Artificial Intelligence Review (2024) 57:226
https://doi.org/10.1007/s10462-024-10859-3

Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets

Ishfaq Hussain Rather[1] · Sushil Kumar[1] · Amir H. Gandomi[2,3]

Accepted: 13 July 2024 / Published online: 2 August 2024

Abstract

Justifiably, while big data is the primary interest of research and public discourse, it is essential to acknowledge that small data remains prevalent. The same technological and societal forces that generate big datasets also produce a more significant number of small datasets. Contrary to the notion that more data is inherently superior, real-world constraints such as budget limitations and increased analytical complexity present critical challenges. Quality versus quantity trade-offs necessitate strategic decision-making, where small data often leads to quicker, more accurate, and cost-effective insights. Concentrating AI research, particularly in deep learning (DL), on big datasets exacerbates AI inequality, as tech giants such as Meta, Amazon, Apple, Netflix and Google (MAANG) can easily lead AI research due to their access to vast datasets, creating a barrier for small and mid-sized enterprises that lack similar access. This article addresses this imbalance by exploring DL techniques optimized for small datasets, offering a comprehensive review of historic and state-of-the-art DL models developed specifically for small datasets. This study aims to highlight the feasibility and benefits of these approaches, promoting a more inclusive and equitable AI landscape. Through a PRISMA-based literature search, 175+ relevant articles are identified and subsequently analysed based on various attributes, such as publisher, country, utilization of small dataset technique, dataset size, and performance. This article also delves into current DL models and highlights open research problems, offering recommendations for future investigations. Additionally, the article highlights the importance of developing DL models that effectively utilize small datasets, particularly in domains where data acquisition is difficult and expensive.

Keywords Deep learning · Small datasets · Data augmentation · Transfer learning · Few-shot learning · Generative adversarial networks

1 Introduction

Although artificial intelligence (AI) is now pervasive and widely used, there is a troubling reality that a handful of tech giants tightly control the technology (Verdegem 2022). One group of tech giants that eliminate numerous friction points in our lives as consumers include the route optimization algorithms used by Google Maps, personalized shopping

Extended author information available on the last page of the article

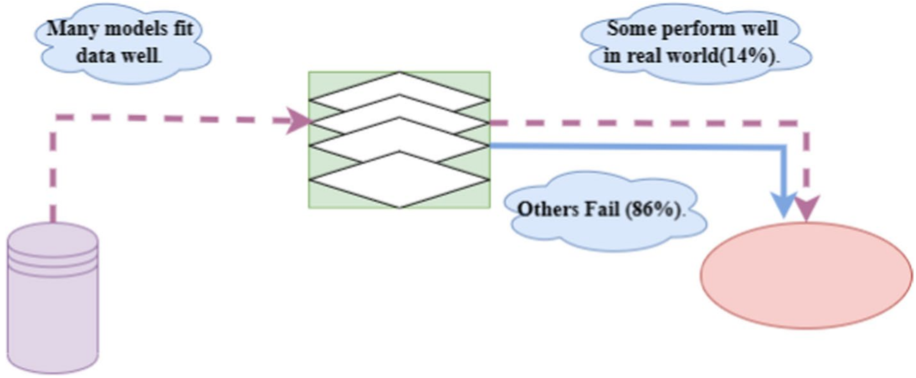
Many models fit data well.
Some perform well in real world(14%).
Others Fail (86%).

Fig. 1 Illustration of real-world ML pipeline

suggestions on Amazon, and the natural language comprehension employed by Apple's Siri. Although we engage with AI regularly, less than 15% implement AI in operational use within the corporate sector, and approximately an equivalent proportion are confident in their possession of the necessary technological framework to sustain AI ventures (c.f. Fig. 1). The disparity between Big Tech. and the rest widens further when the focus is narrowed to Machine Learning (ML).[1] This is because, firstly, big data is necessary for training an ML model, which ultimately allows the model to produce accurate predictions. Additionally, a corporation aiming to automate an internal process could only have 100 relevant samples, unlike Google, which has the advantage of 130 trillion sites to improve its algorithms. Second, engaging a team of ML experts to automate internal operations is impossible for average or small businesses because these experts are a tiny talent pool and can work almost anywhere and at any price. As a result, many companies use off-the-shelf tools from outside vendors rather than developing their own AI tools. This helps small and mid-sized businesses overcome the lack of resources and expertise. Thus, the absolute monopoly of Big Tech companies on AI is primarily narrowed down to one key reason: the shortage of data. Big Tech. Corporations have easy access to sufficient data, and small and mid-size corporations must strive for it, bringing AI inequality.

Considering that data is the fuel for ML models, it is possible to end Big Tech's artificial intelligence (AI) monopoly and promote equality by building AI products with less data. The last several years have seen tremendous developments in deep learning (DL) (Lecun et al. 2015; Menghani 2023; Marcus et al 2018), a subfield of AI and ML that has encouraged enterprises across sectors to incorporate DL solutions into their AI strategy. DL has enabled many sophisticated new AI applications, ranging from chatbots in customer service to image and object identification in retail, among others. The remarkable success of DL algorithms with complicated tasks has made them particularly desirable to many businesses in recent years. However, we live in a world where data is never endless. DL systems often need to generalize beyond the data they are trained on, such as when encountering a new word pronunciation or a different image. Due to the limited nature of data, the use of formal arguments to ensure high-quality performance has constraints.

[1] The Small Data Revolution: AI Isn't Just For The Big Guys Anymore (forbes.com).

The significant contributions of this review paper are subsequently described. This article outlines some significant challenges associated with DL models.

- A PRISMA model search study is conducted to identify relevant studies, considering 175+ research articles.
- The study thoroughly reviewed historical and contemporary DL techniques explicitly designed for small datasets.
- The paper investigated an alternative approach that employs DL techniques specifically for small datasets, deviating from their conventional utilization with large datasets. We concentrated on small, structured datasets and assessed the performance of several DL algorithms explicitly designed for such datasets. Our goal is to evaluate the efficacy and possible benefits of using these specialized approaches in the context of smaller datasets.
- A comparative study of different small dataset techniques using different metrics is performed.
- The article discusses several unresolved research issues and provides recommendations for more studies.

The remaining sections of the paper are organised as follows. Section 2 discusses the search methodology and statistical distribution analysis of publications utilising DL models for small datasets in detail. Section 3 presents the limitations of DL models. Section 4 describes the motivation for this article. Section 5 discusses the detailed techniques for developing DL models for small datasets. Section 6 overviews some open research issues and provides recommendations for more studies. Finally, Sect. 7 summarises the paper's conclusion.

2 Search methodology and statistical distribution analysis of deep learning models utilising small datasets

2.1 PRISMA model design

The Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) guideline offers instructions for writing systematic reviews, incorporating advancements in techniques for identifying, choosing, evaluating, and producing studies (Page et al. 2021). A PRISMA model search study is conducted to identify relevant studies, considering 175+ research articles. There is a considerable amount of existing research focused on DL models that leverage big datasets [e.g., (Marcus et al 2018; Chen and Lin 2014; Ahmed et al. 2023)]. However, much less research has gone into reviewing the usage of small datasets for training DL models [e.g., (Gheisari et al. 2017; Ahmed et al. 2023; Bansal et al. 2022)]. This study performs a systematic search that utilises IEEE, PubMed, Google Scholar, Science Direct and Arxiv. The keywords used for our literature search are "Small datasets" OR "Short datasets" OR "Limited datasets" OR "Low datasets" OR "Few samples" AND "Machine Learning" OR "Deep Learning" OR "Computer Vision". We also performed reference tracking, i.e. the papers cited in the reviews. We carefully examined the papers referenced in the recent review articles (Zhang et al. 2019; ul Sabha et al. 2024). Additionally, we thoroughly investigated the recent articles that cited these review papers

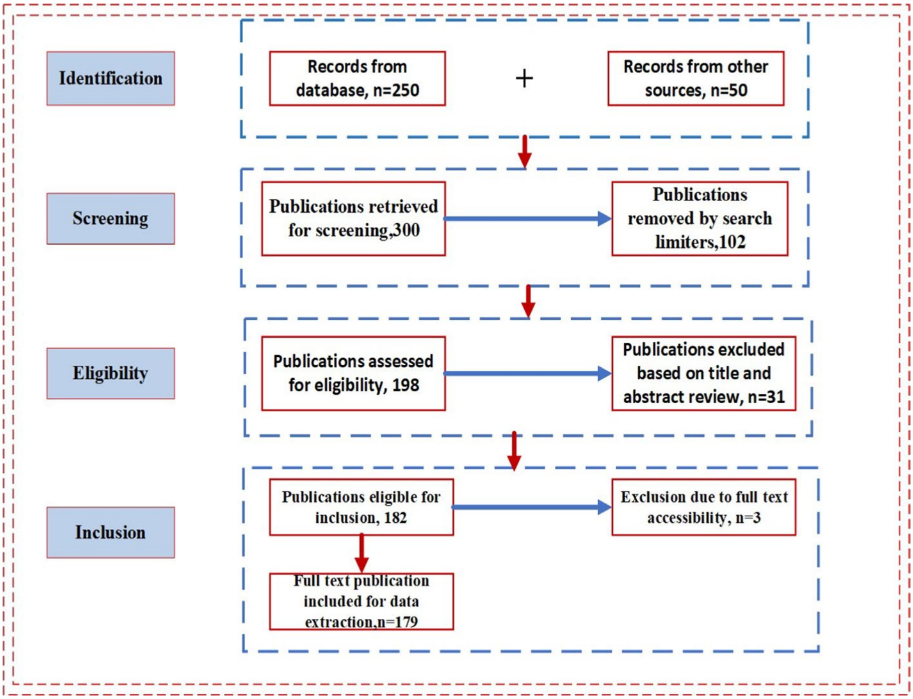
Identification
Records from database, n=250
+
Records from other sources, n=50
Screening
Publications retrieved for screening,300
Publications removed by search limiters,102
Eligibility
Publications assessed for eligibility, 198
Publications excluded based on title and abstract review, n=31
Inclusion
Publications eligible for inclusion, 182
Exclusion due to full text accessibility, n=3
Full text publication included for data extraction,n=179

Fig. 2 PRISMA analysis for selection of studies

on their Google Scholar pages. It is important to note that our search was limited to articles written in English and published in peer-reviewed journals or conference proceedings.

In the literature review process, 300+ papers were initially identified as relevant to DL models for small datasets. Of these, 200 highly related articles were selected by excluding the papers that: (i) do not use AI (including its branches such as Computer Vision (CV), Machine Learning (ML), and Deep Learning (DL)); (ii) have insufficient data, and (iii) are not relevant. Among the selected articles, 42 papers had small, limited, short, low, or few keywords in their title. Figure 2 shows the PRISMA model, illustrating the references related to ML, DL and CV techniques relevant to this review paper.

2.2 Novel contributions compared to existing reviews

Survey articles that focus on applications of DL techniques to small datasets are very scarce**.** This gap in the literature is mainly due to a prevailing misconception that DL techniques cannot be applied to small dataset problems. This notion has led many researchers to disregard small datasets when considering DL applications. Thus, most existing research prioritises large datasets traditionally considered more suitable for DL models. However, several innovative techniques make DL models compatible with small datasets. Our study systematically categorises these small dataset techniques and highlights the diversity of methods available. The existing survey articles, such as Gheisari et al. (2017), Ahmed et al. (2023), Bansal et al. (2022), Zhang et al. 2019, ul Sabha et al. (2024), address only a few small dataset techniques, as shown in Table 1. The

Table 1 Comparison with the existing review articles on deep learning with small datasets

Review study	Small dataset techniques discussed					
	Data augmentation	Transfer learning	GAN(s)	FSL	Loss based	Model architecture
Gheisari et al. (2017)	✓	✓	×	✓	×	×
Ahmed et al. (2023)	✓	✓	✓	✓	×	×
Gheisari et al. (2017)	✓	✓	✓	×	×	×
Ahmed et al. (2023)	✓	✓	✓	×	×	×
Bansal et al. (2022)	×	✓	✓	×	×	✓
Our study	✓	✓	✓	✓	✓	✓

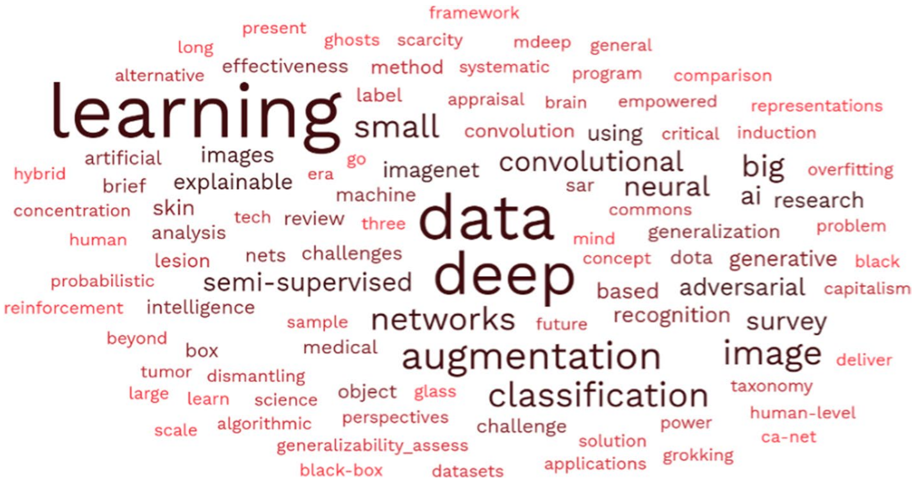

Fig. 3 Word-cloud visualization of the titles of all 175+ articles

comparison illustrates the broader scope and greater depth of our review. By thoroughly examining the methods and categorizing them comprehensively, our study fills a critical gap in the literature. We aim to challenge the prevailing misconceptions and encourage more researchers to explore DL applications with small datasets. Our survey, therefore, stands out by offering a more extensive and detailed overview of the strategies that make DL viable in data-constrained environments.

2.3 Overview of the reviewed studies

In this study, 175+ high-quality papers were reviewed. Figure 3 shows the word-cloud visualisation of the titles of these articles, which offers a thorough and insightful summary of the major areas covered in the reviewed articles, generally and specifically. It can be seen that the most frequently used words are "Data", "Deep", "Learning", "Augmentation", "networks", "classification", "Small", and "Semi-supervised".

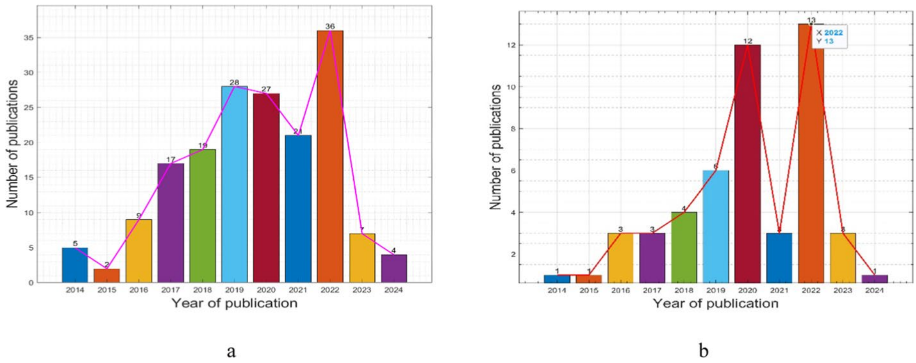

Fig. 4 **a** Number of publications selected per year over the last 10 years. **b** Number of publications per year with keywords, such as “Small” OR “Limited” OR “Shot” OR “few” OR “small” AND “data’’ OR “dataset” OR “sample” in the paper titles

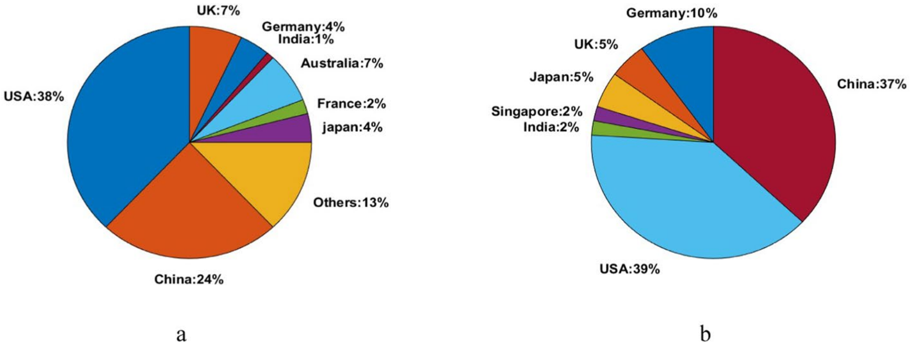

Fig. 5 **a** Country-wise percentage of publications selected in this review article. **b** Country-wise percentage of publications with titles containing keywords like “Small”, “Shot”, “Few samples”, “limited”, and “low” in the article titles

2.4 Statistical distributions

As DL and ML continue to advance globally, knowing the publications and their countries of origin is crucial. Our study was conducted to shed light on this subject based on the number of publications per year for the last 10 years, as depicted in Fig. 4a. While an upward trend in publications occurred from 2015 to 2019, there was a decline between 2020 and 2021. The maximum number of articles was recorded in 2022. Figure 4b shows the yearly number of publications that have “small”, “shot”, “few”, “low”, or “limited” data keywords in the article title. The pie chart in Fig. 6a illustrates the percentage according to the publishers of the articles, indicating that IEEE contributed the highest percentage of papers (43%), followed by arXiv (29%), Springer (9%), and lastly Elsevier (8%). The Association for Computing Machinery (ACM) and Multidisciplinary Digital Publishing Institute (MDPI) each contributed 6% and 3% of publications in this review article, respectively. In addition, 2% of the publications came from Nature, and the remaining 13% came from other sources.

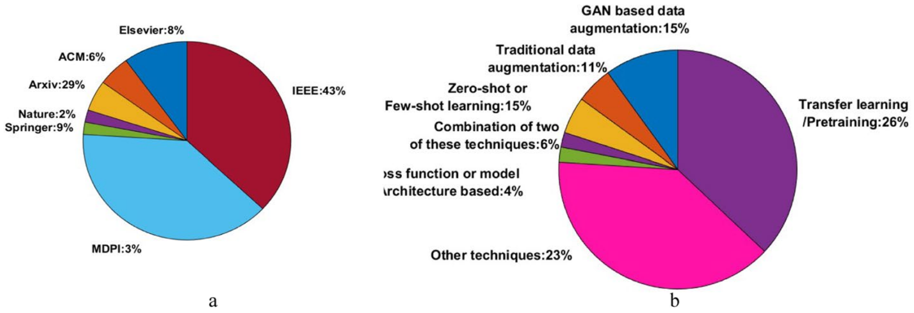

Fig. 6 **a** Proportion of publishers involved in the article review. **b** Proportion of articles employing small dataset techniques

The pie chart in Fig. 5a shows the percentage of publications belonging to a particular country. It can be seen that the USA accounts for 38% of the total publications included in this study, followed by China (24%), UK and Australia (7% each), and Germany (4%). Figure 5b shows the yearly number of publications that have “small”, “shot”, “few”, “low”, or “limited” data keywords in the article title. The data reveal that the US and China are the primary contributors, with 39% and 37% of the total publications, respectively. This suggests that both countries are making significant strides in DL that leverages small datasets. Germany contributed 10%, the UK and Japan contributed 5% each, followed by Singapore and India with 2% each.

Researchers have investigated several techniques to address DL models for small datasets (DLS), such as Data Augmentation (DA), transfer learning (TL), generative adversarial networks (GANs), few-shot learning (FSL), loss function, model architecture, and regularization-based methods. Moreover, these methods are integrated to improve DL performance on limited datasets. Figure 6b shows the contribution percentage in each of these methods. The traditional DA-based techniques contribute 11% of the DLS, followed by TL (26%), GANs-based techniques (15%), and transfer learning methods (26%), and loss function-based methods (2%). Some papers combined these techniques, contributing to 6% of the publications. The articles that used small dataset techniques other than the ones mentioned contributed to 23% of the total articles.

3 Limitations of deep learning

Despite significant achievements in DL models, two aspects of human conceptual understanding have escaped computer systems. First, although humans can acquire new concepts from a few samples with high generalisation ability, traditional DL models require thousands of instances to perform with comparable accuracy. Second, even for basic notions, humans learn richer representations than machines and use them for broader purposes (Lake et al. 2022). Some of the limitations of DL models are mentioned below.

3.1 Deep learning has been data-hungry thus far

DL models often demand significant data to attain optimal results. This is primarily due to the large parameter count for optimisation throughout the training phase. The availability of big datasets improves their generalisation capacity, enabling them to predict and respond to new inputs accurately. The test data are from the same distribution for the model to interpolate new responses between existing ones (Kim et al. 2023; Power et al. 2022). In Krizhevsky et al. (2017), a convolutional neural network (CNN) with nine layers, 60 million parameters, and 650,000 nodes was trained on nearly a million different samples with a thousand classes. On the ImageNet dataset, this sort of brute-force method works well. The quantity of data required for high-quality DL depends on the problem's complexity and network size. For example, GPT-3 has 175 billion parameters and is one of the largest networks ever trained (Chen and Lin 2014). When a model has many parameters that must be optimised exclusively by feeding it training data, we must ensure that it has many training samples. A model that needs extensive data for training clashes with the nature of human intelligence (Świechowski 2022); we can state that no small or medium-sized business can have this much training data.

AI draws inspiration from human intelligence, and most approaches for determining whether or not we have achieved it include analogies to humans. Humans learn to be efficient at an almost infinite variety of tasks. On the other hand, for instance, children do not need to see many different cats to recognise one. While the learning process in children is not entirely comprehended, it is most plausible that the human brain constructs an abstract representation of a concept internally very quickly (Spicer and Sanborn 2019). Furthermore, utilising sizeable DL networks with extensive data results in high computing costs and, as a result, extended training times (OpenAI et al. 2019). It would be absolutely inefficient to train a DL model for an extended period for each job that AI is confronted with. We require quicker methods of training or creating AI models (OpenAI et al. 2019). DL, thus far, has not been regarded as an optimal solution for small dataset problems.

3.2 Deep learning models lack interpretability

The largest compliance barrier for AI is a lack of interpretability. DL models are black box models (Quinn et al. 2022; Rai 2020) whereby the "black box" problem is characterised as difficulty in interpreting and expressing the reasons behind a model's forecast outcome. How can we determine whether a model is adequately trained and tested if we cannot understand its output? The greater the interpretability of an ML model, the simpler it is to know why particular judgements or predictions were made (Quinn et al. 2022). If humans more easily understand a model's outcome than those of other models, the former model is said to be more interpretable or explainable (Huang et al. 2020; Gu et al. 2021; Miller 2017). Moreover, fairness and unbiasedness have lately emerged as crucial auxiliary criteria for model improvement. ML interpretability is a critical tool for testing key features of ML systems.

For ML models, interpretability is critical. When it comes to DL model predictive analysis, there is a trade-off: what prediction is made by the model, for example, the chance that the patient has a brain tumour, or why is the prediction made? In certain circumstances, we do not care why a judgement was made; instead, what truly matters is the prediction performance on test data (Heider and Simmel 1944). However, understanding the "why" provides valuable insights into the problem, the data, and potential model failures. Some approaches may not require explanations as they are used in a low-risk context (e.g., a movie recommender system) or the approaches that are previously widely investigated and evaluated (e.g. optical character recognition). The requirement for interpretability stems from an incompleteness in problem formalisation (Heider and Simmel 1944), which means that for particular issues or tasks, simply getting the prediction is insufficient (the “what”). Because a right forecast only partially answers our initial problem, the model must explain how it arrived at the prediction (the “why”). The more significant the impact of an ML model’s choice on a person’s life, the more vital it is for the machine to explain its behaviour. For example, if a DL model rejects a loan application, the applicants may be utterly surprised (Rayhan and Hashem 2023). They can only reconcile this discrepancy between anticipation and reality by providing some explanation. Incorporating DL-based models into our daily lives is critical for increasing societal acceptability. People attribute beliefs, desires, and intents to these models (Heider and Simmel 1944). Regarding compliance, the black box dilemma impedes AI’s march towards global integration. AI can never fulfil its full potential as long as it remains unexplained.

The majority of applications where interpretability is required involve small dataset problems, such as human illness diagnosis, disaster analysis, defence-related applications, etc. For these applications to be accepted by society, they must be able to explain the predicted behaviours.

3.3 Weakly supervised learning is a problem

Current supervised approaches have succeeded significantly, but getting sufficient supervision information, such as entirely ground-truth labels, is challenging. We require a precollected dataset with ground truth, such as ImageNet (Krizhevsky et al. 2017) or PASCAL VOC (Everingham et al. 2009) datasets, to train a model with many parameters and hundreds of layers (ul Sabha et al. 2024). However, due to the high cost of data labelling processes (such as data labelling of small-scale events from surveillance footage, including crowded large-scale situations) or a lack of expertise (such as annotating MRI scans with tumour and non-tumour or any other disease), it may be challenging to achieve such high-quality annotations for many samples in practice. Additionally, many datasets are gathered using crowdsourcing or search engines to cut the cost of human labour. However, they often have a lot of mediocre annotations (i.e., coarse or even inaccurate). This causes well-known, weakly supervised learning problems (Zhou 2018). One solution to this problem would be a DL model that learns effectively from a few samples. The alternative option is to create a model that can function with weak supervision. Some studies are already being done in this area (Settles 2009; Chen and Wang 2011).

3.4 Long-tail phenomena in big datasets degrade the deep learning model’s performance

Big datasets frequently suffer from the long-tail phenomenon, which occurs when a small number of classes have frequent data, but many more have rare data. Due to this data imbalance, a DL model can perform exceptionally in classes with more data but perform poorly in classes with fewer data (ul Sabha et al. 2024). The number of samples in various classes varies greatly across many datasets, such as credit card fraud detection datasets where the difference is significant. Considering that one fraudulent transaction can occur for every 10,000 valid ones, even if a model forecasts fraudulent transactions as legitimate ones, it would still be 99% accurate.

Rebalancing training data, which involves increasing the frequency of the samples from rare classes or reducing the number of samples from the top-numbered classes, is a straightforward improvement method (Shen et al. 2016). However, this approach is typically heuristic and ineffective. The latter tends to lose crucial feature information inside the classes with more samples, whilst the former tends to develop sample redundancy and confront the issue of over-fitting to the rare classes (Wang et al. 2017). Therefore, it is anticipated that the required small data learning approaches will be helpful in resolving the long-tail training problem by utilising more advantageous previous information from small-sample classes.

3.5 Hype vs reality

Humans have unreasonable short-term expectations of DL and artificial intelligence. Although DL research is advancing at a breakneck pace, the fact remains that relatively little of this knowledge has made its way into the products and processes that make up the world. However, most of the research findings are yet to be applied.[2] Hype may be problematic for emerging technologies because it increases the likelihood of their failure to deliver on projected promises or making exaggerated claims beyond reality. In 2011, producers of the popular game show "Jeopardy"[3] organised a unique competition by pitting IBM's AI supercomputer "Watson" against two of the show's most accomplished champions, Ken Jennings and Brad Rutter. Watson was victorious. Cancer was widely anticipated as Watson's next battle. IBM has partnered with numerous cancer centres since 2012 to apply Watson's abilities to cancer therapy. Watson's entry into cancer care and the interpretation of cancer genomes was well-publicised, with largely positive news coverage: "IBM to team up with UNC, Duke hospitals to fight cancer with big data"[4]; "The future of health care could be elementary with Watson".[5] But, three years after IBM introduced Watson to doctors worldwide to propose optimal cancer treatment, it was revealed that Watson was not living up to the enormous expectations proposed by IBM. The supercomputer is still grappling with the fundamental stage of understanding different types of cancer. Only a few hospitals have accepted the system, which falls significantly short of IBM's goal of dominating the market worth billions of dollars.

Yet, AI is still in its early stages and will take time to reach its full potential. When it does, it will have a long-term societal and economic impact that most people tend to underestimate. AI will change medicine, transportation, science, communication, and culture, becoming our portal to the outside world.

4 Motivation

The acquisition, processing, and privacy costs associated with data must be balanced against the benefits it offers. Technologies or societies that generate big data also produce vast numbers of small datasets. There are cases where small data is preferred over big data. High-quality small data can yield better inference than low-quality big data (Faraway and Augustin 2018). In 1936, for example, the prominent *Literary Digest* magazine polled its

[2] On the importance of democratizing Artificial Intelligence (keras.io).

[3] (31) Miles vs. Watson: The Complete Man Against Machine Showdown—YouTube.

[4] IBM (NYSE: IBM) to team up with UNC, Duke hospitals to fight cancer with Watson, the intelligence platform that appeared on Jeopardy—Triangle Business Journal (bizjournals.com).

[5] The future of health care could be elementary with Watson—PMC (nih.gov).

readers to predict the outcome of the US presidential election.[6] An overwhelming 2.4 million people participated in the poll, with 57% supporting Alfred London and 43% favouring Franklin Roosevelt. During this election, George Gallup's polling organisation (Gallup, Inc.) was getting started. Gallup anticipated a victory for Roosevelt with 56% using a sample size of just thousands. Roosevelt defeated Landon by a landslide margin of 62% to 38%. The small dataset of thousands outperformed the big dataset of millions.

Bias and variance may affect any estimation. In a time of severe economic distress, readers of the *Literary Digest* were more wealthy, had the discretionary income to spend on a magazine, and thus were generally more affluent than the general population. The large sample size did not mitigate the bias. Gallup's tiny sample would have been exposed to more significant variance, but this was a considerably less serious problem than bias. Statistical inference works effectively with small data but not on low-quality extensive data (Martin Lindstrom Company 2016). Developing DL models that are less data-intensive has several advantages. Accordingly, this research was motivated by the following aims.

Reduce capability differences between big and small tech companies: The greater dependence of AI applications on big datasets has created a concern about increasing the differences among organisations' capability to collect, store, and process relevant data. With this scenario, there is the possibility of widening the gap between the AI "haves", such as tech giants that can afford, and "have-nots" that cannot satisfy these demands. The approaches that utilise small datasets to apply AI can break the barrier for smaller organisations.

Minimise the incentive for accumulating much personal information: The use of AI will greatly reduce privacy (DIlmaghani et al. 2019). There are worries that major tech corporations continue to gather increasing amounts of consumer information related to individual identities to train their AI models. By decreasing the requirement to gather big real-world data for training ML models, certain small data approaches offer the ability to alleviate such worries to some extent.

Advance in areas where fewer data points are available: Problems where few data exist can be solved in an AI system, for example, by developing an ML model to detect a rare skin disease like Urticaria where there is no possibility of having a large amount of data. Small data techniques can give a rational strategy to cope with data scarcity.

Address challenges with dirty data: Small data techniques can help businesses with many unclean and unstructured data, making it unfit for analysis. For example, the US Department of Defense has a considerable volume of "dirty data" due to legacy systems, necessitating inefficient, labour-intensive data cleaning, labelling, and organising operations (Chahal et al. 2021). Small data techniques can reduce the quantity of data that needs to be pre-processed, saving labour and time.

There are trade-offs between quality and quantity in the real world of limited budgets. Small data sometimes outperform big data, enabling faster, more reliable, and cost-efficient conclusions. Small data are derived from experimental or intentionally collected data on a human scale, with an emphasis on causality and comprehension rather than prediction (Faraway and Augustin 2018).

[6] That Time the Literary Digest Poll Got the 1936 Election Wrong (proquest.com).

5 Techniques for solving small dataset problems in deep learning

In this section, data augmentation, generative adversarial networks, transfer learning, few-shot learning, and loss function-based techniques are extensively reviewed for solving the problem of small datasets, along with their advantages and drawbacks.

5.1 Data augmentation

The main problem with small data learning is overfitting (Power et al. 2022). In small datasets, the model is not exposed to every possible aspect of the data distribution, which ultimately has an issue of generalisation (Kim et al. 2023; Yousefzadeh 2022; Lemberger 2017; Jiang et al. 2019; Nagarajan 2021). Data augmentation (DA) is one approach to address the issue of overfitting in the DL models. DA aims to generate additional training data from current training samples by augmenting them with a range of random alterations that result in realistic-looking images. The model should never encounter the same image twice during training. As a result, the model is exposed to more aspects of the data and can better generalise.

The more data an ML system can access, the more successful it may be (Wang and Perez 2017). Even if the data is of poorer quality, the model can extract relevant information from the original dataset (Wang and Perez 2017). Text-to-speech and text-based models, for example, have enhanced dramatically due to Google's publication of a trillion-word corpus (Halevy et al. 2009). This is despite the fact that the data was gathered from unfiltered websites and contained several inaccuracies.

DA addresses overfitting at the cause of the problem, namely the training dataset. DA is employed because augmentation extracts additional information from the original data (Shorten and Khoshgoftaar 2019). These augmentations play a crucial role in expanding the training dataset's size through data warping or oversampling. Data warping provides extra training samples by applying transformations to the data space and alters existing pictures to retain their labels. This includes geometric and colour changes, adversarial training, random erasure, and neural style transfer. Synthetic oversampling generates more samples in feature space (Wong et al. 2016). The technique involves mixing images, GANs, and feature space enhancements (Goodfellow et al. 2014).

In the form of data warping, one of the earliest utilisations of DA can be found in LeNet-5 (LeCun et al. 1998) for the classification of handwritten digits. DA is also employed in the AlexNet CNN architecture (Krizhevsky et al. 2017), which revolutionised image classification using convolutional networks on the ImageNet dataset. This method increases the dataset's size by a factor of 2048 and assists in reducing overfitting when training a deep neural network. According to the study, augmentation lowered the error rate by more than 1%. The study Shijie et al. (2017) tested the performance of DA on different classification problems with the pre-trained model AlexNet. The training data are separated into three scales: small, medium and large, with 200, 500 and 1000 samples per class, respectively. The ImageNet dataset was used to identify a subset of ten classes, and the size of the dataset was doubled and tripled with different augmentation techniques. The percentage gain in accuracy is substantially higher with the smaller dataset. According to the paper, triple combinations can reduce performance, which might be due to the images being overly augmented by triple pairings. The DA techniques that involve fundamental image manipulations are subsequently elucidated.

5.1.1 Geometric transformations

These transformations alter an image's shape or geometry by translating the individual pixel values to new locations. Several studies, such as Rodrigues et al. (2019), Majurski et al. (2019) utilised geometric transformations as a DA technique on data. Some effective geometric transformations include flipping, rotation, and cropping methods. For example, Krizhevsky et al. (2017) used flipping to supplement the ImageNet dataset. Due to the popularity of this study, it became one of the most used augmentation schemes. These transformations are computationally efficient and straightforward because only rows of image matrices need to be averted. When applying these transformations, safety consideration is essential concerning label preservation of the data sample after the transformation (Shorten and Khoshgoftaar 2019). For instance, rotation and flips are often safe on ImageNet problems, such as cat vs dog datasets, but not on digit identification datasets, such as MNIST or SVNH datasets, because of digits like "6" versus "9". The ability of a model to respond could potentially be strengthened by non-label transform preservation, indicating that the model is uncertain about its prediction. However, implementing this requires post-augmentation label refinement (Bagherinezhad et al. 2018), which is expensive computationally. One more augmentation strategy involves modifications in the colour channels (Shorten and Khoshgoftaar 2019). Cropping samples by creating new image patches or selecting central patches for image data with varying height and width sizes. Some commonly employed DA techniques include rotation, translation, and noise injection (Moreno-Barea et al. 2019).

Geometric transformations are excellent remedies for positional biases in training data. Many possible causes of bias arise when there are discrepancies in the distribution of training data from testing data. These transformations are valuable when positional biases are present, such as in a facial recognition dataset where every face is properly centred in the image (Zhang et al. 2020). These transformations are effective not just for their remarkable capacity to eliminate positional biases but also for their ease of implementation. However, they have several drawbacks, including increased memory for storing augmented data, computational costs, and increased training time. Finally, the biases between training and test data in applications, such as medical image analysis, often involve more complicated factors beyond positional and translational variances. As a result, the applications of geometric transformations are minimal.

5.1.2 Photometric methods

These transformations modify RGB channels based on predetermined heuristics by shifting each pixel value (r, g, b) to a new pixel value (r ′, g′, b′). This manipulation alters the lighting and colour of the image while preserving the geometry. As a result, the efficiency of colour photometric modifications is relatively simple to grasp (Shorten and Khoshgoftaar 2019). A simple remedy for images that are too bright or too dark is to iterate through them and modify the intensity values by a predefined amount. Some other modification methods include splicing off individual RGB colour matrices and limiting intensity to a particular maximum and minimum value. The intrinsic colour representation in digital images provides a range of augmentation techniques.

5.1.3 Colour space transformation

Transforming the RGB matrices into a single grayscale image simplifies the representation of image datasets, resulting in smaller images of *height* × *width* × *Channel* dimension with less complexity. Nevertheless, it has been demonstrated that this reduces performance accuracy. The study Chatfield et al. (2014) discovered a 3% decline in classification performance between grayscale and RGB images using ImageNet (Deng et al. 2010) and the PASCAL (Everingham, et al. 2009) VOC dataset. Like geometric transformations, colour space conversions have certain disadvantages, such as requiring more memory, high transformation costs, and long training time. Additionally, colour transformations may lead to omitting important colour information and, thus, may potentially impact label preservation. The study by Wah et al. (2011) showed that DA increased CNN classification performance. In terms of Top-1 and Top-5 scores, geometric augmentation schemes performed better than the photometric schemes.

5.1.4 Kernel filter-based data augmentation

Kernel filtering techniques are used to sharpen and pre-process images. For example, a Gaussian Blur Filter produces blurred images, but sharper images are formed using a high-contrast vertical or horizontal edge filter. Image sharpening for DA may capture additional details about things of interest. The augmentation procedure, termed PatchShufe Regularization, proposed in Kang et al. (2017) employs a kernel filter that swaps intensity values in a sliding window at random. Experiments were conducted using ResNet CNN architecture with varying filter widths and adjusting the pixel shuffling probabilities at each step. This method achieved a 5.66% error rate on CIFAR-10 compared to a 6.33% error rate without including PatchShufe Regularization. In DA, kernel filters are relatively less explored.

5.1.5 Mixing images to augment the dataset

Averaging the pixel intensities by combining images is an unconventional DA approach. In the research Inoue (2018), two or more images were cropped randomly from 256 × 256 to 224 × 224 and flipped horizontally. The samples were combined by taking an average of the pixel values for each RGB channel. Consequently, a mixed image was created, which was then utilised to train the classifier algorithm. The new image's label was found to be identical to the label applied to the first randomly picked image. After employing the SamplePairing DA approach on the CIFAR-10 dataset, there was a drop in the error rate from 8.22 to 6.93%. When the researchers tested a smaller dataset, CIFAR-10 decreased to 100 samples per category, i.e. 1000 total samples. SamplePairing lowered the error rate from 43.1 to 31.0% using the smaller dataset.

5.1.6 Random erasing

Random erasing (Zhong et al.) is another intriguing DA approach inspired by dropout regularisation processes. Specifically, random erasing is similar to dropout except that it occurs in the input data space rather than being integrated into the network architecture. This approach was created primarily to address image identification issues caused by occlusion, which occurs when some aspects of an item are obscured (Shorten and Khoshgoftaar

Table 2 Overview of DA techniques for addressing small data challenges

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on accuracy or error rate	Remarks
Krizhevsky et al. (2017)	ILSVRC (subset of ImageNet)	Roughly 1000 images per class	Traditional DA translation and pixel values of RGB channels in train data are modified	Reduces the error rate by 1%	Since only rows of the matrix (image) are reversed it is efficient computationally and reduces overfitting
Shijie et al. (2017)	(1) Subset of CIFAR 10 (2) ImageNet 10	Three scales of training datasets are employed: (a) small-scale 200 samples in each class (b) Medium-scale: 1000 samples in each class (c) Large-scale 5000 samples in each class	Traditional and GAN-based DA GAN/WGAN, flipping, shifting, color jittering, cropping, noise, PCA jittering, rotation, and their combinations	Accuracy on (1) dataset Flipping + cropping=+ 3% Flipping + WGAN=+ 3.5% WGAN +cropping=+ 2% Flipping + cropping + rotation = +0.9% WGAN + flipping + cropping =− 1% (decrease) Accuracy on (2) dataset Flipping + cropping=+ 2% flipping + WGAN=+ 2.5%	Pair combinations outperform triple combinations in terms of total performance
Chatfield et al. (2014)	(1) PASCAL VOC 2007 dataset (2) ILSVRC-2012 challenge dataset (3) Caltech-101 (4) Caltech-256	The (1) dataset has 10,000 samples with 20 classes The dataset (2) has 1000 classes obtained from the ImageNet dataset that consists of 1.2 M training images, Datasets (3) and (4) has randomly 30 and 60 training samples per class, respectively	Traditional DA (1) Flip (2) Flip and cropping With deep and shallow CNN network. The CNN network has three versions CNN-fast (CNN-F), CNN-medium (CNN-M) and CNN-slow (CNN-S) Improved Fisher Vector (IFV)	(1) DA improves the performance by + 3% consistently for both IFV and CNN (2) Both CNN-M and CNN-S showed better results than CNN-F by a 2–3% margin (3) CNN-S with hinge loss for a small VOC dataset showed a + 2.7% improvement in accuracy	DA significantly improves the performance of shallow CNN Deep architectures' performance is higher than shallow architectures

Table 2 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on accuracy or error rate	Remarks
Inoue (2018)	(1) ILSVRC 2012 (2) CIFAR-10 (3) CIFAR-100 (4) SVNH	In (1), 1000 object classes from ImageNet. The dataset is also reduced and used with 100 classes In (2), there are 50,000 training images; in (3) dataset, there are 100 categories of images for CIFAR The (4) dataset consists of 73,257 training images A reduced CIFAR 10 dataset with only 100 samples per label was also used	DA approach SamplePairing, to synthesize a new sample by overlaying another image randomly selected from the train set over another image By overlapping two images, N^2 images can be created from N images	The error rate decreases from 33.5 to 29.0% for the (1) dataset utilizing GoogleNet, and for the (2) dataset, the error rate is reduced from 8.22 to 6.93% With this technique, there is an improvement in the error rate of all the testing datasets With only 100 samples per class, the technique helps to reduce the error rate from 43.1 to 31.0%	The study shows that the small CIFAR-10 dataset with only 1000 images decreased the error rate from 43.1 to 31.0% The study shows that the proposed method is more suitable for small data problems such as medical data classification
Zhang et al. (2021a)	(1) CIFAR-10 (2) CIFAR-100 (3) SVNH	In (1) dataset, only 50 to 100 samples are selected from each category The training is performed for the (2) dataset with 50, 80, 100, and 200 samples	The proposed deep adversarial DA (DADA) approach solves the small data problem Pre-trained VGG-16 as a classifier ResNet-56 as a replacement to VGG-16 to train deeper classifiers It also proposes a new loss for the GAN discriminator called $2k$ loss	The pre-trained VGG baseline accuracy is 82.86%, and DADA attains a higher accuracy of + 85.71% On the (1) dataset DADA achieves + 6% more accuracy than TDA DADA outperforms TDA with 500 and 1000 samples. With less than 200 samples, there is extreme overfitting On (3) dataset using extensively small data regimes, DADA gives the best performance DADA obtained a minor negative impact with many labelled samples (500 per class) than TDA	DADA outperforms TDA in small sample sizes (without augmentation The combination of DADA + TDA achieves the highest results With very little data, deeper classifiers are not the right option

Table 2 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on accuracy or error rate	Remarks
Karras et al. (2020)	(1) METFACES (2) BRECAHAD (3) Animal faces (AFHQ) (4) CIFAR-10	The (1) dataset has 1336 high-quality face images The (2) dataset has 162 breast cancer histopathology samples The (3) dataset has 5000 close-ups per class for dogs' cats, and wildlife The (4) has 50*k* images in 10 classes	The study proposes Adaptive Discriminator Augmentation (ADA) method significantly stabilizes training in short data problems	On (3) dataset SOTA FID improved from 5.59 to 2.42, and Inception Score from $9.58to10.24$ On (1) dataset (FID) Baseline = 57.26 ADA = 18.22 On (2) dataset (FID) Baseline = 97.72 ADA = 15.71	The ADA improves over baseline StyleGAN2

2019). Random erasing randomly selects a rectangular section in an image and replaces its pixels with random values during training. This approach generates training samples with varying degrees of occlusion, which decreases the danger of overfitting and makes the model resistant to occlusion. Random erasing does not need parameter learning, is easy to implement, and can be utilised with most CNN-based recognition algorithms. Although it is simple, random erasing complements commonly used DA techniques, such as flipping and random cropping, and offers consistent enhancement over robust baselines in image classification, object recognition, and human reidentification. Table 2 shows the summary of approaches based on DA for solving small data problems.

Traditional DA techniques significantly enhance the predictive performance of DL models and are used extensively across various applications. However, these methods often involve laborious manual efforts. Identifying the optimal DA strategy is highly dependent on the specific dataset and task, necessitating testing a virtually infinite number of augmentation permutations to discover the most effective approach. Additionally, manually crafted augmentations are typically limited in variety. Furthermore, different types of augmentations yield better results for different DL tasks, making the selection of an appropriate augmentation method a complex and challenging problem. Additionally, techniques that enhance generalization on one dataset may not be effective on others. For example, research (DeVries and Taylor 2017) showed that while CutOut (DeVries and Taylor 2017) boosts performance on CIFAR-10, it does not have the same effect on the ImageNet dataset. Furthermore, another study (Raileanu et al. 2021) suggests that traditional DA methods are not well-suited for reinforcement learning tasks.

Extensive research is being conducted to automate the process of DA. Automated Machine Learning (AutoML) (Kim et al. 2022) aims to automate all aspects of designing, training, deploying, and monitoring ML solutions. AutoML frameworks can perform DA, feature engineering, and even construct the network architecture of DL models.

The concept of automated DA involves creating a variety of basic transformation functions (e.g., rotations, flipping, color jittering, solarization, scaling) and then using AutoML techniques to algorithmically apply different combinations of these operations to the data. The goal is to select the most effective set of DA operations. Typically, black-box optimization techniques are employed to determine the best augmentation strategies. This optimization process must identify not only the relevant transformations but also the optimal levels for each transformation. For image augmentation, these levels might include rotation angles, translation offsets, and saturation values. The AutoML field is relatively new, and extensive research is required in this field to enhance its capability of selecting augmentation strategies (Mumuni and Mumuni 2024).

5.2 GANS for solving limited data problems

> "GANs are the most interesting idea in the last ten years in Machine Learning".
> – Facebook AI research director Wang (2020)

DA may not be enough to train DL models efficiently when training data are *lacking* (dos Santos Tanaka and Aranha 2019). It frequently fails to produce the variance needed to accurately reflect the whole task distribution (www.causaLens.com), which leads to model overfitting. In order to ensure that a larger portion of the task distribution may be represented, similar data can be generated to enhance the variance in the training data. The term "Generative Adversarial Network(s)" or GAN(s) for short, was initially introduced by Goodfellow (Goodfellow, et al. 2014). GANs are a generative model that synthesises new images based on training data. The study Marchesi (2017) created high-resolution photorealistic images (up to 1024 × 1024 pixels) using less than 2000 images. They used the DCGAN (Gao et al. 2018) version of generative modelling. The generated photorealistic images can prove to be a good asset for commercial use of the samples. The study proposed Deep Adversarial Data Augmentation (DADA) model, or learning-based DA on the GAN model (Zhang et al. 2021a). The paper also offers a novel loss for the GAN discriminator, referred to as $2k$ loss, compared to the $k+1$ loss employed by many existing GANs. The experiments were conducted on the CIFAR-10, CIFAR100 and SVNH datasets, which were sampled to simulate very low data regimes (less than 1000). The study compared augmentation based on DADA, traditional methods with and without augmentation by measuring the classifier's performance. The experimental findings reveal that DADA outperforms both TDA and a few GAN-based models significantly.

In another study dos Santos Tanaka and Aranha (2019) on augmenting small datasets using the GAN model, high-quality skin lesion samples were synthesised by employing the style-GAN model. This study utilised the classification challenge International Skin Imaging Collaboration (ISIC 2018) dataset (Codella et al. 2018), which consists of 10,015 images of skin lesions images. The dataset is imbalanced and categorised among seven classes. More than 77% of samples belong to only two categories, namely melanocytic nevi and melanoma. The other classes only have a few hundred samples, for instance, vascular skin lesions have only 115 samples. GAN-based models were used to synthesise images, which is very challenging with such a small dataset. The CNN model exhibited improved classification accuracy when synthesised images were added to the training set (Frid-Adar et al. 2018).

One of the drawbacks of training GANs on limited datasets is that only images with small variances are produced around a restricted number of modes, which characterise the manifold learned by the generator (Bowles et al. 2018). There must be sufficient data to enable the learning of a smooth manifold. However, with so much annotated data available, performing augmentation is probably unnecessary, making the use of a GAN unnecessary. One study (Bowles et al. 2018) aimed to transition to learning this smooth manifold from a considerably smaller set of labelled images. This was accomplished by applying a method influenced by TL, many unlabelled images, and a few labelled images. In small datasets, the parameters of the network are not fully determined, resulting in poor generalisation (Antoniou et al. 2017). The study by Antoniou et al. (2017) demonstrated the possibility of a much more comprehensive range of augmentations. The Data Augmentation Generative Adversarial Network(s) (DAGAN) model is based on the CGAN model and shows a noteworthy enhancement in the overall performance and generalisation when applied to augment data in low data regimes. The paper Frid-Adar et al. (2018) utilised the GAN model to augment CT images of the liver on a dataset with only 182 liver images of three categories (65 haemangiomas, 64 metastasis, and 53 cysts images). CNN was used for classification purposes, and its performance was evaluated by comparison with TDA and GAN-based augmentation. The study revealed better performance with GAN-based augmentation.

The authors of Karras et al. (2020) suggest an adaptive discriminator augmentation technique that significantly stabilises training in restricted data regimes. Without changing loss functions or network designs, the approach can be utilised to start over and improve an existing GAN on a different dataset. With a number of restricted data problems, the authors used this technique and produced noteworthy results, consequently in applications where the restricted data synthesis appears to yield poor-quality synthetic images. This method

Table 3 Overview of GAN-based techniques for addressing small data challenges

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Marchesi (2017)	Two datasets of images collected from Social Media and Magazines	1807 images are in the first dataset, and the second dataset has 1796 images	DCGAN-based optimized image synthesis process to create high-resolution photorealistic images (up to 1024 × 1024 pixels)	–	The generated images are of excellent quality and show that GANs can be exploited to generate data for applications with small datasets available
Qin et al. (2020)	(1) ISIC 2018	The (1) dataset has 10,015 skin lesion images categorized into seven classes It is imbalanced Sample count ranges from 115 to 6705 per class	Image synthesis-based DA Style GAN for the synthesis of skin lesion images	Increase in accuracy + 1.6%, sensitivity by +24.4%, specificity by 3.6% and average precision by 5.6% than the CNN-based classifier	High-quality skin lesion images are generated with a limited dataset, leading to classification improvement
Bowles et al. (2018)	(1) MICCAI 2013 Grand Challenge on Multi-Atlas 35 images from the OASIS-1 dataset Unlabelled dataset	The dataset has 35 images, of which five are discarded and 24 are used for training The unlabelled dataset contains 436 images	Before producing training samples, the model combines unlabelled and labelled images with train data within the GAN framework Segmentation is evaluated on DSC	Experiments on a single fold show decreased DSC with more samples (12 and 24) Overall average DSC is 0.68, 0.73, 0.73, 0.68 and 0.64 when training on 1, 3, 6, 12 and 24 images, respectively	The model shows significant segmentation improvement with limited data

Table 3 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Antoniou et al. (2017)	1. Omniglot 2. EMNIST 3. VGG-Face	In (1), the classifier is trained with 5,10,15 samples per class From the (2) dataset, 15,20,25,100 samples per class are selected From the (3), 5, 15, and 25 samples per class are used	DAGAN: A CGAN-based DA model	Dataset (1) shows +13% increase in accuracy (69% to 82%) EMNIST + 2.1% increase in accuracy VGG-Face + 7.5% (from 4.5 to 12%) On Matching networks An accuracy increases + 0.5% and 1.8% on (1) and (2) datasets	DAGAN is a flexible model that learns the DA automatically DAGAN improve the performance of classifiers even after TDA is applied
Frid-Adar et al. (2018)	CT images of the Liver	The dataset has a total of 182 images with three classes (65 haemangiomas, 64 metastasis and 53 cysts)	Synthetic DA based on generating synthetic images by utilizing GAN models The CNN is used for the classification	TDA shows +78% sensitivity and +88.4% specificity. The results improved with a sensitivity of +85.7% and specificity of +92.4% after augmenting with synthetic data	The classifier's performance improves when synthesized samples are added to the dataset

Table 3 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Song et al. (2022), Ma et al. (2022)	HRSSRD: High-Resolution SAR Ship Recognition Dataset	The dataset consists of 3355 military ships and 2565 civil ships	Two-stage TL approach that combines data-level and feature-level knowledge transfer In the first stage, Cycle GAN is adopted A new network named Domain Transfer using Adversarial learning and Metric learning (DTAM) is introduced The network aims to help in the recognition task of the military and civilian ship	The approach used achieves a test accuracy of 71%	The two-stage model showed potential cross-modality transfer abilities by demonstrating promising performance on the dataset
Bargshady et al. (2022)	1. Extensive COVID-19 X-ray 2. CT Chest Images	The (1) dataset contains 5500 normal and 4044 abnormal chest scans The (2) dataset has 2628 normal CT Scans and 5427 COVID-19 CT Scans	Semi-supervised Cycle-GAN to augment the training dataset Finetuned Inception V3 TL model trained specifically for COVID-19 detection	The study showed an accuracy of 94.2%, AUC 92.2%, MSE 0.27 and MAE 0.16.	The study claims that the novel augmentation method can assist the model in being employed for smaller databases

can be applied to improve the data. Table 3 shows the summary of techniques that consist of GAN-based techniques for solving small data problems applied in the literature.

GANs have emerged as the most versatile generative models due to their unparalleled data synthesis capabilities across various domains. However, training GANs remains highly unstable for several reasons, including vanishing or exploding gradients and oscillatory or diverging dynamics when attempting to find Nash equilibrium. Although recent studies have proposed various solutions to enhance training stability, achieving stable GAN training continues to be an open research challenge. Another issue with GANs is mode collapse, where the model generates limited diversity in its outputs. Numerous solutions have been proposed to mitigate this problem, including modifications to network structures, optimized loss functions, and improved training algorithms. While these techniques have partially addressed mode collapse, further research is needed to enhance the diversity of generated data, particularly in large-scale datasets (Ahmad et al. 2024).

5.3 Transfer learning

> "Transfer Learning will be the next ML Success".
> – Andrew Ng NIPS 2016 Tutorial

In traditional ML models, it is typically assumed that the training and testing data are from the same distribution. However, in many real-world scenarios, this assumption does not always hold. The solutions discussed in Sects. 5.1 and 5.2 address the one problem of these models, i.e. insufficient data. The other challenge, like incomparable computation, can be solved with the help of cloud computing and distributed learning. Nevertheless, there are several drawbacks to these mentioned solutions, such as high cost, inefficiency, and security. TL addresses all three challenges and has recently become a viable approach to mitigate such problems (Chen et al. 2021). Some recent survey papers that utilised TL can be referred to in this regard (Niu et al. 2020; Tan et al. 2018; Zhuang et al. 2021). To avoid starting from the beginning with big datasets, TL primarily seeks to complete the target task by leveraging the information gained from source tasks across multiple domains (Pan et al. 2011). TL is frequently used to minimise the impact of small datasets (Pan et al. 2011; Ibragimov and Xing 2017; Interian et al. 2018). On small datasets, the algorithms overfit easily (Yosinski et al. 2014). It has been demonstrated that feature transfer performance degrades as the source and target become increasingly dissimilar (Yosinski et al. 2014). The relation between a model's training data size and trainable parameters count significantly impacts model performance (Romero et al. 2019). As a result, there is an increasing interest in employing TL to train big models, such as CNNs, in areas with a dearth of training data or other limitations. It has been proven that small target datasets are considerably subtler to variations in TL hyperparameters; hence, it is helpful to distinguish across target dataset sizes (Plested and Gedeon 2019a). The research (Plested and Gedeon 2019a) demonstrates that the frequently used TL protocols for small target datasets lead to increased overfitting and dramatically lower accuracy than optimum protocols (Goceri 2021). The relationship between the appropriate layers count to transfer, and the hyperparameters used for fine-tuning is shown in the study. The work of Yosinski et al. (2014) represents the most organised and extensive examination of TL on CNNs to-date. After being pre-trained on a comparable dataset, they demonstrated that fine-tuned networks generalise better than those trained directly on a massive target dataset. Performance on the target dataset improves with more source datasets (Plested and Gedeon 2022). However, pre-training on bigger, more broad source datasets can sometimes outperform source data that has been carefully selected to more closely resemble target data (Singh et al. 2022; Mormont et al. 2018). A preliminary study (Huh et al. 2016) demonstrates that extra pre-training data is only helpful if it is highly related to the target task. In certain circumstances, augmenting with irrelevant training data degrades performance.

The study (Zhao et al. 2022) presents a deep TL strategy utilising CNN to address the cross-domain diagnostic challenge. The technique extracts features with CNN from the source domain data and generates a pre-trained model. Subsequently, the model is fine-tuned with a small dataset from the target domain through TL strategy, leading to the final intelligent diagnostic model. The paper Zhao et al. (2022) employed a massive-training artificial neural network (MTANN) to detect lung nodules in a small dataset of lung CT scans. The model performed considerably better than TL-based AlexNet.

Nevertheless, pre-training with the smaller source dataset resulted in much lower performance when ImageNet 5k or 9k, or the more problem-specific Caltech Birds (Wah et al. 2011) and Places365 (Zhou et al. 2018), was used as the target dataset. There are just a handful of large image datasets unrelated to the image classification tasks often used for pre-training Places365 with 1.8 million training images as a source task. It was demonstrated that when the source and target datasets were less connected, the larger and more diversified the source training dataset, the better the results on the target dataset.

5.3.1 Short target datasets

When the magnitude of the target dataset diminishes, TL becomes more heavily reliant on it. TL hyperparameters significantly influence performance as the target dataset magnitude decreases (Plested and Gedeon 2019b). Two challenging variables affect TL's performance as the target dataset's size diminishes: (1) the empirical risk estimate becomes less trustworthy, increasing the likelihood of overfitting in the target dataset; and (2) the pre-trained weights implicitly regularise the fine-tuned model, and the final weights do not deviate much from their pre-trained values (Raghu et al. 2019; Neyshabur et al. 2020). As a result of Point 1, there is a growing need to apply TL and other approaches to prevent overfitting. The implicit regularisation mentioned in Point 2 may help to reduce the overfitting of the empirical risk estimate mentioned in Point 1. If the transferred weights from the source dataset are inappropriate for the target dataset, it might have a negative impact (negative transfer). Point 1 (Plested and Gedeon 2019b) can exacerbate the negative influence on performance when the weight and features created are confined to being far from ideal.

5.3.2 Smaller target datasets with similar tasks

A well-known work on TL (Yosinski et al. 2014) used an AlexNet (Krizhevsky et al. 2017) with vast, tightly connected source and target datasets. In Plested and Gedeon (2019b), the same experiments were performed with various datasets but with a smaller target dataset size than used by Yosinski et al. (2014). Compared to conventional hyperparameters from Yosinski et al. (2014), there was a considerable improvement when employing more optimum TL hyperparameters. As the sample size shrank, the improvement in accuracy was significant. The average accuracy increased from 20.86 to 30.12% for the lowest target dataset of only ten samples for each of the 500 classes while employing optimal instead of commonly utilised hyperparameters. The study also demonstrates that the conventional method of transferring all but the final classification layer is not the best. The improvement of TL over random initialisation correlates positively when the target and source datasets are closely related. This was demonstrated in Deng et al. (2010) by transferring the pre-trained CNN model to significantly smaller datasets. However, the improvements correlate negatively with the target dataset size. The study Kornblith et al.(2019) states that the performance improvement of the model trained from scratch is marginal for CARS and FGVC AIRCRAFT (Kornblith et al. 2019) datasets, which are approximately 0.6% and 0.2%, respectively. This is because the similarity between the source (ImageNet 1k) and the target datasets is very low. According to Kornblith et al. (2019), there is a negative association between target dataset size and the improvement over the baseline, but the lower the baseline accuracy numbers, the greater the gain in accuracy since there is more space for development.

5.3.3 Smaller target datasets with less similar tasks

TL often works better on smaller target datasets that are more closely connected to the source dataset than on big datasets that are less related (Kornblith et al. 2019). Self-supervised learning approaches customised to a specific task and applied to more comparable but unlabelled source datasets usually outperform supervised learning techniques used for less similar source datasets (Azizi et al. 2021; Zoph et al. 2020b). Recent research shows that TL may accelerate convergence even when the source and target datasets are vastly dissimilar (Azizi et al. 2021; Siuly and Zhang 2016). Some tasks that rely on TL because of having significantly less target datasets include face detection (Zhang et al. 2020), Facial Expression Recognition (FER) (Li and Deng 2022; Revina and Emmanuel 2021) and medical image diagnosis (Siuly and Zhang 2016; Chen et al. 2018; Singha et al. 2021; Anwar et al. 2018; Xu et al. 2021; Afshar et al. 2019). These applications usually have very few training datasets available. Some unique challenges arise with this type of research. In the case of face recognition, there is minimal variation among the samples within class, because each class represents only one individual. One more challenge in this research is that there can be hundreds of thousands or even millions of classes, which is much higher than the classes in the ImageNet dataset. TL plays a significant role in these types of problems. A DL model can be trained with celebrity faces that are publicly available, and then TL can be used for limited datasets (Plested and Gedeon 2022). In the case of FER, the data are often limited, making these problems challenging. Less than 10,000 images or videos make up the majority of popular FER datasets. Even the largest ones that are frequently used only feature 100 different subjects, which results in a great correlation between the individual images. The fact that there are significant intraclass variances caused by many personal characteristics, such as age, gender, ethnic origin, and expressiveness degree, presents an additional obstacle specific to facial expression recognition (Li and Deng 2022).

The study Deng et al. (2010) demonstrated that pre-training with source data that are more closely related to the target dataset improves performance. Pre-training on a sizable facial recognition dataset outperformed training on the more general and distantly related ImageNet 1k (Deng et al. 2010). Performance was enhanced by a multi-stage pre-training pipeline in Ng et al. (2015), which uses an extensive FER dataset for interim fine-tuning before the final fine-tuning on the short target dataset. Medical imaging tasks are another use of TL. Regarding medical imaging, DL models face two issues: (1) data scarcity—the medical image training databases are frequently in the hundreds or thousands, which is insufficient for a DL model to get effectively trained (Mazurowski et al. 2019); and (2) imbalanced datasets—there are frequently many more examples of healthy data samples than unhealthy ones. DL models are not adequately trained as a result of these issues. TL techniques are used in every medical imaging modality, including CT scans, pathological samples, X-rays, CET, and MRI (Mazurowski et al. 2019). Despite this, there is relatively little research on the optimal practices for deep TL in medical scan identification. In Tajbakhsh et al. (2016), the researchers investigated AlexNet pre-trained on ImageNet 1k with and without fine-tuning, an AlexNet trained from scratch, and traditional models with hand-created features. Employing a pre-trained AlexNet CNN with appropriate fine-tuning regularly outperformed or is on par with training from random initialisation and conventional methods. While the performance advantage from utilising a trained and fine-tuned AlexNet was slight for comparatively bigger target datasets, it became considerably more critical when the target dataset size was lowered. The study Tajbakhsh et al. (2016) utilised a simplified AlexNet (AlexNet with lower parameters) architecture as a DL model. The paper constructed a classifier for online face expressions using a short dataset of only 480 images. As measured by average fivefold cross-validation, the model achieved an accuracy of 78.69%. It is also pertinent to mention that expanding the dataset boosted the classifier's accuracy.

The study Keshari et al. (2020) presents a Dynamic Attention Pooling (DAP) method that helps to extract global knowledge from the most discriminative sub-part of the feature map. The performance of the DAP was analysed with a ResNet model on comparatively small publicly existing datasets, such as SVHN, C10, C100, and Tiny-ImageNet. The proposed ResNet-based DAP showed an improvement of 1.75%, 0.47%, and 1.87% on C10, C100, and TinyImageNet, respectively. However, several recent results fit this category where deep TL shows slight or no increase over random initialisation (Raghu et al. 2019; He et al. 2019; Zoph et al. 2020a). The findings of the study Barbero-Aparicio et al. (2024) highlight the potential of deep transfer learning as a cutting-edge approach for protein fitness prediction. Researchers can attain performance levels that exceed those of traditional supervised and semi-supervised methods by utilising pre-trained models and fine-tuning them on small datasets. Table 4 presents the summary of techniques based on TL for solving small data problems applied in the literature. The success of TL is not always assured. When the source and target tasks are unrelated, or if the transferred representation lacks sufficient information relevant to the target task, TL may fail to improve and could even degrade the performance compared to training from scratch on the target task, a phenomenon known as negative transfer (Zhang et al. 2023). Research focused on understanding when and what to transfer between tasks to ensure the effectiveness of transfer learning is an essential area of study (Tan et al. 2024). A current trend in transferability research (Tan et al. 2024) focuses on efficiently predicting transfer performance beforehand with minimal or no training of the transfer model. Several effective transferability metrics have been introduced, such as negative conditional entropy (NCE) (Tran et al. 2019) and the H-score (Bao et al. 2019). The study (Barbero-Aparicio et al. 2024) presents a DL model that leverages TL, using the pre-trained Inception V3 network to apply its knowledge to a small (Barbero-Aparicio et al. 2024), labeled dataset within the construction context. This enables the model to effectively learn meaningful representations from the limited training data, thereby enhancing its accuracy in classifying material conditions. Moreover, GLCM-based texture features are extracted from the images to capture textural variations in construction materials. The proposed approach achieved an accuracy of 97% with 208 images and 71% with 70 images, respectively.

Table 4 Overview of TL-based techniques for addressing small data challenges

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Goceri (2021)	(1) DermWeb (2) DermatoWeb (3) DermQuest	A total of 725 and 145 images per class	Pre-trained on ImageNet weights SqueezeNet, MobileNet, ShuffleNet etc	Skin disease prediction accuracy 94.76%	Pre-training as an efficient approach for solving small dataset problems
Huh et al. (2016)	1. Necrosis 2. ProliferativePattern 3. CellInclusion 4. MouseLba 5. HumanLba 6. Lung 7. Breast 8. Glomeruli	(1), (2), (3), (7), and (8) datasets have 695, 1179, 1644, 14055, and 12157 train images with two classes, respectively. Dataset (4) has 1722, (5) has 4051 and (6) has 4881 training samples with 8, 9,10 classes, respectively	Feature Extraction by utilizing Deep CNN pre-trained (VGG16, VGG19, Inception, InceptionV3, ResNet50 ResNetV2, DenseNet201, and MobileNet) on ImageNet classification with the help of SVM	On multiclass datasets (4), (5) and (6) fine-tuning shows the highest accuracy, 87%, 94% and 86%, respectively. InnerDenseNet and MergeNetworks show the highest accuracy on (p) 89.84% in both, Feature selection leads the highest accuracy on (N), which is 99%, and fine-tuning shows the highest accuracy on the rest	The study observed that residual and dense networks yield the best performances The study also shows that fine-tuning outperforms features from off-the-shelf networks' last layer
Zhao et al. (2022)	Rolling bearing data set from the Bearing Data Center at Case Western Reserve University (CWRU)	The data is divided into four datasets A, B, C, and D, where A, B, and C are considered source domain data, each with 24 samples, and each dataset contains 96 examples Dataset D is the target domain data divided into training and test sets	CNN-based Deep TL strategy to handle the cross-domain diagnostic challenge	The model showed an accuracy improvement of 20% with AB, AC, and BC source domain data	The study deals with the reduction in diagnostic accuracy Because of domain difference

Table 4 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Zhou et al. (2018)	1. PASCAL-VOC 2007 2. PASCAL-VOC 2012 3. SUN dataset	The (1) dataset has 10,000 samples with 20 classes The (2) dataset contains 1000 object categories from ImageNet with roughly 1.2 M images	TL	By training five AlexNet models (1) with 1000, 500, 250 samples per class is found as mAP 58.3, 57.0 and 54.6, and the similar trend is seen with (2) and (3) dataset Pre-training with only 127 categories instead of 1000 results in a performance drop by 2.8 mAP on the PASCAL-DET dataset	The study claims that increasing training data does not continuously improve performance and sometimes even worsens it
Neyshabur et al. 2020)	1. Retina dataset, consists of retinal fundus images 2. CheXpert that contain X-ray images	In the case of a limited dataset, this study utilizes dataset (1) with only 5000 data points selected for training	TL for medical imaging	TL has a more significant effect with very small datasets, and model size has a confounding effect The study shows that TL helps larger models like ResNet50 with an improvement from 92.2% (random initialization) to 94.6% (TL), and for smaller CBR models, the improvement is 0.3% (CBR-Large T) and 0.1% (CBR-Large W) only	Surprisingly, TL delivers little improvement to performance, as simple, lightweight models perform on par with ImageNet on two large-scale medical imaging tasks

Table 4 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Kornblith et al. (2019)	1. Food-101 2. CIFAR 10 3. CIFAR 100 4. Birdsnap 5. SUN397 6. Stanford Cars 7. FGVC Aircraft 8. PASCAL VOC 2007 8. Oxford-IIIT Pets 9 Caltech-101 10. Oxford 102 Flowers 11. Describable Textures (DTD)	(1), (2), (3), (4) and (5) datasets are comparatively larger. However, in (6), 8144 training samples are used, totalling 196 classes. In (7) dataset, there are 6667 training samples and 20 classes. (8) dataset has 5011 training samples with 37 classes. The dataset (9) has 3060 training samples and 102 classes. The Dataset (10) has 2040 training samples, and (11) has a total of 3760 training samples	TL for classification across 16 modern CNNs	TL with linear classification, more effective ImageNet networks deliver better penultimate layer features (r = 0.99), and when the whole network is fine-tuned, the results are better (r = 0.96) Logistic regression is an effective baseline when data is limited (47–800 total instances), delivering accuracy on par with or higher than fine-tuning	Pre-training on ImageNet yields negligible improvement on small fine-grained image classification datasets, signifying that ImageNet learned features do not transfer well to fine-grained tasks
Azizi et al. (2021)	1. Dermatology skin images 2. CheXpert	In (1) dataset total of 15,340 samples were for training. There are 419 unique condition labels in the dataset In (2) dataset, the study utilized a total of 67,429 training images that belong to 5 pathologies	Self-supervised learning as a pre-training strategy for medical image classification The study proposes a Multi-Instance Contrastive Learning (MICLe) to generalise existing contrastive learning approaches MICL enhances the performance of self-supervised models	On (1) dataset classification, the proposed self-supervised technique improves by 6.7% in top-1 accuracy In terms of mean AUC, self-supervised learning surpasses robust supervised baselines pre-trained on ImageNet by 1.1%	Self-supervised pre-training on unlabeled medical images significantly perform better than standard ImageNet pre-training and random initialization

Table 4 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Ng et al. (2015)	1. FER28 2. FER32 3. FER32 + EmotiW	The (1) dataset has a total of 28,709 training images In (2), a total of 32,298 total training images are present The (3) dataset is the combination of FER32 and EmotiW datasets	The TL technique for deep CNN models employs a two-stage supervised fine-tuning, first on datasets relevant to facial expressions and second on the contest's dataset	The study demonstrates that a validation set achieves an overall accuracy of 48.5% and 55.6%, and the test set achieves 55.6%; these scores are notably higher than the 35.96% and 39.13%, respective of the baseline	The cascade fine-tuning method outperforms single-stage fine-tuning
Tajbakhsh et al. (2016)	(1) 40 short videos of colonoscopy	The data is divided into 3800 frames of polyps and 15,100 frames without polyps Due to large negative frames a set of 1,00,000 training patches were utilised	Pre-trained deep CNNs with sufficient fine-tuning	(1) Pre-trained CNN with sufficient fine-tuning performed better or, in the worst case, showed comparable performance with a CNN trained from scratch (2) The layer-wise fine-tuning provides a practical approach to achieve optimal performance for the application considering available data	Shallow tuning of the pre-trained CNNs leads to a reduced performance than CNNs trained from scratch. Deeper fine-tuning shows comparable or even better performance than the CNNs trained from scratch The performance gap widens when size of the trained dataset is reduced
Ma et al. (2022)	1. Battery A 2. Battery B	The (lithium-ion batteries) LIB degradation experiments are carried out to form two small datasets with different discharging rates and ambient, respectively	Personalized state of health prediction of LIBs through TL-based method	The efficacy of the approach is demonstrated by comparison with ML techniques, such as CNN, RNN, long short-term memory (LSTM) CNN + TL obtains more minor prediction errors	The TL-based approach proposed in this study shows good generalizability, and a notable improvement in precision

Table 4 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Chien et al. (2022)	Historical data collected from the etching process	The data consists of 15 SVIDs for 1652 wafers. 1444 Normal wafers and 208 abnormal wafers	Fault detection and classification that employed CNN: Pre-trained VGG 16	The proposed method achieves better results in accuracy, recall, and F1-score	The strategy employs CNN TL to redefine new Fault detection and classification monitoring rules. The cycle time for yield ramping is effectively shortened with new small data

5.4 Few short learning

In 1950, Alan Turing posed a query, "can machines think?" in his famous paper "Computing Machinery and Intelligence" (Tsai and Salakhutdinov 2017). The paper states that "the idea behind digital computers may be explained by saying that these machines are intended to carry out any operations that could be done by a human computer". Turing was referring to the idea that these machines are capable of performing any task that a human computer could perform. The ultimate aim of machines is to match human intelligence. Numerous DL methods have helped AI to surpass human accuracy levels. CNN (Krizhevsky et al. 2017) and LSTM (Hochreiter and Schmidhuber 1997) are two examples of such models that have contributed to this advancement in AI. Big datasets like ImageNet, which contains 1000 categories (Krizhevsky et al. 2017), are readily available in the era of big data and are used to train DL models. Additionally, AI has advanced thanks to the development of distributed platforms and powerful processing hardware like GPUs.

ML models must generalise from a small number of instances and learn from experience in order to narrow the gap between AI and humans (Fei-Fei et al. 2006). A novel ML paradigm called Few Shot Learning (FSL) allows for the learning of new information from small datasets. ResNet (He et al. 2016) surpasses humans in ImageNet classification; however, each class in the dataset must have enough samples, which may not be feasible for all applications. For data-intensive applications, FSL can reduce data collection effort. Examples that FSL is used in include face recognition, image classification (Liu et al. 2019), object tracking (Bertinetto et al. 2016) image retrieval (Triantafillou et al. 2017), video event detection (Zhang et al. 2019), language modelling (Vinyals et al. 2016; Bansal et al. 2019), and gesture recognition (Pfister et al. 2014; Feng and Duarte 2019).

The study Sun et al. (2021) devised a model based on FSL that investigates discriminative features by emphasising critical areas in the image. The model employs the focus-area localisation method to identify visually comparable areas among different objects. Furthermore, a real-world fine-grained dataset miniPPlankton, a typical FSL dataset in marine ecological environments, was constructed and extensively validated. In this dataset, fine-grained phytoplankton images were collected using an electron microscope. However, there are only a few samples in the dataset. Image classification of plankton is becoming increasingly crucial for marine observations and aquaculture. Medical datasets are another application where FSL can have a significant impact because most of these datasets available are small and, thus, insufficient for training. FSL approaches can be very helpful in resolving these issues. Another study in this field (Medela et al. 2019) validated FSL approaches for knowledge transfer. The study focused on knowledge transfer from a well-defined source domain of colon tissue to a more general domain comprising colon, lung, and breast tissue using only a small number of training samples. With only 60 training images, FSL achieved a balanced accuracy of 90%. Other studies that have utilised limited medical datasets to investigate the use of FSL are Cai et al. (2020), Chen et al. (2020a, b), Wibowo et al. (2022), Feyjie et al. (2020). The paper Feng and Duarte (2019) presented a few shot human activity recognition technique that employs a DL approach to extract features and perform classification, while knowledge transfer is done via model parameter transfer. Due to the expensive nature of obtaining human-generated activity data and the inherent similarities among activity modes, borrowing information may be more efficient from existing activity recognition models than collecting additional data for training a new model from scratch when only limited training data are available.

In Iwata and Kumagai (2020), the authors offer an FSL approach that can predict the future value of a time series in an objective task based on a limited number of time series data in the target domain. The study Bansal et al. (2019) presents a new approach, LEOPARD, that enables optimisation-based meta-learning across tasks with distinct categories and analysis of alternative strategies for generalisation to various NLP classification problems. LEOPARD is trained using cutting-edge transformer architecture and exhibits improved generalisation to unseen tasks during training, with as little as four examples per class. In an evaluation that involved 17 NLP tasks, spanning diverse domains of entity, sentiment analysis typing, natural language inference, and many other text classification tasks, LEOPARD outperformed several robust baseline approaches by more effectively learning initial parameters for FSL than self-supervised pre-training or multi-task training, for example, yielding a 14.6% average relative improvement in accuracy on unseen problems with only four samples per class. The research Zhao et al. (2023) investigated the FSL model with TL using a short dataset of import and export commodities. They used a ResNet18 as the backbone and DA to enlarge the tiny initial dataset before training, which helped to mitigate the CNN model's overfitting issue. Also, the attention module is included in the backbone.

The paper Drumond et al. (2023) provides a few-shot motion prediction model incorporating the underlying network structure. The model employs heterogeneous sensors, showing a considerable performance increase overall relevant baselines from 10.4 to 39.3%. The study attempted to anticipate motion for previously unknown actions using only a few labelled instances. The benefit of the model is that the end users can contribute additional movements by showing an activity a few times before the model can reliably categorise and forecast future frames. Another study Zheng et al. (2022) proposed ANomaly dEtection framework with Multi-scale cONtrastive lEarning (ANEMONE), a broad framework based on contrastive learning for graph anomaly detection. The approach uses multi-scale information at the patch and context levels to detect abnormal patterns concealed in complicated networks. Comprehensive trials with ANEMONE and its variation ANEMONE-FS in totally unsupervised and few-shot anomaly detection situations show that both approaches consistently outperformed state-of-the-art methods on six benchmark datasets. Table 5 provides a summary of techniques that consist of FSL-based approaches for solving small data challenges applied in the literature.

In recent years, innovative FSL approaches have been developed to address various computer vision challenges, including object detection, 3D reconstruction, and video inputs. Moreover, current few-shot image classification methods primarily focus on general datasets (Gharoun et al. 2024). Due to data security concerns and data collection challenges, there is limited research on specialized datasets. Consequently, constructing larger-scale, higher-quality image datasets and developing dedicated datasets for specific fields is a critical research challenge in few-shot image classification. FSL techniques often encounter overfitting issues because they rely on very few samples. The limited diversity of these samples makes it difficult to learn complex features effectively. Consequently, considerable research is required to overcome these challenges and enhance the performance of FSL techniques. FSL is still in its early stages, particularly in its application to multimodal data. Some research in this area, such as studies Peng et al. (2019), Xing et al. (2019) has explored their use for image classification. For instance, study Drumond et al. (2023) combined image features and semantic features for FSL classification. Similarly, the AM3 method proposed in Zheng et al. (2022) adaptively and selectively combines semantic and visual features, significantly enhancing the classification performance compared to the original algorithm.

Table 5 Overview of FSL techniques for addressing small data challenges

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Drumond et al. (2023)	1. Human 3.6 M dataset	The experiment is performed with eleven actions in meta-training and four actions in meta-testing The task is to forecast the successive ten frames (400 ms) given the previous 50 frames (2 s) across the given set of sensors	Few-shot motion prediction model incorporating the underlying network structure The study uses heterogeneous sensors	The study shows a considerable performance increase overall relevant baselines, with performance lifts ranging from 10.4 to 39.3%	End users can contribute additional movements by showing an activity a few times before the model can reliably categorise and forecast future frames
Sun et al. (2021)	1. Caltech-UCSD Birds dataset (Chahal et al. 2021) 2. mini DogsNet 3. mini PPlankton	11,788 images categorized into 1200 fine-grained classes 2. The study selects ten classes randomly to form the training data The dataset (3) consists of 20 classes with 70 samples. Ten classes are used for training, and the rest are used for novel classes to evaluate the model	FSL Exploration of features using a feature fusion model focusing on critical regions Proposes a Center Neighbor (CN) Loss function to form robust feature space distributions for synthesizing discriminative features	SVM and Cosine classifiers achieve outstanding performance On miniDogs dataset the proposed model improves accuracy by 2.96% on 1-shot learning and 4.24% on 5-shot learning. A slight accuracy increase of around 1% on the (3) dataset	The paper presents the Feature Fusion Model and CN Loss for feature extraction on such challenging tasks

Table 5 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Medela et al. (2019)	1. High-grade primary tumour dataset collected from University Medical Center Mannheim 2. Biopool Colon, Breast and Lung Hematoxylin–Eosin dataset (B-CBL-HE)	In dataset (1), 5000 image tiles are distributed among eight classes and 625 images per sample The study uses the subset of the original data from the Basque Biobank (BIOEF—Spain) collected at five local hospitals of the Basque Public Health system-Osakidetza The subset of this dataset contains 1755 samples of images (healthy, low-grade and high-grade tumours)	The Siamese network learns features and distance representations from tumoural tissue and transmits that information to SVM for classification	With only 60 training samples, the model reached a balanced accuracy (BAC) of 90% The model even achieved excellent results for the Lung and Breast tissues absent in the training set	The proposed model performs better than the Finetune TL method that obtained 73% BAC with the exact sample count. It needs 600 samples to achieve 81% BAC
Iwata and Kumagai (2020)	UCR time series classification archive for obtaining 90-ttime-series datasets	The effectiveness of the model is shown by using 90 time-series datasets The values were obtained at the first 100-time steps for each time series The data is divided into 55 training, ten validation and 25 target tasks, each with 50-time series	An FSL approach that predicts a future value of a time series in a target task with limited time series data in the target task	The presented approach achieved comparable performance to the best approach in 62 among 90 target tasks, the most among comparing methods	Generally, LSTM outperformed NN, and NN outperformed Linear. The study reveals that LSTM-based RNN are suitable for forecasting time series

Table 5 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Triantafillou et al. (2017)	1. Omniglot 2. mini-ImageNet 3. CUB	The (1) dataset contains 1623 characters of 50 different alphabets From the (1) dataset, 100 classes were selected, out of which 64 were selected for testing. The dataset has 60,000 colour images	FSL for information retrieval	Overall, the model demonstrates a comparable performance with the cutting-edge results shown by the classification benchmarks while having the capability of few-shot retrieval, achieving better results over a robust baseline	It is observed that when there are few examples per class, the proposed model has an advantage over the all-pairs of Siamese
Feng and Duarte (2019)	1. Opportunity activity recognition dataset (OPP) 2. PAMAP2 physical activity monitoring dataset	In (1) dataset, a total of 202 samples per class and in dataset (2) 129 samples per class. In (3), all 9 classes are distributed in 3 classes to increase the dataset size	Human activity recognition with the help of FSL approach for wearables A framework alleviates negative transfer	With only few samples per class satisfying results are achieved utilizing the proposed framework	Increasing the size of training samples from 1 to 5 for each target class improves the performance from 10 to 15%

Table 5 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Chen et al. (2020a)	1. MRI dataset from Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset 2. The unpaired CT scans were collected from the CQ500 dataset (Jiang et al. 2019)	The dataset consists of eight pairs of MRI-CT scans, 50 unpaired MRI scans and 50 unpaired CT scans	One-shot GAN approach to solve the short-paired training data task A semi-supervised synthesis sub-network to learn the cross-modality mappings between MRI and CT One-shot learning neighbour anchoring approach for effectively reducing the space of alternative translation mappings for unpaired images	The proposed model shows an improvement of (85.66 ± 3.26) on DSC and an Average Symmetric (1.04± 0.19)	To address the scarcity of the data, the study makes full use of unpaired data, which are typically abundant along with single-paired MRI-CT data
Feyjie et al. (2020)	1. FSS-1000 dataset 2. The ISIC 2018 dataset is provided by the International 2018 Skin Imaging Collaboration Grand Challenge 3. PH^2 dataset	The dataset (1) consists of 1000 classes containing ten images The dataset (2) contains 2594 demographic images The PH^2 dataset contains a total of 200 RGB dermoscopic images of melanocytic lesions	A new FSL technique for semantic segmentation, where unlabelled samples are also available The study suggests integrating surrogate activities that use powerful supervisory signals produced from data to learn semantic features	The proposed technique improves from the FSL baseline by a margin of 6–7% With (3) dataset, the difference is around 15%	The results show that the performance gain is more significant when fewer labelled samples, such as 1-shot versus 5-shot

5.5 Loss function, regularisation, and architecture-based methods

The one consistent finding in the present DL discourse: categorical cross-entropy loss following softmax activation is the preferred technique for classification. One study Barz and Denzler (2020) revealed that the cosine loss function performs significantly better than cross-entropy on datasets with only a few samples per class. The authors demonstrated a 30% gain in accuracy without pre-training on the CUB 200-2011 dataset compared to the cross-entropy loss.

The Orthogonal Softmax Layer (OSL), which keeps the weight vectors in the classification layer orthogonal during both the training and test phases, is proposed as a solution in Li et al. (2020). The suggested OSL shows superior performance compared to the techniques used for comparison on four benchmark datasets with small samples, and experimental findings also suggest that it applies to datasets with large samples. Table 6 provides a summary of techniques, including loss function, DL architecture, and regularisation-based techniques, for solving small data challenges in the literature.

Table 7 shows the comparative study of recent DL approaches applied to various small datasets. Each reviewed paper is categorised based on three characteristics: reference abbreviated as Ref.; the approach employed to address small dataset problems abbreviated as TDA, ODA, GAN, TL, FSL, L&A, and OT; and the evaluation metric utilised. The complete form of these abbreviations is given in Table 7. Table 6 reveals that TDA was the least used technique, while TL and FSL were the most frequently employed techniques, highlighting their popularity and effectiveness for small dataset scenarios. A substantial number of studies have explored other methods, such as cosine loss function, changes in network architecture like adding orthogonal softmax layer, and regularisation-based variations, also demonstrating that a good number of papers used these methods to tackle small dataset challenges. The primary evaluation metric employed across the studies is classification accuracy, emphasising its importance as a measure of model effectiveness.

In Table 8, the dataset size is distributed into five categories, namely D1, D2, D3, D4, and D5, representing different ranges of dataset sizes based on the number of samples per class in the training dataset. D1 includes datasets with less than 100 samples per class. D2 and D3 encompass datasets with 101 to 1000 and 1001 to 3000 samples per class, respectively. D4 includes datasets with 3000 to 10,000 samples per class, and D5 represent datasets with 10,001 or more samples per class. The distribution summary depicted in Fig. 7 shows that 33% of the studies comprise D1 datasets, indicating datasets with less than 100 samples per class. The next most common category is D2, accounting for 29% of the studies, representing datasets with 101 to 1000 samples per class. D3, which includes datasets with 1001 to 3000 samples per class, contributes to 18% of the studies. D4 and D5, representing bigger datasets, comprise 10% of the studies. Table 9 provides additional information, such as references (cited as "Ref."), the NDS column denoting the number of datasets studied, the technique employed to address small dataset problems, and an additional method apart from DA, TL, GAN, FSL, L&A. It is worth noting that the datasets marked with (*) in the NDS column indicate imbalanced datasets.

Table 6 Overview of loss function, architectural changes, and regularisation-based techniques for addressing small data challenges in literature

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Barz and Denzler (2020)	1. CUB 2. NAB 3. Stanford Cars 4. Oxford 5. Flowers 6. MIT Indoor Scenes 7. CIFAR 100	The (1) dataset has 200 classes with 29–30 samples per class. The (2) dataset has 555 classes with 4–60 samples per class. The (3), (4), (5), and (6) datasets have 24–68, 20, 77–83 and 500 samples per class respectively	Using Cosine Loss function without Pre-Training	30% and 21% improvements were observed on (1) and (2) datasets. On (3) and (5), the improvement was 8% and 6%, respectively. Cross entropy and softmax performed equally on a big dataset with 500 samples per class	Classification accuracy obtained with cosine loss shows considerably better performance over cross-entropy after softmax on small datasets
Li et al. (2020)	1. UIUC-Sports dataset (UIUC) (Qin et al. 2020) 2. 15 Scenes 3. Subset of the Scenes dataset on AI Challenger (80-AI) 4. Caltech101	The (1) dataset has 1579 images in eight classes, the dataset has less than 200 samples per class In the (2) dataset, there are 200 samples per class In (3), there are also 200 images per class In the (4) dataset, there are a total of 40–800 images per class	OSL-NET (Orthogonal SoftMax layer)	It achieves more accuracy with a higher mean and lower variation than the comparative baseline	It is more appropriate for thin and shallow networks than a fully connected network It also shows excellent performance on big datasets

Table 6 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Chen et al. (2018)	1. CT scans collected from two local hospitals (Permission from the Imperial College Joint Research Office) 2. BraTS 2017	A total of 781 2-D images randomly selected from 101 subjects in the (1) dataset with 500 for training In (2), 50 subjects from 285 are chosen for training	A new CNN architecture called Dense-Res-Inception Net (DRINet)	The architecture proposed shows improved U-Net performance in three challenging tasks: multi-class cerebrospinal fluid (CSF) segmentation on brain CT images, multi-organ segmentation on abdominal CT images, and multi-class brain tumour segmentation on MR images	Due to its three strong blocks, DRINet provides higher segmentation results than U-Net in terms of DSC, sensitivity, and Hausdorff distances
Gao et al. (2022)	1. CIFAR-10 2. SVNH 3. STL-10 4. MNIST	In (1) dataset there are 50,000 training images. In (2) dataset there are 73257 training images In (3), only 5000 training images categorized into ten different classes The (4) dataset has 60,000 training images	The paper introduces PatchShuffle "a new regularization method for generalizable CNN training" and compares CNN models trained with and without PatchShuffle	PatchShuffle outperforms standard Backpropagation methods. The error rate is reduced by approx. 4% with 9000 CIFAR-10 training samples. On the whole CIFAR-10 dataset, the error rate is reduced by 0.67% The test error rate improves by approx. 1% on the SVNH dataset. The test error rate also improves in the case of STL-10 and MNIST	Under rising pollution levels, both normal BP and PatchShuffle function poorly

Table 6 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Zhang et al. (2020)	Subset of following Datasets (1) Extended Yale B Database (2) CMU PIE Database (3) AR Database (4) FRGC Database	In (1) and (2) dataset, 2414, 41,368 face images. 15, 20, 25, and 30 samples per individual are randomly chosen for training The (3) dataset contains 2600 images from 50 male and 50 female subjects. 8, 11, 14, and 17 samples per class for training are chosen The (4) dataset has 220 persons and 20 images per person. Only Ten samples per person are used for training	An end-to-end deep cascade model (DCM) based on Sparse Representation based Classification (SRC), Nuclear-norm Matrix Regression (NMR), and (DL) with hierarchical learning, non-linear transformation and multi-layer structure for corrupted face recognition	The recognition rate of DCM is 4% higher than that of RSC, which has the second-best rank among the compared methods The proposed model shows higher accuracy on the (2) dataset On the (3) dataset, the proposed model showed the highest performance with 14 training samples. However, the increase was minimal compared to the second-best algorithm	DCM based on SRC DCM(S) outperforms DCM based on NMR DCM(N) in recognition when there is no corrupt test data However, when the test data is corrupt, DCM(N) outperforms DCM(S)
Keshari et al. (2020)	1. CIFAR 10 2. CIFAR 100 3. SVNH 4. TinyImageNet	There are 50,000 training images in the (1) and (2) datasets, 73,257 in (3) and 100k in the (4) dataset	The Dynamic Attention Pooling (DAP) method aims to retrieve global knowledge from the feature map's most discriminative sub-part	The proposed ResNet-based DAP shows an improvement of 1.75%, 0.47%, and 1.87% on (1), (2), and (4) datasets, respectively	The model compares its results with the state-of-the-art ResNet model and considers all four datasets as small
Rahadian and Yusuf (2023)	The dataset collected from four college students consists of different mood images	The dataset has a total of 480 images	Simplified AlexNet	The classifier showed an accuracy of 78.6%	The model's accuracy increased to 90% after the DA

Table 6 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Zheng et al. (2022)	1 Cora 2. CiteSeer 3. PubMed 4. ACM 5. BlogCatalog 6. Flickr	Dataset (1) has 2708 nodes and 5429 edges (2) has 3327 nodes and 4732 edges. There are 19717 nodes and 44338 edged in (3) dataset. Dataset (4) has 16484 nodes and 71980 edges. (5) has 5175 nodes and 171743 edges, and (6) dataset consists of 7575 nodes and 239738 edges	ANomaly dEtection framework with Multi-scale cONtrastive lEarning (ANEMONE) is a broad framework based on contrastive learning for graph anomaly detection	Comprehensive trials with ANEMONE and its variation ANEMONE-FS in totally unsupervised and few-shot anomaly detection situations show that ANEMONE and its variant ANEMONE-FS consistently beat state-of-the-art approaches on six benchmark datasets	The method enhances the flexibility of contrastive learning for anomaly detection and facilitates broader applications
Chen et al. (2020b)	The dataset is collected from 2013 to 2018 from a Shanghai hospital	The dataset has three classes with a total of 340 patient data; there are 112 Alzheimer's disease patients, 145 Mild cognitive impairment patients and 83 NC patients	Multi-modal feature fusion. The network fuses the features extracted at the vector level. KNN attention pooling layer and CNN are employed to classify small samples	Accuracy and F1-score based on multi-modality are increased by more than 10% compared to the single modality	Under the same multimodal data in a small sample learning method, the accuracy and F1-score improved by 8.2% and 8.4%, respectively

Table 6 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Wibowo et al. (2022)	2017 MICCAI Medical Image Computing and Computer-Assisted Intervention dataset	The dataset consists of slices examined from 150 patients and distributed among five classes, each composed of the data of 30 patients	The fully convolutional EfficientNetB5-UNet is enhanced to perform the MRI slice semantic segmentation in the encoder-decoder network. A two-dimensional thickness algorithm combines the segmented output for the 2D feature map of images	The segmentation performance is above 86.92% for all five classes, and the classification accuracy is 92%	The FSL approach is employed for classification to handle limited data
Zhu et al. (2022)	Rail vehicle bogie performance test bench	The training data has only 4.31% of the total dataset size, i.e. 12 samples, and the testing data has 220 samples	Based on the multi-information fusion technique, an unsupervised representation alignment deep network (URADQN) is proposed to tackle the issues of generalization and overfitting	The algorithm obtained an average fault diagnostic task accuracy of up to 98.44%	The model shows a better performance for small dataset of fault diagnosis

Table 6 (continued)

References	Dataset	Dataset size	Approach for solving small dataset problem	Effect on the accuracy or error rate	Remarks
Gao et al. (2022)	1. Houston 2013 (HS13), 2. Botswana (BO), 3. Kennedy Space Center (KSC), 4. Chikusei (CH), 5. University of Pavia (UP), 6. Pavia Center (PC), 7. Salinas (SA), 8. Indian Pines (IP)	In (1) dataset 15029 samples are present in 15 classes. The (2) dataset has 3248 samples and 14 classes, (3) has 13 classes and 5211 samples. The (4) dataset has 77592 samples in 19 classes. (5) has 42776 samples in 9 classes. The datasets (6), (7) and (8) have 148152, 54129 and 10249 samples in 9,16 and 16 classes, respectively	The unsupervised meta-learning approach with multiview constraints for hyperspectral image (HSI) small sample set classification	The experiment demonstrates that the proposed method outperforms state-of-art supervised meta-learning methods and other advanced classification models in small sample tasks	Overall, the classification performance of ML models is inferior to that of DL models The DL models perform better by extracting more discriminative features
Suzuki (2022)	1. Low-dose CT images with lung nodules	The (1) dataset comprises 38 low-dose thoracic helical CT (LDCT) scans obtained from 31 patients	MTANN for lung nodule detection CT	MTANN outperforms the best-performing CNN MTANN generates 2.37 (False Positives) FPs per patient at 100% sensitivity, which is much lower than the best-performing CNN, which achieves 32.50 FPs per patient	The research explores many cases of small-data DL in CT lung cancer diagnosis

Table 7 Comparative assessment of deep learning models for small datasets

References	Approach for solving small dataset problems							Metrics							
	TDA	ODA	GAN	TL	FSL	L&A	OT	Acc	F1	ME	ER	CE	DSC	S	Sp
Krizhevsky et al. (2017)	✓										✓				
Shijie et al. (2017)	✓		✓					✓							
Chatfield et al. (2014)		✓						✓							
Inoue (2018)		✓													
Zhang et al. (2021a)		✓													
Karras et al. (2020)		✓													
Marchesi (2017)			✓												
Qin et al. (2020)			✓					✓						✓	✓
Bowles et al. (2018)			✓										✓		
Antoniou et al. (2017)			✓					✓						✓	✓
Frid-Adar et al. (2018)			✓												
Song et al. (2022), Ma et al. (2022)			✓					✓							
Bargshady et al. (2022)			✓	✓				✓			✓				
Goceri (2021)				✓				✓							
Huh et al. (2016)				✓				✓							
Zhao et al. (2022)				✓				✓							
Zhou et al. (2018), Neyshabur et al. (2020)				✓											
Neyshabur et al. (2020)				✓				✓							
Kornblith et al. (2019)				✓				✓							
Azizi et al. (2021)				✓				✓							
Ng et al. (2015)				✓				✓							
Tajbakhsh et al. (2016)				✓				✓							
Ma et al. (2022)				✓							✓				
Chien et al. (2022)				✓				✓	✓	✓		✓			
Drumond et al. (2023)					✓			✓							

Table 7 (continued)

References	Approach for solving small dataset problems							Metrics							
	TDA	ODA	GAN	TL	FSL	L&A	OT	Acc	F1	ME	ER	CE	DSC	S	Sp
Sun et al. (2021)					✓		✓	✓							
Medela et al. (2019)					✓		✓	✓							
Iwata and Kumagai (2020)					✓			✓							
Triantafillou et al. (2017)					✓			✓							
Feng and Duarte (2019)					✓		✓	✓							
Chen et al. (2020a)			✓		✓					✓			✓		
Feyjie et al. (2020)					✓								✓		
Barz and Denzler (2020)						✓						✓			
Li et al. (2020)						✓		✓							
Chen et al. (2018)							✓						✓	✓	
Kang et al. (2017)							✓				✓				
Zhang et al. (2020)							✓	✓							
Keshari et al. (2020)							✓	✓							
Rahadian and Yusuf (2023)							✓	✓							
Chen et al. (2020b)							✓	✓	✓						
Iwata and Kumagai (2020)							✓	✓							
Zhu et al. (2022)							✓	✓							
Gao et al. (2022)				✓			✓	✓						✓	

Table 8 Size distribution of small datasets and the wide range of methods used in recent research articles

References	NDS	D1	D2	D3	D4	D5	Technique	Additional methods
Krizhevsky et al. (2017)	1		✓				TDA	
Shijie et al. (2017)	2	✓	✓		✓		TDA + GAN	
Chatfield et al. (2014)	4	✓	✓	✓			TDA	
Inoue (2018)	4	✓	✓			✓	DA	SP
Zhang et al. (2021a)	3	✓					TL + DA	DADA
Karras et al. (2020)	4	✓		✓	✓		DA	ADA
Marchesi (2017)	2			✓			GAN	
Qin et al. (2020)	1*		✓	✓			GAN	
Bowles et al. (2018)	1	✓					GAN	
Antoniou et al. (2017)	3	✓					GAN	DAGAN
Frid-Adar et al. (2018)	1	✓					GAN	
Song et al. (2022), Ma et al. (2022)	1*			✓			TL + GAN	DTAM
Bargshady et al. (2022)	2			✓	✓		TL + GAN	
Goceri (2021)	3		✓				TL	
Huh et al. (2016)	8		✓	✓		✓	TL	
Zhao et al. (2022)	1	✓					TL	
Deng et al. (2010), Zhou et al. (2018), Neyshabur et al. (2020), Kornblith et al. (2019), Azizi et al. (2021), Zoph et al. (2020b), Li and Deng (2022), Revina and Emmanuel (2021), Siuly and Zhang (2016), Chen et al. (2018), Singha et al. (2021), Anwar et al. (2018), Xu et al. (2021), Afshar et al. (2019), Ng et al. (2015), Mazurowski et al. (2019), Tajbakhsh et al. (2016), Keshari et al. (2020), He et al. (2019), Zoph et al. (2020a), Barbero-Aparicio et al. (2024), Zhang et al. (2023), Tan et al. (2024), Tran et al. (2019), Bao et al. (2019)	3		✓	✓			TL	
Neyshabur et al. (2020)	2				✓		TL	
Kornblith et al. (2019)	11	✓			✓		TL	
Azizi et al. (2021)	2	✓				✓	TL	MICL

Table 8 (continued)

References	NDS	D1	D2	D3	D4	D5	Technique	Additional methods
Ng et al. (2015)	3			✓			TL	
Tajbakhsh et al. (2016)	1					✓	TL	
Ma et al. (2022)	2		✓				TL	
Chien et al. (2022)	1*		✓	✓			TL	
Drumond et al. (2023)	1	✓					FSL	
Sun et al. (2021)	3	✓					FSL	CNL
Medela et al. (2019)	2		✓				FSL (SN)	
Iwata and Kumagai (2020)	1	✓					FSL	
Triantafillou et al. (2017)	3	✓	✓				FSL	
Tan et al. (2024)	2		✓				FSL	
Chen et al. (2020a)	2	✓					FSL	OS-GAN
Feyjie et al. (2020)	3	✓	✓	✓			FSL	
Barz and Denzler (2020)	7	✓	✓				CLF	
Li et al. (2020)	4		✓				OSL	
Chen et al. (2018)	2	✓	✓				TL	DRINET
Kang et al. (2017)	4		✓		✓		RM	PS
Zhang et al. (2020)	4	✓					DCM	SRC, NMR
Keshari et al. (2020)	4				✓	✓		DAP
Li et al. (2020)	1		✓				TL	SA
Zheng et al. (2022)	6			✓		✓	CL	ANEMONE
Chen et al. (2020b)	1*	✓	✓				MMFU	
Wibowo et al. (2022)	1	✓					FSL	
Feng and Duarte (2019)	1	✓					MIFT	URADQN
Gao et al. (2022)	8		✓	✓		✓	UML	
Suzuki (2022)	1	✓					MTANN	

6 Open issues and future research directions

6.1 Open issues

A study conducted by Gartner, Inc.[7] reported that 70% of enterprises will shift their focus from big datasets to small and wide datasets by 2025, giving more context for analytics and making AI less data-hungry. Nevertheless, there are several open issues in DL with small datasets. Subsequently, some of the major outstanding problems in DL with limited datasets are discussed.

6.1.1 Poor generalisation with small datasets

The theoreticians of ML have focused on the Independent and Identical Distribution (IID) assumptions, which state that the test cases are likely drawn from the same distribution as the training samples. Unfortunately, in the actual world, this is not a reasonable assumption. As a result, the performance of today's state-of-the-art AI systems suffers when they transition from the controlled laboratory to the field.

Our goal is to improve the model's robustness when challenged with variations in sample distribution. Generalisation refers to the model's capacity to perform effectively on unknown data after being trained on a small dataset. One major reason this impacts generalisation is overfitting, in which the model gets overly specialised to the training data and fails to generalise successfully to new cases. Overfitting arises when the model is excessively complex for the available data, causing it to memorise the training instances instead of learning generalisable patterns. Overfitting also arises when the training data do not accurately represent the population the model is meant to serve, resulting in biased predictions.

Recent research assists us in understanding how different DL architectures perform in terms of systematic generalisation capability. How can we develop future ML systems with improved generalisation capabilities and adapt faster in scenarios where the data is out-of-distribution? Some studies that discuss generalisation in DL in detail (Kawaguchi et al. 2022; Bousquet and Elisseeff 2002; Zhang et al. 2021b; Olson et al. 2018; Power et al. 2022; Caro, et al. 2022; Chatterjee and Zielinski 2022).

Enabling higher-level cognition in deep learning models.

DL models frequently lack the ability to reason like humans do. Researchers are creating new strategies that allow DL models to reason and learn from small datasets in order to enable higher-level cognition. Incorporating symbolic reasoning into DL models is one interesting method. The capacity to handle abstract symbols and utilise logical principles to execute tasks has long been a characteristic of human cognition. Researchers hope to give DL models the ability to reason about relationships and abstract concepts and generalise to new tasks and domains. For instance, it has demonstrated that DL models that include symbolic reasoning perform well on tasks like answering visual questions and exercising common sense (Storrs and Kriegeskorte 2019a; Perconti and Plebe 2020; Goyal and Bengio 2022).

[7] Gartner Says 70% of Organizations Will Shift Their Focus From Big to Small and Wide Data By 2025.

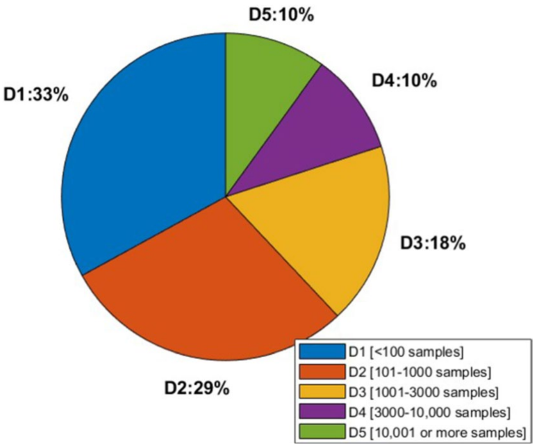

Fig. 7 The percentage distribution of studies conducted based on different dataset sizes

Incorporating previous information into DL models is another strategy for allowing higher-level cognition. This may entail integrating information from subject-matter specialists or from different sources, including language models.

Overall, allowing higher-level cognition in DL models is an active research topic, and there is still more to be done to allow DL models to reason and generalise in a way that is more akin to human cognition (Goyal and Bengio 2022; Storrs and Kriegeskorte 2019b; Battleday et al. 2021).

6.1.2 Robustness

DL models are sometimes vulnerable to changes in the input data, including adversarial attacks. Even a small number of adversarial cases can greatly impact the model's performance in the case of small datasets, which is particularly problematic (Battleday et al. 2021; Qian et al. 2022; Allen-Zhu and Li 2022; Shaukat et al. 2022). Further study is required to determine how to strengthen DL models, particularly for limited datasets.

6.1.3 Unsupervised learning

Unsupervised learning is one fundamental solution beyond supervised, data-hungry DL versions. DL and unsupervised learning are not in logical opposition. DL is generally utilised in a supervised scenario with labelled data, but there are applications of the unsupervised version where excellent results can be obtained by employing DL techniques. There are certainly reasons in many sectors to shift away from the huge data needs that supervised DL typically necessitates.

Unsupervised learning seeks to build usable data representations without explicit labelling or supervision. Autoencoders, variational autoencoders, and generative adversarial networks are DL algorithms that have shown remarkable strides in unsupervised learning tasks, including clustering, anomaly detection, and dimensionality reduction. Nevertheless, employing DL for unsupervised learning still has a lot of obstacles and unanswered problems. Developing efficient training algorithms, creating appropriate architecture, and understanding the theoretical foundations of deep unsupervised learning are some of these difficulties (Agarwal et al. 2022; Akcakaya et al. 2022; Tao et al. 2022).

Table 9 Abbreviations used in comparison tables

Abbreviation	Full form
Acc	Accuracy
ADA	Adaptive Discriminative DA
ANEMONE	Anomaly Detection Framework with Multi-scale Contrastive Learning
CE	Cross Entropy (Binary Crossentropy and categorical crossentropy)
CFL	Cosine Loss function
CL	Contrastive Learning
CNL	Center Neighbour Loss
DADA	Deep Adversarial Data Augmentation
DAGAN	Data Augmentation Generative Adversarial Networks
DAP	The Dynamic Attention Pooling
DRINET	Dense-Res-Inception Net
DSC	Dice Score
DTAM	Domain Transfer using Adversarial learning and Metric learning
ER	Error rate
F1	F1-Score
FSL	Few-Shot Learning
GAN	Generative adversarial Networks
L&A	Loss function and Architecture based methods
ME	Mean Square Error
MICL	Multi-instance Contrastive Learning
MIFT	Multi-information Fusion Technique
MMFU	Multi modal Feature Fusion
MTANN	Massive-training artificial neural network
NDS	Number of Datasets
NMR	Nuclear-norm Matrix Regression
ODA	Other Data Augmentations
OS-GAN	One-Shot Generative Adversarial Networks
OSL	Orthogonal softmax layer
OT	Other
PS	PatchShuffle
RM	Regularization Method
S	Sensitivity
SA	Simplified AlexNet
SP	SamplePairing
Sp	Specificity
SRC	Sparse Representation-based Classification
TDA	Traditional Data Augmentation
TL	Transfer Learning
UML	Unsupervised meta Learning
URADQN	Unsupervised Representation Alignment deep network

6.1.4 Data diversity problem

The data diversity problem is still open in DL models, especially with small datasets. It can be minimised by using techniques like DA, TL, regularisation, and ensemble learning; although, these methods may not always work or be feasible in all circumstances. The data diversity issue might also get worse as DL models get more complicated and there are more data to choose from. This is because complex models need more varied instances to acquire robust input representations since they are more prone to overfitting the training data.

Another problem is that defining or quantifying the data diversity issue is not always straightforward. It might be challenging to distinguish between bias, noise, labelling mistakes, or a lack of variety in the training data as the cause of a model's poor performance on fresh samples. As a result, considerable studies are still being done in the domain of data diversity in DL, and as the subject develops, new methods and techniques are expected to be developed.

6.1.5 How small is actually "small" in deep learning models?

It is uncertain what minimum amount of data is required for successful DL models to function well. The concept of small datasets differs for specific applications and model architectures.

6.1.6 Effective data augmentation and regularisation

DA techniques, such as rotation, scaling, and cropping, can help generate additional training data, but it is unclear which augmentation technique is the most effective and how the model balances augmentation with overfitting. While regularisation approaches reduce overfitting, it is difficult to ascertain which strategies are most effective on small datasets. Some techniques like weight decay and dropout are extensively used in the literature.

6.2 Future research directions

Some primary future research directions other than the ones discussed in the study articles, such as Zhang et al., ul Sabha et al. (2024), are mentioned as follows:

- Future research on the above-mentioned small dataset techniques should extensively validate these small dataset techniques on real-world applications. Applying these techniques to diverse fields such as healthcare, agriculture, and finance can demonstrate their practicality and lead to the development of tailored solutions for specific domains.
- Over the past decade, DL models, particularly deep neural networks, have seen substantial advancements. In many practical applications, the model architecture has matured to a point where it can be considered a solved problem. Consequently, it is now more beneficial to maintain a fixed neural network architecture and direct research efforts towards enhancing the quality and quantity of data. Future research should, therefore, prioritize data-centric AI, focusing on innovative ways to improve data to boost model performance.
- Another promising future research direction is the targeted use of GAN-based data synthesis for problem-solving. For example, consider a model trained on a dataset with five

classes, where it performs well on four classes but poorly on one. Instead of enhancing the overall model or dataset, we can specifically improve the data for that underperforming class using GAN-based data augmentation. This targeted approach to addressing specific weaknesses can be a valuable strategy for future research, allowing for more precise and effective improvements.

- Another future research direction could be the development of a tool designed to identify the most beneficial subset of a big dataset for model training. This tool would select a small, representative dataset that maximizes training efficiency. Additionally, it could pinpoint specific areas where data augmentation is needed, thereby reducing the effort and resources required to collect additional data across the board. Instead of gathering more data indiscriminately, this targeted approach would focus on augmenting data for only those classes that truly need it, streamlining the data collection process.
- A significant research direction in the area of small datasets involves developing customized loss functions and model architectures specifically suited for limited data scenarios. Future research could explore adaptive loss functions that dynamically adjust based on data characteristics and the learning stage. Additionally, lightweight model architectures requiring fewer parameters should be investigated to enhance robustness and generalisation when working with small datasets. This approach aims to optimize model performance and efficiency, making DL more effective in data-constrained environments.

7 Conclusion

This study comprehensively analyses the advancements in DL models trained on small datasets. The state-of-the-art techniques used in this area were thoroughly reviewed, illustrating their advantages and disadvantages. The PRISMA model search was performed to identify 165 relevant studies, which were subsequently analysed based on various attributes, such as publisher, country, utilisation of small dataset technique, dataset size, and performance. A comparative analysis of different small dataset techniques using different metrics was then conducted. According to our findings, several critical paths for future research in DL on small datasets were identified. Overall, this publication is anticipated to be a helpful resource for academicians and industry professionals interested in this area and inspire more studies to address the challenges of DL with small datasets. Besides the limitations caused by the lack of data, there is significant interest in investigating cutting-edge methods to enhance the effectiveness of DL models.

Author contributions Ishfaq Hussain Rather: conceptualization, formal analysis, methodology, validation, writing—review and editing. Sushil Kumar: investigation, supervision, validation, writing—review and editing. Amir H. Gandomi: investigation, validation, writing—review and editing.

Funding Open access funding provided by Óbuda University.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

Afshar P, Mohammadi A, Plataniotis KN, Oikonomou A, Benali H (2019) From handcrafted to deep-learning-based cancer radiomics: challenges and opportunities. IEEE Signal Process Mag 36(4):132–160. https://doi.org/10.1109/MSP.2019.2900993

Agarwal P, Aghaee M, Tamer M, Budman H (2022) A novel unsupervised approach for batch process monitoring using deep learning. Comput Chem Eng 159:107694. https://doi.org/10.1016/J.COMPCHEMENG.2022.107694

Ahmad Z, ul Abidin Jaffri Z, Chen M, Bao S (2024) Understanding GANs: fundamentals, variants, training challenges, applications, and open problems. Multimed Tools Appl. https://doi.org/10.1007/S11042-024-19361-Y

Ahmed SF et al (2023) Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artif Intell Rev 2023:1–97. https://doi.org/10.1007/S10462-023-10466-8

Akcakaya M, Yaman B, Chung H, Ye JC (2022) Unsupervised deep learning methods for biological image reconstruction and enhancement: an overview from a signal processing perspective. IEEE Signal Process Mag 39(2):28–44. https://doi.org/10.1109/MSP.2021.3119273

Allen-Zhu Z, Li Y (2022) Feature purification: how adversarial training performs robust deep learning. In: Proceedings—annual IEEE symposium on foundations of computer science, FOCS, vol 2022-February. pp 977–988. https://doi.org/10.1109/FOCS52979.2021.00098

Antoniou A, Storkey A, Edwards H (2017) Data augmentation generative adversarial networks. http://arxiv.org/abs/1711.04340

Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK (2018) Medical image analysis using convolutional neural networks: a review. J Med Syst. https://doi.org/10.1007/s10916-018-1088-1

Azizi S et al (2021) Big self-supervised models advance medical image classification. In: Proceedings of the IEEE international conference on computer vision. pp 3458–3468. https://doi.org/10.48550/arxiv.2101.05224

Bagherinezhad H, Horton M, Rastegari M, Farhadi A (2018) Label refinery: improving ImageNet classification through label progression. https://doi.org/10.48550/arxiv.1805.02641

Bansal T, Jha R, McCallum A (2019) Learning to few-shot learn across diverse natural language classification tasks. pp 5108–5123. https://doi.org/10.48550/arxiv.1911.03863

Bansal A, Sharma R, Kathuria M (2022) A systematic review on data scarcity problem in deep learning: solution and applications. ACM Comput Surv (CSUR) 54(10s):1–29. https://doi.org/10.1145/3502287

Bao Y, Li Y, Huang SL, Zhang L, Zheng L, Zamir A, Guibas L (2019) An information-theoretic approach to transferability in task transfer learning. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 2309–2313

Barbero-Aparicio JA, Olivares-Gil A, Rodríguez JJ, García-Osorio C, Díez-Pastor JF (2024) Addressing data scarcity in protein fitness landscape analysis: a study on semi-supervised and deep transfer learning techniques. Inf Fusion 102:102035. https://doi.org/10.1016/J.INFFUS.2023.102035

Bargshady G, Zhou X, Barua PD, Gururajan R, Li Y, Acharya UR (2022) Application of CycleGAN and transfer learning techniques for automated detection of COVID-19 using X-ray images. Pattern Recognit Lett 153:67–74. https://doi.org/10.1016/J.PATREC.2021.11.020

Barz B, Denzler J (2020) Deep learning on small datasets without pre-training using cosine loss. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp 1371–1380

Battleday RM, Peterson JC, Griffiths TL (2021) From convolutional neural networks to models of higher-level cognition (and back again). Ann N Y Acad Sci 1505(1):55–78. https://doi.org/10.1111/NYAS.14593

Bertinetto L, Henriques JF, Valmadre J, Torr P, Vedaldi A (2016) Learning feed-forward one-shot learners. In: Advances in neural information processing systems, vol 29

Bousquet O, Elisseeff A (2002) Stability and generalization. J Mach Learn Res 2(3):499–526. https://doi.org/10.1162/153244302760200704

Bowles C, Gunn R, Hammers A, Rueckert D (2018) GANsfer learning: combining labelled and unlabelled data for GAN based data augmentation. http://arxiv.org/abs/1811.10669

Cai A, Hu W, Zheng J (2020) Few-shot learning for medical image classification. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 12396 LNCS. pp 441–452. https://doi.org/10.1007/978-3-030-61609-0_35/COVER

Caro MC et al (2022) Generalization in quantum machine learning from few training data. Nat Commun 13(1):1–11. https://doi.org/10.1038/s41467-022-32550-3

Chahal H, Toner H, Rahkovsky I (2021) Small data's big AI potential. Center for Security and Emerging Technology. https://doi.org/10.51593/20200075

Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details: delving deep into convolutional nets. In: BMVC 2014—proceedings of the British machine vision conference 2014. https://doi.org/10.48550/arxiv.1405.3531

Chatterjee S, Zielinski P (2022) On the generalization mystery in deep learning. arXiv Preprint. https://arxiv.org/abs/2203.10036

Chen XW, Lin X (2014) Big data deep learning: challenges and perspectives. IEEE Access 2:514–525. https://doi.org/10.1109/ACCESS.2014.2325029

Chen K, Wang S (2011) Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions. IEEE Trans Pattern Anal Mach Intell 33(1):129–143. https://doi.org/10.1109/TPAMI.2010.92

Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D (2018) DRINet for medical image segmentation. IEEE Trans Med Imaging 37(11):2453–2462. https://doi.org/10.1109/TMI.2018.2835303

Chen X et al (2020a) One-shot generative adversarial learning for MRI segmentation of craniomaxillofacial bony structures. IEEE Trans Med Imaging 39(3):787–796. https://doi.org/10.1109/TMI.2019.2935409

Chen DH, Zhang L, Ma C (2020b) A multimodal diagnosis predictive model of Alzheimer's disease with few-shot learning. In: Proceedings—2020 international conference on public health and data science (ICPHDS 2020). pp 273–277. https://doi.org/10.1109/ICPHDS51617.2020.00060

Chen S, Cao Y, Kang Y, Li P, Sun B (2021) Deep feature representation based imitation learning for autonomous helicopter aerobatics. IEEE Trans Artif Intell 2(5):437–446. https://doi.org/10.1109/TAI.2021.3053511

Chien CF, Hung WT, Liao ETY (2022) Redefining monitoring rules for intelligent fault detection and classification via CNN transfer learning for smart manufacturing. IEEE Trans Semicond Manuf 35(2):158–165. https://doi.org/10.1109/TSM.2022.3164904

Codella N et al (2019) Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). http://arxiv.org/abs/1902.03368. Accessed 28 Nov 2022

Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2010) ImageNet: a large-scale hierarchical image database. pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848

DeVries T, Taylor GW (2017) Improved regularization of convolutional neural networks with cutout. https://arxiv.org/abs/1708.04552v2. Accessed 12 June 2024

DIlmaghani S, Brust MR, Danoy G, Cassagnes N, Pecero J, Bouvry P (2019) Privacy and security of Big Data in AI systems: a research and standards perspective. In: Proceedings—2019 IEEE international conference on Big Data, Big Data 2019. pp 5737–5743. https://doi.org/10.1109/BIGDATA47090.2019.9006283

dos Santos Tanaka FHK, Aranha C (2019) Data augmentation using GANs. In: Proceedings of machine learning research. pp 1–16

Drumond RR, Brinkmeyer L, Schmidt-Thieme L (2023) Few-shot human motion prediction for heterogeneous sensors. In: Pacific-Asia conference on knowledge discovery and data mining. Springer Nature Switzerland, Cham, pp 551–563

Everingham M et al (2009) The Pascal Visual Object Classes (VOC) CHALLENGE. Int J Comput Vis 88(2):303–338. https://doi.org/10.1007/S11263-009-0275-4

Faraway JJ, Augustin NH (2018) When small data beats big data. Stat Probab Lett 136:142–145. https://doi.org/10.1016/j.spl.2018.02.031

Fei-Fei L, Fergus R, Perona P (2006) One-shot learning of object categories. IEEE Trans Pattern Anal Mach Intell 28(4):594–611. https://doi.org/10.1109/TPAMI.2006.79

Feng S, Duarte MF (2019) Few-shot learning-based human activity recognition. Expert Syst Appl 138:112782. https://doi.org/10.1016/J.ESWA.2019.06.070

Feyjie AR, Azad R, Pedersoli M, Kauffman C, Ayed IB, Dolz J (2020) Semi-supervised few-shot learning for medical image segmentation. arXiv Preprint. https://arxiv.org/abs/2003.08462

Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H (2018) GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 321:321–331. https://doi.org/10.1016/j.neucom.2018.09.013

Gao F et al (2018) A deep convolutional generative adversarial networks (DCGANs)-based semi-supervised method for object recognition in synthetic aperture radar (SAR) images. mdpi.com. https://doi.org/10.3390/rs10060846

Gao K, Liu B, Yu X, Yu A (2022) Unsupervised meta learning with multiview constraints for hyperspectral image small sample set classification. IEEE Trans Image Process 31:3449–3462. https://doi.org/10.1109/TIP.2022.3169689

Gharoun H, Momenifar F, Chen F, Gandomi A (2024) Meta-learning approaches for few-shot learning: a survey of recent advances. ACM Comput Surv. https://doi.org/10.1145/3659943

Gheisari M, Wang G, Bhuiyan MZA (2017) A survey on deep learning in Big Data. In: Proceedings—2017 IEEE international conference on computational science and engineering and IEEE/IFIP international conference on embedded and ubiquitous computing, CSE and EUC 2017. Institute of Electrical and Electronics Engineers Inc., pp 173–180. https://doi.org/10.1109/CSE-EUC.2017.215

Goceri E (2021) Diagnosis of skin diseases in the era of deep learning and mobile technology. Comput Biol Med. https://doi.org/10.1016/J.COMPBIOMED.2021.104458

Goodfellow IJ et al (2014) Generative adversarial nets. In: Advances in neural information processing systems, vol 27. http://www.github.com/goodfeli/adversarial. Accessed 20 Aug 2022

Goyal A, Bengio Y (2022) Inductive biases for deep learning of higher-level cognition. Proc R Soc A. https://doi.org/10.1098/RSPA.2021.0068

Gu R et al (2021) CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans Med Imaging 40(2):699–711. https://doi.org/10.1109/TMI.2020.3035253

Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24(2):8–12. https://doi.org/10.1109/MIS.2009.36

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. pp 770–778. http://image-net.org/challenges/LSVRC/2015/. Accessed 13 Dec 2022

He K, Girshick R, Dollár P (2019) Rethinking imagenet pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 4918–4927

Heider F, Simmel M (1944) An experimental study of apparent behavior. Am J Psychol 57(2):243. https://doi.org/10.2307/1416950

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/NECO.1997.9.8.1735

Huang Z, Datcu M, Pan Z, Lei B (2020) A hybrid and explainable deep learning framework for SAR images. In: International geoscience and remote sensing symposium (IGARSS). Institute of Electrical and Electronics Engineers Inc., pp 1727–1730. https://doi.org/10.1109/IGARSS39084.2020.9323845

Huh M, Agrawal P, Efros AA (2016) What makes ImageNet good for transfer learning? https://doi.org/10.48550/arxiv.1608.08614

Ibragimov B, Xing L (2017) Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 44(2):547–557. https://doi.org/10.1002/MP.12045

Inoue H (2018) Data augmentation by pairing samples for images classification. https://doi.org/10.48550/arxiv.1801.02929

Interian Y et al (2018) Deep nets vs expert designed features in medical physics: an IMRT QA case study. Med Phys 45(6):2672–2680. https://doi.org/10.1002/MP.12890

Iwata T, Kumagai A (2020) Few-shot learning for time-series forecasting. arXiv Preprint. https://arxiv.org/abs/2009.14379

Jiang Y, Neyshabur B, Mobahi H, Krishnan D, Bengio S (2019) Fantastic generalization measures and where to find them. http://arxiv.org/abs/1912.02178

Kang G, Dong X, Zheng L, Yang Y (2017) PatchShuffle regularization. arXiv. https://arxiv.org/abs/1707.07103

Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T (2020) styleGAN_with limited data. In: Conference on neural information processing systems (NeurIPS 2020), Vancouver, Canada. pp 12104–12114

Kawaguchi K, Bengio Y, Kaelbling L (2022) Generalization in deep learning. In: Mathematical aspects of deep learning. pp 112–148. https://doi.org/10.1017/9781009025096.003

Keshari R, Ghosh S, Chhabra S, Vatsa M, Singh R (2020) Unravelling small sample size problems in the deep learning world. In: Proceedings—2020 IEEE 6th international conference on multimedia big data, BigMM 2020. pp 134–143. https://doi.org/10.1109/BIGMM50055.2020.00028

Kim D, Koo J, Kim UM (2022) A survey on automated machine learning: problems, methods and frameworks. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 13302 LNCS. pp 57–70. https://doi.org/10.1007/978-3-031-05311-5_4

Kim SY, Malatesta JL, Lee WC (2023) Generalizability theory and applications. Int Encycl Educ 59-71

Kornblith S, Shlens J, Le QV (2019) Do better ImageNet models transfer better? pp 2661–2671

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM. https://doi.org/10.1145/3065386

Lake BM, Salakhutdinov R, Tenenbaum JB (2022) Human-level concept learning through probabilistic program induction, vol 21. p 2022, https://www.science.org. Accessed 22 Oct 2022

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2323. https://doi.org/10.1109/5.726791

Lecun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Lemberger P (2017) On generalization and regularization in deep learning. http://arxiv.org/abs/1704.01312

Li S, Deng W (2022) Deep facial expression recognition: a survey. IEEE Trans Affect Comput 13(3):1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Li X et al (2020) OSLNet: deep small-sample classification with an orthogonal softmax layer. IEEE Trans Image Process 29:6482–6495. https://doi.org/10.1109/TIP.2020.2990277

Liu B, Yu X, Yu A, Zhang P, Wan G, Wang R (2019) Deep few-shot learning for hyperspectral image classification. IEEE Trans Geosci Remote Sens 57(4):2290–2304. https://doi.org/10.1109/TGRS.2018.2872830

Ma G et al (2022) A transfer learning-based method for personalized state of health estimation of lithium-ion batteries. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2022.3176925

Majurski M et al (2019) Cell image segmentation using generative adversarial networks, transfer learning, and augmentations. https://nei.nih.gov/eyedata/amd. Accessed 15 June 2023

Marchesi M (2017) Megapixel size image creation using generative adversarial networks. http://arxiv.org/abs/1706.00082

Marcus G (2018) Deep learning: a critical appraisal http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial

Martin Lindstrom Company (2016) Small data: the tiny clues that uncover huge trends. John Murray Press. ISBN 9781473630154. https://books.google.co.in/books?id=UtJbCgAAQBAJ

Mazurowski MA, Buda M, Saha A, Bashir MR (2019) Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. J Magn Reson Imaging 49(4):939–954. https://doi.org/10.1002/JMRI.26534

Medela A et al (2019) Few shot learning in histopathological images: reducing the need of labeled data on biological datasets. In: Proceedings—international symposium on biomedical imaging, vol 2019-April. pp 1860–1864. https://doi.org/10.1109/ISBI.2019.8759182

Menghani G (2023) Efficient deep learning: a survey on making deep learning models smaller, faster, and better. ACM Comput Surv 55(12):1–37

Miller T (2017) Explanation in artificial intelligence: insights from the social sciences. http://arxiv.org/abs/1706.07269

Moreno-Barea FJ, Strazzera F, Jerez JM, Urda D, Franco L (2019) Forward noise adjustment scheme for data augmentation. In: Proceedings of the 2018 IEEE symposium series on computational intelligence (SSCI 2018). pp 728–734. https://doi.org/10.1109/SSCI.2018.8628917

Mormont R, Geurts P, Maree R (2018) Comparison of deep transfer learning strategies for digital pathology. pp 2262–2271

Mumuni A, Mumuni F (2024) Data augmentation with automated machine learning: approaches and performance comparison with classical data augmentation methods. https://arxiv.org/abs/2403.08352v1. Accessed 8 June 2024

Nagarajan V (2021) Explaining generalization in deep learning: progress and fundamental limits. http://arxiv.org/abs/2110.08922

Neyshabur B, Sedghi H, Zhang C (2020) What is being transferred in transfer learning? In: Advances in neural information processing systems, vol 33. pp 512–523

Ng HW, Nguyen VD, Vonikakis V, Winkler S (2015) Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction. pp 443–449. https://doi.org/10.1145/2818346.2830593

Niu S, Liu Y, Wang J, Song H (2020) A decade survey of transfer learning (2010–2020). IEEE Trans Artif Intell 1(2):151–166. https://doi.org/10.1109/TAI.2021.3054609

Olson M, Wyner A, Berk R (2018) Modern neural networks generalize on small data sets. In: Advances in neural information processing systems, vol 31

OpenAI et al (2019) Dota 2 with large scale deep reinforcement learning. http://arxiv.org/abs/1912.06680

Page MJ et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Int J Surg 88:105906. https://doi.org/10.1016/J.IJSU.2021.105906

Pan SJ, Tsang IW, Kwok JT, Yang Q (2011) Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 22(2):199–210. https://doi.org/10.1109/TNN.2010.2091281

Peng Z, Li Z, Zhang J, Li Y, Qi G-J, Tang J (2019) Few-shot image recognition with knowledge transfer. pp 441–449

Perconti P, Plebe A (2020) Deep learning and cognitive science. Cognition 203:104365. https://doi.org/10.1016/J.COGNITION.2020.104365

Pfister T, Charles J, Zisserman A (2014) Domain-adaptive discriminative one-shot learning of gestures. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 8694 LNCS, no PART 6. pp 814–829. https://doi.org/10.1007/978-3-319-10599-4_52

Plested J, Gedeon T (2019a) An analysis of the interaction between transfer learning protocols in deep neural networks. In: Neural information processing: 26th international conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, proceedings, part I 26. Springer International Publishing, pp 312–323

Plested J, Gedeon T (2019b) An analysis of the interaction between transfer learning protocols in deep neural networks. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 11953 LNCS. pp 312–323. https://doi.org/10.1007/978-3-030-36708-4_26/COVER

Plested J, Gedeon T (2022) Deep transfer learning for image classification: a survey. http://arxiv.org/abs/2205.09904

Power A, Burda Y, Edwards H, Babuschkin I, Misra V (2022) Grokking: generalization beyond overfitting on small algorithmic datasets. arXiv Preprint. http://arxiv.org/abs/2201.02177

Qian Z, Huang K, Wang QF, Zhang XY (2022) A survey of robust adversarial training in pattern recognition: fundamental, theory, and methodologies. Pattern Recognit 131:108889. https://doi.org/10.1016/J.PATCOG.2022.108889

Qin Z, Liu Z, Zhu P, Xue Y (2020) A GAN-based image synthesis method for skin lesion classification. Comput Methods Programs Biomed 195:105568

Quinn TP, Jacobs S, Senadeera M, Le V, Coghlan S (2022) The three ghosts of medical AI: can the blackbox present deliver? Artif Intell Med 124:102158. https://doi.org/10.1016/J.ARTMED.2021.102158

Raghu M, Zhang C, Kleinberg J, Bengio S (2019) Transfusion: understanding transfer learning for medical imaging. Adv Neural Inf Process Syst. https://doi.org/10.48550/arxiv.1902.07208

Rahadian A, Yusuf R (2023) Online learning facial expression detection using simplified AlexNet deep learning architecture: image data samples comparison experiment. pp 83–88. https://doi.org/10.1109/ICSET57543.2022.10011131

Rai A (2020) Explainable AI: from black box to glass box. J Acad Mark Sci 48(1):137–141. https://doi.org/10.1007/S11747-019-00710-5/TABLES/1

Raileanu R, Goldstein M, Yarats D, Kostrikov I, Fergus R (2021) Automatic data augmentation for generalization in reinforcement learning. In: Advances in neural information processing systems, vol 34. pp 5402–5415. https://github.com/rraileanu/auto-drac. Accessed 12 June 2024

Rayhan Y, Hashem T (2023) AIST: an interpretable attention-based deep learning model for crime prediction. ACM Trans Spat Algorithms Syst 9(2):1–31

Revina IM, Emmanuel WRS (2021) A survey on human face expression recognition techniques. J King Saud Univ Comput Inf Sci 33(6):619–628. https://doi.org/10.1016/J.JKSUCI.2018.09.002

Rodrigues PLC, Jutten C, Congedo M (2019) Riemannian procrustes analysis: transfer learning for brain-computer interfaces. IEEE Trans Biomed Eng 66(8):2390–2401. https://doi.org/10.1109/TBME.2018.2889705

Romero M, Interian Y, Solberg T, Valdes G (2019) Targeted transfer learning to improve performance in small medical physics datasets. https://doi.org/10.1002/mp.14507

Settles B (2009) Active learning literature survey. Technical report TR-1648. University of Wisconsin-Madison Department of Computer Sciences

Shaukat K, Luo S, Varadharajan V (2022) A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks. Eng Appl Artif Intell 116:105461. https://doi.org/10.1016/J.ENGAPPAI.2022.105461

Shen L, Lin Z, Huang Q (2016) Relay backpropagation for effective learning of deep convolutional neural networks. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 9911 LNCS. pp 467–482. https://doi.org/10.1007/978-3-319-46478-7_29/TABLES/6

Shijie J, Ping W, Peiyi J, Siping H (2017) Research on data augmentation for image classification based on convolution neural networks. Chin Autom Congr 2017:4165–4170

Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data. https://doi.org/10.1186/s40537-019-0197-0

Singh M et al (2022) Revisiting weakly supervised pre-training of visual perception models. pp 804–814. https://github.com/facebookresearch/SWAG. Accessed 9 Dec 2022

Singha A, Thakur RS, Patel T (2021) Deep learning applications in medical image analysis. Biomed Data Min Inf Retr 2021:293–350. https://doi.org/10.1002/9781119711278.ch11

Siuly S, Zhang Y (2016) Medical big data: neurological diseases diagnosis through medical data analysis. Data Scie Eng 1(2):54–64. https://doi.org/10.1007/s41019-016-0011-3

Song Y, Li J, Gao P, Li L, Tian T, Tian J (2022) Two-stage cross-modality transfer learning method for military-civilian SAR ship recognition. IEEE Geosci Remote Sens Lett. https://doi.org/10.1109/LGRS.2022.3162707

Spicer J, Sanborn AN (2019) What does the mind learn? A comparison of human and machine learning representations. Curr Opin Neurobiol 55:97–102. https://doi.org/10.1016/J.CONB.2019.02.004

Storrs KR, Kriegeskorte N (2019a) Deep learning for cognitive neuroscience. Cognit Neurosci. https://doi.org/10.7551/mitpress/11442.003.0077

Storrs KR, Kriegeskorte N (2019b) Deep learning for cognitive neuroscience. Cognit Neurosci. https://doi.org/10.48550/arxiv.1903.01458

Sun X, Xv H, Dong J, Zhou H, Chen C, Li Q (2021) Few-shot learning for domain-specific fine-grained image classification. IEEE Trans Ind Electron 68(4):3588–3598. https://doi.org/10.1109/TIE.2020.2977553

Suzuki K (2022) Small data deep learning for lung cancer detection in CT.In: Proceedings—IEEE 8th international conference on big data computing service and applications, BigDataService 2022. pp 114–118. https://doi.org/10.1109/BIGDATASERVICE55688.2022.00025

Świechowski M (2022) Deep learning and artificial general intelligence: still a long way to go. http://arxiv.org/abs/2203.14963

Tajbakhsh N et al (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312. https://doi.org/10.1109/TMI.2016.2535302

Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 11141 LNCS. pp 270–279. https://doi.org/10.1007/978-3-030-01424-7_27/COVER

Tan Y, Li Y, Huang SL, Zhang XP (2024) Transferability-guided cross-domain cross-task transfer learning. IEEE Trans Neural Netw Learn Syst

Tao X, Gong X, Zhang X, Yan S, Adak C (2022) Deep learning for unsupervised anomaly localization in industrial images: a survey. IEEE Trans Instrum Meas. https://doi.org/10.1109/TIM.2022.3196436

Tran AT, Nguyen CV, Hassner T (2019) Transferability and hardness of supervised classification tasks. In: Proceedings of the IEEE/CVF international conference on computer vision. pp 1395–1405

Triantafillou E, Zemel R, Urtasun R (2017) Few-shot learning through an information retrieval lens. In: Advances in neural information processing systems, vol 30

Tsai YHH, Salakhutdinov R (2017) Improving one-shot learning through fusing side information. arXiv Preprint. https://arxiv.org/abs/1710.08347

ul Sabha S, Assad A, Shafi S, Din NMU, Dar RA, Bhat MR (2024) Imbalcbl: addressing deep learning challenges with small and imbalanced datasets. Inte J Syst Assur Eng Manag 1:1–13. https://doi.org/10.1007/S13198-024-02346-3/TABLES/10

Verdegem P (2022) Dismantling AI capitalism: the commons as an alternative to the power concentration of Big Tech. AI Soc 1:1–11. https://doi.org/10.1007/S00146-022-01437-8/TABLES/1

Vinyals O, Deepmind G, Blundell C, Lillicrap T, KKavukcuoglu K, Wierstra D (2016) Matching networks for one shot learning. In: Advances in neural information processing systems, vol 29

Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology

Wang Y (2020) A mathematical introduction to generative adversarial nets (GAN). arXiv 2020. arXiv preprint arXiv:2009.00169.

Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. http://arxiv.org/abs/1712.04621

Wang Y, Ramanan D, Hebert M (2017) Learning to model the tail. NIPS

"Why machine learning 'succeeds' in development but fails in deployment." www.causaLens.com

Wibowo A et al (2022) Cardiac disease classification using two-dimensional thickness and few-shot learning based on magnetic resonance imaging image segmentation. J Imaging 8(7):194. https://doi.org/10.3390/JIMAGING8070194

Wong SC, Gatt A, Stamatescu V, McDonnell MD (2016) Understanding data augmentation for classification: when to warp? In: 2016 international conference on digital image computing: techniques and applications (DICTA 2016). https://arxiv.org/abs/1609.08764v2

Xing C, Rostamzadeh N, Oreshkin B, Pinheiro POO (2019) Adaptive cross-modal few-shot learning. In: Advances in neural information processing systems, vol 32

Xu Y, Li Z, Wang S, Li W, Sarkodie-Gyan T, Feng S (2021) A hybrid deep-learning model for fault diagnosis of rolling bearings. Measurement (London). https://doi.org/10.1016/j.measurement.2020.108502

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? In: Advances in neural information processing systems, vol 4, no January. pp 3320–3328. https://doi.org/10.48550/arxiv.1411.1792

Yousefzadeh R (2022) Deep learning generalization, extrapolation, and over-parameterization. http://arxiv.org/abs/2203.10366

Zhang C, Butepage J, Kjellstrom H, Mandt S (2019) Advances in variational inference. IEEE Trans Pattern Anal Mach Intell 41(8):2008–2026. https://doi.org/10.1109/TPAMI.2018.2889774

Zhang L, Liu J, Zhang B, Zhang D, Zhu C (2020) Deep cascade model-based face recognition: when deep-layered learning meets small data. IEEE Trans Image Process 29:1016–1029. https://doi.org/10.1109/TIP.2019.2938307

Zhang X, Wang Z, Liu D, Lin Q, Ling Q (2021a) Deep adversarial data augmentation for extremely low data regimes. IEEE Trans Circuits Syst Video Technol 31(1):15–28. https://doi.org/10.1109/TCSVT.2020.2967419

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O (2021b) Understanding deep learning (still) requires rethinking generalization. Commun ACM 64(3):107–115. https://doi.org/10.1145/3446776

Zhang C, Costa-Perez X, Patras P (2022) Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. IEEE/ACM Trans Netw 30(3):1294–1311. https://doi.org/10.1109/TNET.2021.3137084

Zhang W, Deng L, Zhang L, Wu D (2023) A survey on negative transfer. IEEE/CAA J Autom Sin 10(2):305–329. https://doi.org/10.1109/JAS.2022.106004

Zhang P, Zhong Y, Deng Y, Tang X, Li X (2019) A survey on deep learning of small sample in biomedical imageanalysis. arXiv:190800473

Zhao J, Yuan M, Cui J, Dong S, Qu Y, Xu B (2022) A small-sample intelligent fault diagnosis method based on deep transfer learning; a small-sample intelligent fault diagnosis method based on deep transfer learning. https://doi.org/10.1109/DSIT55514.2022.9943875

Zhao Q, Yu H, Chu J, Li T (2023) Few-shot learning with attention mechanism and transfer learning for import and export commodities classification. pp 125–130. https://doi.org/10.1109/CCIS57298.2022.10016358

Zheng Y, Jin M, Liu Y, Chi L, Phan KT, Pan S, Chen YPP (2022) From unsupervised to few-shot graph anomaly detection: a multi-scale contrastive learning approach. arXiv Preprint. https://arxiv.org/abs/2202.05525

Zhong Z, Zheng L, Kang G, Li S, Yang Y (2020) Random erasing data augmentation. In Proceedings of the AAAI conference on artificial intelligence, Vol 34, No 07, pp 13001-13008

Zhou Z-H (2018) A brief introduction to weakly supervised learning. Natl Sci Rev 5(1):44–53. https://doi.org/10.1093/nsr/nwx106

Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zhu Y, Liang X, Wang T, Xie J, Yang J (2022) Multi-information fusion fault diagnosis of bogie bearing under small samples via unsupervised representation alignment deep Q-learning. IEEE Trans Instrum Meas. https://doi.org/10.1109/TIM.2022.3225008

Zhuang F et al (2021) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76. https://doi.org/10.1109/JPROC.2020.3004555

Zoph B, Ghiasi G, Lin TY, Cui Y, Liu H, Cubuk ED, Le Q (2020a) Rethinking pre-training and self-training. In: Advances in neural information processing systems, vol 33. pp 3833–3845

Zoph B et al (2020b) Rethinking pre-training and self-training. In: Advances in neural information processing systems, vol. 2020-December. https://doi.org/10.48550/arxiv.2006.06882

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ishfaq Hussain Rather[1] · Sushil Kumar[1] · Amir H. Gandomi[2,3]

✉ Amir H. Gandomi
gandomi@uts.edu.au

Ishfaq Hussain Rather
ishfaq76_scs@jnu.ac.in

Sushil Kumar
skdohare@mail.jnu.ac.in

[1] School of Computer & Systems Sciences, Jawaharlal Nehru University, New Delhi, India

[2] Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia

[3] University Research and Innovation Center (EKIK), Óbuda University, Budapest 1034, Hungary