozorgtabar et al. (2020) [72] addressed the challenges of data annotation in medical databases by employing SSL methods for anomaly classification in X-ray

images. Tian et al. (2021) [73] identified clinical anomalies in fundus and colonoscopy datasets using SSL models, emphasizing the benefits of unsupervised anomaly detection in large-scale health screening programs. Ouyang et al. (2021) [74] introduced longitudinal neighborhood embedding SSL models for classifying Alzheimer's disease-related neurological problems, enhancing the understanding of brain disorders. Liu et al. (2021) [75] proposed an SSMT-SiSL hybrid model for chest X-ray data classification, highlighting the potential of SSL techniques to expedite data annotation and improve model performance. Li et al. (2021) [76] addressed data imbalances in medical datasets with an SSL approach, enhancing lung cancer and brain tumor detection. Manna et al. (2021) [77] demonstrated the practicality of SSL pre-training in improving downstream operations in medical data classification. Zhao and Yang (2021) [78] utilized radiomics-based SSL approaches for precise cancer diagnosis, showcasing SSL's vital role in medical classification tasks.

4.2. SSL Models for Engineering Contexts

In the field of engineering, SSL models may provide contributory practicalities, especially when prediction tasks in mechanical, industrial, electrical, or other engineering domains are necessary without the need for massive data annotations to train and test conventional models to accomplish this task accurately and flexibly.

In this context, Esrafilian and Haghighat (2022) [79] explored the critical workabilities of SSL models in providing sufficient control systems and intelligent monitoring frameworks for heating, ventilation, and air-conditioning (HVAC) systems. Typically, ML and DL models may not contribute to noteworthy advantages since complicated relationships, patterns, and energy consumption behaviors are not directly and clearly provided. The controller was created by employing a model-free reinforcement learning technique recognized with a double-deep Q-network (DDQN). Long et al. (2023) [80] proposed an SSL-based defect prognostics-trained DL model, SSDL, addressing the challenges of costly data annotation in industrial health prognostics. SSDL dynamically updates a sparse auto-encoder classifier with reliable pseudo-labels from unlabeled data, enhancing prediction accuracy compared with static SSL frameworks. Yang et al. (2023) [81] developed an SSL-based fault identification model for machine health prognostics, leveraging vibrational signals and one-class classifiers. Their SSL model, utilizing contrastive learning for intrinsic representation derivation, outperformed novel numerical models in fault prediction accuracy during simulations. Wei et al. (2021) [82] utilized SSL models for rotary machine failure diagnosis, employing 1-D SimCLR to efficiently encode patterns with a few unlabeled samples. Their DTC-SimCLR model combined data transformation combinations with a fixed feature encoder, demonstrating effectiveness in diagnosing cutting tooth and bearing faults with minimal labeled data. Overall, DTC-SimCLR had improved diagnosis accuracy and fewer samples. **Figure 12** depicts a low-sample machine failure diagnosis approach.

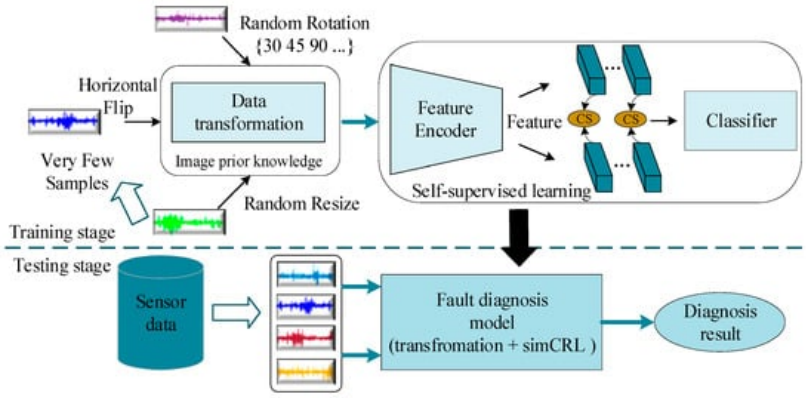

Figure 12. The formulated system for machine failure diagnosis needs very few samples [82].

Furthermore, the procedure related to the SSL in SimCLR can be expressed in **Figure 13**.

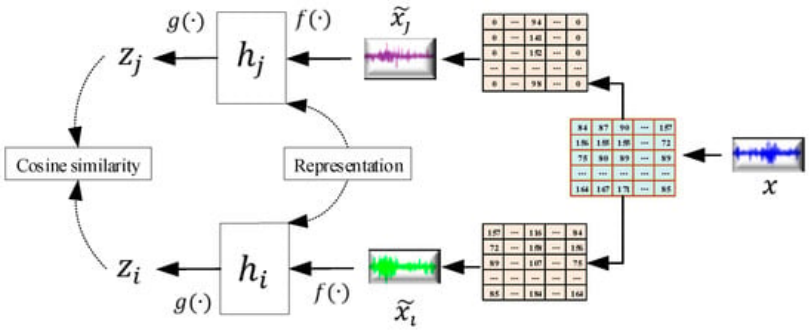

Figure 13. The procedure related to the SSL in SimCLR [82].

Simultaneously, **Table 1** indicates the critical variables correlated with the 1D SimCLR.

Table 1. The major variables linked to the 1D SimCLR [82].

No.	Variable Category	Magnitude
1	Input Data	A Length of 1024 Data Points
2	Temperature	10
3	Feature Encoder	Sixteen Convolutional Layers
4	Output Size	128
5	Training Epoch	200

Above these examples, Lei et al. (2022) [83] addressed SSL models in predicting the temperature of aluminum correlated with industrial engineering applications. Through their numerical analysis, they examined how changing the temperature of the pot or electrolyte could affect the overall yield of aluminum during the reduction process through their proposed deep long short-term memory (D-LSTM).

On the other side, Xu et al. (2022) [84] identified the contributory rationale of functional SSL models to offer alternative solutions to conventional human defect detection methods that became insufficient. Bharti et al. (2023) [85] remarked that deep SSL (DSSL) contributed to significant relevance in the industry owing to its potency in reducing the time and effort required by humans for data annotation by manipulating operational procedures carried out by robotic systems, taking into account the CIFAR-10 dataset. Hannan et al. (2021) [86] implemented SSL prediction to estimate the state of charge (SOC) correlated with lithium-ion (Li-ion) batteries precisely in electric vehicles (EVs) to ensure their maximum cell lifespan.

4.3. Patch Localization

Regarding the critical advantages and positive gains of SSL models in conducting active processes of patch localization, several authors confirmed the significant effectiveness and valuable merits of innovative SSL schemes in accomplishing optimum activities of recognition and detection related to a defined dataset of patches. For instance, Li et al. (2021) [87] estimated the substantial contributions of SSL in identifying visual defects or irregularities in an image without relying on abnormal training data. The patch localization of visual defects involves grid classes, wood, screws, metal nuts, hazelnuts, and bottles.

Although SSL has made great strides in the field of image classification, there is moderate effectiveness in making precise object recognition. Through their analysis, Yang et al. (2021) [88] aimed to improve self-supervised, pre-trained models for object detection. They proposed a novel self-supervised pretext algorithm called instance localization, proposing an augmentation strategy for the image-bounding boxes. Their results confirmed that their pre-trained algorithm for object detection was improved, but it became less effective in ImageNet semantic classification and more so in image patch localization. Object detection considering the PASCAL VOC and MSCOCO datasets revealed that their method achieved state-of-the-art transfer learning outcomes.

The red box in their result, expressed in **Figure 14**, indicates the base truth bounding box linked to the foreground image. However, the right-hand photo shows a group of anchor boxes positioned in the central area related to a singular spatial location. By improving the multiple anchors using variant scales, positions, and aspect ratios, the base truth pertaining to the blue boxes can be augmented, offering an intersection over union (IoU) level greater than 0.5.

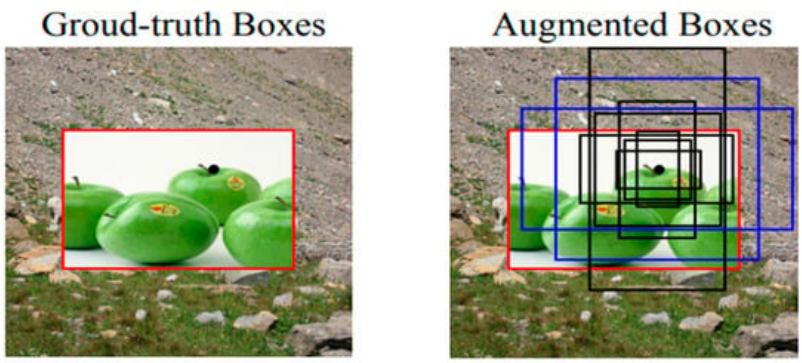

Figure 14. Bounding boxes addressed for spatial modeling [88].

To train an end-to-end model for anomaly identification and localization using only normal training data, Schlüter et al. (2022) [89] created a flexible self-supervision patch categorization model called natural synthetic anomalies (NSA). Their NSA harnessed Poisson photo manipulation to combine scaled patches of varying sizes from multiple photographs into a single coherent entity. Compared with other data augmentation methods for unsupervised anomaly identification, this aspect helped generate a wider variety of synthetic anomalies that were more akin to natural sub-image inconsistencies. Natural and medical images were employed to test their proposed technique, including the MVTec AD dataset, indicating the efficient capability of identifying various unknown manufacturing defects in real-world scenarios.

4.4. Context-Aware Pixel Prediction

Learning visual representations from unlabeled photographs has recently witnessed a rapid evolution owing to self-supervised instance discrimination techniques. Nevertheless, the success of instance-based objectives in medical imaging is unknown because of the large variations in new patients' cases compared with previous medical data. Context-aware pixel prediction focuses on understanding the most discriminative global elements in an image (such as the wheels of a bicycle). According to the research investigation conducted by Taher et al. (2022) [90], instance discrimination algorithms have poor effectiveness in downstream medical applications because the global anatomical similarity of medical images is excessively high, resulting in complicated identification tasks. To address this shortcoming, scholars have innovated context-aware instance discrimination (CAiD), a lightweight but powerful self-supervised system, considering: (a) generalizability and transferability; (b) separability in embedding space; and (c) reusability and systematic reusability. The authors addressed the dice similarity coefficient (DSC) as a measure related to the similarity between two datasets that are often indicated as binary arrays. Similarly, authors in [91] proposed a teacher-student strategy for representation learning, wherein a perturbed version of an image serves as an input for training a neural net to reconstruct a bag-of-visual-words (BoW) representation referring to the original image. The BoW targets are generated by the teacher network, and the student network learns representations while simultaneously receiving online training and an updated visual word vocabulary.

Liu et al. (2018) [50] distinguished some beneficial yields of SSL models in identifying information from defined datasets of context-aware pixel databases. To train the CNN models necessary for depth evaluation from monocular endoscopic data without a priori modeling of the anatomy or coloring, the authors implemented the SSL technique, considering a multiview stereo reconstruction technique.

4.5. Natural Language Processing

Fang et al. (2020) [8] considered SSL to classify essential information in certain defined datasets related to natural language processing. Scholars explained that pre-trained linguistic models, such as bidirectional encoder representations from transformers (BERTs) and generative pre-trained transformers (GPTs), have proved their considerable effectiveness in executing active linguistic classification tasks. Existing pretraining techniques rely on auxiliary prediction tasks based on tokens, which may not be effective for capturing sentence-level semantics. Thus, they proposed a new approach that recognizes contrastive self-supervised encoder representations using transformers (CERTs). Baevski et al. (2023) [92] highlighted critical SSL models' relevance to high-performance data identification correlated with NLP. They explained that currently available techniques of unsupervised learning tend to rely on resource-intensive and modal-specific aspects. They added that the Data2vec model expresses a practical learning paradigm that can be generalized and broadened across several modalities. Their study aimed to improve the training efficiency of this model to help handle the precise classification of NLP problems. Park and Ahn (2019) [93] inspected the vital gains of SSL to lead to efficient detection of NLP. Researchers proposed a new approach dedicated to data augmentation that considers the intended context of the data. They suggested a label-asked language model (LMLM), which can effectively employ the masked language model (MLM) in data with label information by including label data for the mask tokens adopted in the MLM. Several text classification benchmark datasets were examined in their work, including the Stanford sentiment treebank-2 (SST2), multi-perspective question answering (MPQA), text retrieval conference (TREC), Stanford sentiment treebank-5 (SST5), subjectivity (Subj), and movie reviews (MRs).

4.6. Auto-Regressive Language Modeling

Elnaggar et al. (2022) [94] published a paper shedding light on valuable SSL roles in handling the active classification of datasets connected to the modeling of auto-regressive language. The scholars trained six models, four auto-encoders (BERT, Albert, Electra, and T5), and two auto-regressive prototypes (Transformer-XL and XLNet) on up to 393 billion amino acids from UniRef and BFD. The Summit supercomputer was utilized to train the protein LMs (pLMs), which

required 5616 GPUs and a TPU Pod with up to 1024 cores. Lin et al. (2021) [95] performed numerical simulations, exploring the added value of three SSL models, notably (I) autoregressive predictive coding (APC), (II) contrastive predictive coding (CPC), and (III) wav2vec 2.0, in performing flexible classification and reliable recognition of datasets engaged in auto-regressive language modeling. Several any-to-any voice conversion (VC) methods have been proposed, like AUTOVC, AdaINVC, and FragmentVC. To separate the feature material from the speaker information, AUTOVC and AdaINVC utilize source and target encoders. They proposed a new model, known as S2VC, which harnesses SSL by considering multiple features of the source and target linked to the VC model. Chung et al. (2019) [96] proposed an unsupervised auto-regressive neural model to help students learn generalized representations of speech. Their speech representation learning approach was developed to maintain information for various downstream applications to remove noise or speaker variability.

5. Commonly-Utilized Feature Indicators of SSL Models' Performance

Specific formulas in [97][98] were investigated to examine different SSL paradigms in carrying out their classification task, particularly the prediction and identification of faults and errors in machines, which can support maintenance specialists in selecting the most appropriate repair approach. These formulas formulate practical feature indicators to monitor signals that can be prevalently utilized by maintenance engineers to identify the health state of machines. Twenty-four typical feature indicators were addressed, referring to Zhang et al. (2022) [99]. These indices can enable maintenance practitioners to locate optimum maintenance strategies to apply to industrial machinery, helping to handle current failure issues flexibly.

References

1. Lai, Y. A Comparison of Traditional Machine Learning and Deep Learning in Image Recognition. J. Phys. Conf. Ser. 2019, 1314, 012148.
2. Rezaeianjouybari, B.; Shang, Y. Deep learning for prognostics and health management: State of the art, challenges, and opportunities. Measurement 2020, 163, 107929.
3. Thoppil, N.M.; Vasu, V.; Rao, C.S.P. Deep Learning Algorithms for Machinery Health Prognostics Using Time-Series Data: A Review. J. Vib. Eng. Technol. 2021, 9, 1123–1145.
4. Zhang, L.; Lin, J.; Liu, B.; Zhang, Z.; Yan, X.; Wei, M. A Review on Deep Learning Applications in Prognostics and Health Management. IEEE Access 2019, 7, 162415–162438.
5. Deng, W.; Nguyen, K.T.P.; Medjaher, K.; Gogu, C.; Morio, J. Bearings RUL prediction based on contrastive self-supervised learning. IFAC-PapersOnLine 2023, 56, 11906–11911.
6. Akrim, A.; Gogu, C.; Vingerhoeds, R.; Salaün, M. Self-Supervised Learning for data scarcity in a fatigue damage prognostic problem. Eng. Appl. Artif. Intell. 2023, 120, 105837.
7. Nadif, M.; Role, F. Unsupervised and self-supervised deep learning approaches for biomedical text mining. Brief. Bioinform. 2021, 22, 1592–1603.
8. Fang, H.; Wang, S.; Zhou, M.; Ding, J.; Xie, P. Cert: Contrastive self-supervised learning for language understanding. arXiv 2020, arXiv:2005.12766.
9. Jaiswal, A.; Babu, A.R.; Zadeh, M.Z.; Banerjee, D.; Makedon, F. A Survey on Contrastive Self-Supervised Learning. Technologies 2020, 9, 2.
10. Shurrab, S.; Duwairi, R. Self-supervised learning methods and applications in medical imaging analysis: A survey. PeerJ Comput. Sci. 2022, 8, e1045.
11. Ohri, K.; Kumar, M. Review on self-supervised image recognition using deep neural networks. Knowl. Based Syst. 2021, 224, 107090.
12. He, Y.; Carass, A.; Zuo, L.; Dewey, B.E.; Prince, J.L. Autoencoder based self-supervised test-time adaptation for medical image analysis. Med. Image Anal. 2021, 72, 102136.
13. Huang, S.-C.; Pareek, A.; Jensen, M.; Lungren, M.P.; Yeung, S.; Chaudhari, A.S. Self-supervised learning for medical image classification: A systematic review and implementation guidelines. NPJ Digit. Med. 2023, 6, 74.
14. Baek, S.; Yoon, G.; Song, J.; Yoon, S.M. Self-supervised deep geometric subspace clustering network. Inf. Sci. 2022, 610, r235–r245.

15. Zhang, X.; Mu, J.; Zhang, X.; Liu, H.; Zong, L.; Li, Y. Deep anomaly detection with self-supervised learning and adversarial training. Pattern Recognit. 2022, 121, 108234.
16. Ciga, O.; Xu, T.; Martel, A.L. Self supervised contrastive learning for digital histopathology. Mach. Learn. Appl. 2022, 7, 100198.
17. Liu, Y.; Zhou, S.; Wu, H.; Han, W.; Li, C.; Chen, H. Joint optimization of autoencoder and Self-Supervised Classifier: Anomaly detection of strawberries using hyperspectral imaging. Comput. Electron. Agric. 2022, 198, 107007.
18. Hou, Z.; Liu, X.; Cen, Y.; Dong, Y.; Yang, H.; Wang, C.; Tang, J. GraphMAE: Self-Supervised Masked Graph Autoencoders. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 14–18 August 2022; ACM: New York, NY, USA, 2022; pp. 594–604.
19. Li, Y.; Lao, Q.; Kang, Q.; Jiang, Z.; Du, S.; Zhang, S.; Li, K. Self-supervised anomaly detection, staging and segmentation for retinal images. Med. Image Anal. 2023, 87, 102805.
20. Wang, T.; Wu, J.; Zhang, Z.; Zhou, W.; Chen, G.; Liu, S. Multi-scale graph attention subspace clustering network. Neurocomputing 2021, 459, 302–314.
21. Li, J.; Ren, W.; Han, M. Variational auto-encoders based on the shift correction for imputation of specific missing in multivariate time series. Measurement 2021, 186, 110055.
22. Sun, C. HAT-GAE: Self-Supervised Graph Auto-encoders with Hierarchical Adaptive Masking and Trainable Corruption. arXiv 2023, arXiv:2301.12063.
23. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked Autoencoders Are Scalable Vision Learners. In IEEE/CVF Conference on Computer Vision and Pattern Recognition; Ernest N. Morial Convention Center: New Orleans, LA, USA; IEEE: Amsterdam, The Netherlands, 2022; pp. 16000–16009.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020, arXiv:2010.11929.
25. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. Commun. ACM 2020, 63, 139–144.
26. Fekri, M.N.; Ghosh, A.M.; Grolinger, K. Generating Energy Data for Machine Learning with Recurrent Generative Adversarial Networks. Energies 2019, 13, 130.
27. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv 2015, arXiv:1511.06434.
28. Berg, P.; Pham, M.-T.; Courty, N. Self-Supervised Learning for Scene Classification in Remote Sensing: Current State of the Art and Perspectives. Remote Sens. 2022, 14, 3995.
29. Doersch, C.; Gupta, A.; Efros, A.A. Unsupervised Visual Representation Learning by Context Prediction. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; IEEE: Amsterdam, The Netherlands, 2015; pp. 1422–1430.
30. Zhang, R.; Isola, P.; Efros, A.A. Colorful Image Colorization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 649–666.
31. Gidaris, S.; Singh, P.; Komodakis, N. Unsupervised Representation Learning by Predicting Image Rotations. In Proceedings of the Sixth International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018; ICLR 2018. Cornel University: Ithaca, NY, USA, 2018.
32. Noroozi, M.; Favaro, P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In European Conference on Computer Vision; Springer Nature: Cham, Switzerland, 2016; pp. 69–84.
33. Dosovitskiy, A.; Springenberg, J.T.; Riedmiller, M.; Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. Adv. Neural Inf. Process Syst. 2014, 27, 1–9.
34. Lee, C.P.; Lim, K.M.; Song, Y.X.; Alqahtani, A. Plant-CNN-ViT: Plant Classification with Ensemble of Convolutional Neural Networks and Vision Transformer. Plants 2023, 12, 2642.
35. Dong, X.; Shen, J. Triplet Loss in Siamese Network for Object Tracking. In European Conference on Computer Vision (ECCV); Springer: Munich, Germany, 2018; pp. 459–474.
36. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. In 37th International Conference on Machine Learning, PMLR 119; PLMR: Vienna, Austria, 2020; pp. 1597–1607.
37. Helber, P.; Bischke, B.; Dengel, A.; Borth, D. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 2019, 12, 2217–2226.

38. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition; IEEE: Seattle, WA, USA, 2020; pp. 9729–9738.

39. Li, X.; Zhou, Y.; Zhang, Y.; Zhang, A.; Wang, W.; Jiang, N.; Wu, H.; Wang, W. Dense Semantic Contrast for Self-Supervised Visual Representation Learning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; ACM: New York, NY, USA, 2021; pp. 1368–1376.

40. Fini, E.; Astolfi, P.; Alahari, K.; Alameda-Pineda, X.; Mairal, J.; Nabi, M.; Ricci, E. Semi-Supervised Learning Made Simple with Self-Supervised Clustering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; IEEE: New York, NY, USA, 2023; pp. 3187–3197.

41. Khan, A.; AlBarri, S.; Manzoor, M.A. Contrastive Self-Supervised Learning: A Survey on Different Architectures. In Proceedings of the 2022 2nd International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 30–31 March 2022; IEEE: New York, NY, USA, 2022; pp. 1–6.

42. Liu, Y.; Zhu, L.; Yamada, M.; Yang, Y. Semantic Correspondence as an Optimal Transport Problem. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New York, NY, USA, 2020; pp. 4462–4471.

43. Shvetsova, N.; Petersen, F.; Kukleva, A.; Schiele, B.; Kuehne, H. Learning by Sorting: Self-supervised Learning with Group Ordering Constraints. arXiv 2023, arXiv:2301.02009.

44. Li, H.; Liu, J.; Cui, L.; Huang, H.; Tai, X.-C. Volume preserving image segmentation with entropy regularized optimal transport and its applications in deep learning. J. Vis. Commun. Image Represent. 2020, 71, 102845.

45. Li, R.; Lin, G.; Xie, L. Self-Point-Flow: Self-Supervised Scene Flow Estimation from Point Clouds with Optimal Transport and Random Walk. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 19–25 June 2021; IEEE: New York, NY, USA, 2021; pp. 15577–15586.

46. Scetbon, M.; Cuturi, M. Low-rank optimal transport: Approximation, statistics and debiasing. Adv. Neural Inf. Process Syst. 2022, 35, 6802–6814.

47. Zhang, C.; Zhang, C.; Zhang, K.; Zhang, C.; Niu, A.; Feng, J.; Yoo, C.D.; Kweon, I.S. Decoupled Adversarial Contrastive Learning for Self-supervised Adversarial Robustness. In European Conference on Computer Vision; Springer Nature: Cham, Switzerland, 2022; pp. 725–742.

48. Liu, W.; Li, Z.; Zhang, H.; Chang, S.; Wang, H.; He, J.; Huang, Q. Dense lead contrast for self-supervised representation learning of multilead electrocardiograms. Inf. Sci. 2023, 634, 189–205.

49. Wang, X.; Zhang, R.; Shen, C.; Kong, T. DenseCL: A simple framework for self-supervised dense visual pre-training. Vis. Inform. 2023, 7, 30–40.

50. Liu, X.; Sinha, A.; Unberath, M.; Ishii, M.; Hager, G.D.; Taylor, R.H.; Reiter, A. Self-Supervised Learning for Dense Depth Estimation in Monocular Endoscopy. In OR 2.0 Context-Aware. Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin. Image Analysis: First International Workshop, OR 2.0 2018, 5th International Workshop, CARE 2018, 7th International Workshop, CLIP 2018, Third International Workshop, ISIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16 and 20, 2018, Proceedings 5; Springer International Publishing: Cham, Switzerland, 2018; pp. 128–138.

51. Kar, S.; Nagasubramanian, K.; Elango, D.; Nair, A.; Mueller, D.S.; O'Neal, M.E.; Singh, A.K.; Sarkar, S.; Ganapathysubramanian, B.; Singh, A. Self-Supervised Learning Improves Agricultural Pest Classification. In Proceedings of the AI for Agriculture and Food Systems, Vancouver, BC, Canada, 28 February 2021.

52. Niizumi, D.; Takeuchi, D.; Ohishi, Y.; Harada, N.; Kashino, K. BYOL for Audio: Self-Supervised Learning for General-Purpose Audio Representation. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; IEEE: New York, NY, USA, 2021; pp. 1–8.

53. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Pires, B.A.; Guo, Z.D.; Azar, M.G.; et al. Bootstrap your own latent-a new approach to self-supervised learning. Adv. Neural Inf. Process. Syst. 2020, 33, 21271–21284.

54. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. arXiv 2015, arXiv:1503.02531.

55. Wang, J.; Zhu, T.; Gan, J.; Chen, L.L.; Ning, H.; Wan, Y. Sensor Data Augmentation by Resampling in Contrastive Learning for Human Activity Recognition. IEEE Sens. J. 2022, 22, 22994–23008.

56. Wu, J.; Gong, X.; Zhang, Z. Self-Supervised Implicit Attention: Guided Attention by The Model Itself. arXiv 2022, arXiv:2206.07434.

57. Haresamudram, H.; Essa, I.; Plötz, T. Investigating Enhancements to Contrastive Predictive Coding for Human Activity Recognition. In Proceedings of the 2023 IEEE International Conference on Pervasive Computing and Communications (PerCom), Atlanta, GA, USA, 13–17 March 2023; IEEE: New York, NY, USA, 2023; pp. 232–241.

58. Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: New York, NY, USA, 2021; pp. 9630–9640.

59. Balestriero, R.; Ibrahim, M.; Sobal, V.; Morcos, A.; Shekhar, S.; Goldstein, T.; Bordes, F.; Bardes, A.; Mialon, G.; Tian, Y.; et al. A cookbook of self-supervised learning. arXiv 2023, arXiv:2304.12210.

60. Chen, Y.; Liu, Y.; Jiang, D.; Zhang, X.; Dai, W.; Xiong, H.; Tian, Q. SdAE: Self-distillated Masked Autoencoder. In European Conference on Computer Vision; Springer Nature: Cham, Switzerland, 2022; pp. 108–124.

61. Alfaro-Contreras, M.; Ríos-Vila, A.; Valero-Mas, J.J.; Calvo-Zaragoza, J. Few-shot symbol classification via self-supervised learning and nearest neighbor. Pattern Recognit. Lett. 2023, 167, 1–8.

62. Lee, D.; Aune, E. VIbCReg: Variance-invariance-better-covariance regularization for self-supervised learning on time series. arXiv 2021, arXiv:2109.00783.

63. Mialon, G.; Balestriero, R.; LeCun, Y. Variance covariance regularization enforces pairwise independence in self-supervised representations. arXiv 2022, arXiv:2209.14905.

64. Bardes, A.; Ponce, J.; LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv 2021, arXiv:2105.04906.

65. Chen, S.; Guo, W. Auto-Encoders in Deep Learning—A Review with New Perspectives. Mathematics 2023, 11, 1777.

66. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 2015, 115, 211–252.

67. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Open Long-Tailed Recognition in A Dynamic World. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 46, 1836–1851.

68. Liu, Y.; Jin, M.; Pan, S.; Zhou, C.; Zheng, Y.; Xia, F.; Yu, P. Graph Self-Supervised Learning: A Survey. IEEE Trans. Knowl. Data Eng. 2022, 35, 5879–5900.

69. Krishnan, R.; Rajpurkar, P.; Topol, E.J. Self-supervised learning in medicine and healthcare. Nat. Biomed. Eng. 2022, 6, 1346–1352.

70. Azizi, S.; Mustafa, B.; Ryan, F.; Beaver, Z.; Freyberg, J.; Deaton, J.; Loh, A.; Karthikesalingam, A.; Kornblith, S.; Chen, T.; et al. Big Self-Supervised Models Advance Medical Image Classification. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV); IEEE: New York, NY, USA, 2021; pp. 3458–3468.

71. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive learning of medical visual representations from paired images and text. In Machine Learning for Healthcare Conference; PMLR: Vienna, Austria, 2022; pp. 2–25.

72. Bozorgtabar, B.; Mahapatra, D.; Vray, G.; Thiran, J.-P. SALAD: Self-supervised Aggregation Learning for Anomaly Detection on X-rays. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, 4–8 October 2020, Proceedings, Part I 23; Springer International Publishing: Cham, Switzerland, 2020; pp. 468–478.

73. Tian, Y.; Pang, G.; Liu, F.; Chen, Y.; Shin, S.H.; Verjans, J.W.; Singh, R.; Carneiro, G. Constrained Contrastive Distribution Learning for Unsupervised Anomaly Detection and Localisation in Medical Images. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part V 24; Springer International Publishing: Cham, Switzerland, 2021; pp. 128–140.

74. Ouyang, J.; Zhao, Q.; Adeli, E.; Sullivan, E.V.; Pfefferbaum, A.; Zaharchuk, G.; Pohl, K.M. Self-supervised Longitudinal Neighbourhood Embedding. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part II 24; Springer International Publishing: Cham, Switzerland, 2021; pp. 80–89.

75. Liu, F.; Tian, Y.; Cordeiro, F.R.; Belagiannis, V.; Reid, I.; Carneiro, G. Self-supervised Mean Teacher for Semi-supervised Chest X-ray Classification. In International Workshop on Machine Learning in Medical Imaging; Springer International Publishing: Cham, Switzerland, 2021; pp. 426–436.

76. Li, H.; Xue, F.F.; Chaitanya, K.; Luo, S.; Ezhov, I.; Wiestler, B.; Zhang, J.; Menze, B. Imbalance-Aware Self-supervised Learning for 3D Radiomic Representations. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part II 24; Springer International Publishing: Cham, Switzerland, 2021; pp. 36–46.

77. Manna, S.; Bhattacharya, S.; Pal, U. Interpretive Self-Supervised pre-Training. In Twelfth Indian Conference on Computer Vision, Graphics and Image Processing; ACM: New York, NY, USA, 2021; pp. 1–9.

78. Zhao, Z.; Yang, G. Unsupervised Contrastive Learning of Radiomics and Deep Features for Label-Efficient Tumor Classification. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021, Proceedings, Part II 24; Springer International Publishing: Cham, Switzerland, 2021; pp. 252–261.

79. Esrafilian-Najafabadi, M.; Haghighat, F. Towards self-learning control of HVAC systems with the consideration of dynamic occupancy patterns: Application of model-free deep reinforcement learning. Build. Environ. 2022, 226, 109747.

80. Long, J.; Chen, Y.; Yang, Z.; Huang, Y.; Li, C. A novel self-training semi-supervised deep learning approach for machinery fault diagnosis. Int. J. Prod. Res. 2023, 61, 8238–8251.

81. Yang, Z.; Huang, Y.; Nazeer, F.; Zi, Y.; Valentino, G.; Li, C.; Long, J.; Huang, H. A novel fault detection method for rotating machinery based on self-supervised contrastive representations. Comput. Ind. 2023, 147, 103878.

82. Wei, M.; Liu, Y.; Zhang, T.; Wang, Z.; Zhu, J. Fault Diagnosis of Rotating Machinery Based on Improved Self-Supervised Learning Method and Very Few Labeled Samples. Sensors 2021, 22, 192.

83. Lei, Y.; Karimi, H.R.; Chen, X. A novel self-supervised deep LSTM network for industrial temperature prediction in aluminum processes application. Neurocomputing 2022, 502, 177–185.

84. Xu, R.; Hao, R.; Huang, B. Efficient surface defect detection using self-supervised learning strategy and segmentation network. Adv. Eng. Inform. 2022, 52, 101566.

85. Bharti, V.; Kumar, A.; Purohit, V.; Singh, R.; Singh, A.K.; Singh, S.K. A Label Efficient Semi Self-Supervised Learning Framework for IoT Devices in Industrial Process. IEEE Trans. Ind. Inform. 2023, 20, 2253–2262.

86. Hannan, M.A.; How, D.N.T.; Lipu, M.S.H.; Mansor, M.; Ker, P.J.; Dong, Z.Y.; Sahari, K.S.M.; Tiong, S.K.; Muttaqi, K.M.; Mahlia, T.M.I.; et al. Deep learning approach towards accurate state of charge estimation for lithium-ion batteries using self-supervised transformer model. Sci. Rep. 2021, 11, 19541.

87. Li, C.L.; Sohn, K.; Yoon, J.; Pfister, T. CutPaste: Self-Supervised Learning for Anomaly Detection and Localization. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, North America; IEEE: Washington, DC, USA, 2021; pp. 9664–9674.

88. Yang, C.; Wu, Z.; Zhou, B.; Lin, S. Instance Localization for Self-Supervised Detection Pretraining. In CVF Conference on Computer Vision and Pattern Recognition, North America; IEEE: Washington, DC, USA, 2021; pp. 3987–3996.

89. Schlüter, H.M.; Tan, J.; Hou, B.; Kainz, B. Natural Synthetic Anomalies for Self-supervised Anomaly Detection and Localization. In European Conference on Computer Vision; Springer Nature: Cham, Switzerland, 2022; pp. 474–489.

90. Taher, M.R.H.; Haghighi, F.; Gotway, M.B.; Liang, J. CAiD: Context-Aware Instance Discrimination for Self-Supervised Learning in Medical Imaging. In International Conference on Medical Imaging with Deep Learning; MIDL Foundation: Zürich, Switzerland, 2022; pp. 535–551.

91. Gidaris, S.; Bursuc, A.; Puy, G.; Komodakis, N.; Cord, M.; Pérez, P. Obow: Online Bag-of-Visual-Words Generation for Self-Supervised Learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, North America; IEEE: Washington, DC, USA, 2021; pp. 6830–6840.

92. Baevski, A.; Babu, A.; Hsu, W.N.; Auli, M. Efficient Self-Supervised Learning with Contextualized Target Representations for Vision, Speech and Language. In Proceedings of the 40th International Conference on Machine Learning, PMLR 2023, Honolulu, HI, USA, 23–29 July 2023; PMLR: Vienna, Austria, 2023; pp. 1416–1429.

93. Park, D.; Ahn, C.W. Self-Supervised Contextual Data Augmentation for Natural Language Processing. Symmetry 2019, 11, 1393.

94. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44, 7112–7127.

95. Lin, J.H.; Lin, Y.Y.; Chien, C.M.; Lee, H.Y. S2VC: A Framework for Any-to-Any Voice Conversion with Self-Supervised Pretrained Representations. arXiv 2021, arXiv:2104.02901.

96. Chung, Y.-A.; Hsu, W.-N.; Tang, H.; Glass, J. An Unsupervised Autoregressive Model for Speech Representation Learning. In Interspeech 2019; ISCA: Singapore, 2019; pp. 146–150.

97. Pan, T.; Chen, J.; Zhou, Z.; Wang, C.; He, S. A Novel Deep Learning Network via Multiscale Inner Product with Locally Connected Feature Extraction for Intelligent Fault Detection. IEEE Trans. Ind. Inform. 2019, 15, 5119–5128.

98. Chen, J.; Wang, C.; Wang, B.; Zhou, Z. A visualized classification method via t-distributed stochastic neighbor embedding and various diagnostic parameters for planetary gearbox fault identification from raw mechanical data. Sens. Actuators A Phys. 2018, 284, 52–65.

99. Zhang, T.; Chen, J.; He, S.; Zhou, Z. Prior Knowledge-Augmented Self-Supervised Feature Learning for Few-Shot Intelligent Fault Diagnosis of Machines. IEEE Trans. Ind. Electron. 2022, 69, 10573–10584.