Please quote as: Karst, F., Li, M. M., Reinhard, P. & Leimeister, J. M. (2025). Not Enough Data to Be Fair? Evaluating Fairness Implications of Data Scarcity Solutions. Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS) (p./pp. 6886-6895), Hawaii, USA.

Not enough Data to be Fair? Evaluating Fairness Implications of Data Scarcity Solutions

Fabian Sven Karst
University of St.Gallen
fabian.karst@unisg.ch

Mahei Manhai Li
University of St.Gallen; University of Kassel
mahei.li@unisg.ch

Philipp Reinhard
University of Kassel
philipp.reinhard@uni-kassel.de

Jan Marco Leimeister
University of St.Gallen; University of Kassel
janmarco.leimeister@unisg.ch

Abstract

This study explores the implications of the use of data scarcity solutions on fairness in machine learning, specifically in consumer credit interest rate prediction. We develop a comprehensive taxonomy of Data Scarcity Solutions (DSS) by analyzing academic literature, data science competitions, and practical implementations. We identify six distinct DSS clusters: Data Extension, Pre-Training, Public Data Inclusion, Data Sharing, Federated Learning, and Active Learning. Our evaluation shows that most DSS enhance both performance and fairness, with minimal negative correlation between the two. Notably, approaches incorporating external or synthetic data significantly improve fairness. This research contributes to understanding DSS beyond algorithmic performance, providing a framework for evaluating their societal impact. Furthermore, it offers practitioners a taxonomy to select the right method for tackling data scarcity and addresses fairness concerns in real-world scenarios.

Keywords: Data Scarcity, Fairness, Machine Learning, Consumer Credit, Taxonomy

Introduction

Access to affordable credit is linked to positive life outcomes such as upward class mobility (Wydick, 1999), college enrolment (Solis, 2017), and well-being (Dobridge, 2016). However, significant discrimination in credit costs for minority borrowers results in a yearly disadvantage of over $450 million (Bartlett et al., 2022), reducing the benefits of credit access or preventing borrowing altogether. Despite new technologies in the credit origination process, this fairness problem persists (Bartlett et al., 2022). One reason for these biases (Y. Zhang & Long, 2021) as well as a major limitation of machine learning (ML) performance in the field of private lending is the limited availability of training data (Bhatore et al., 2020). Consequently, robust strategies to combat data scarcity are needed, and ensuring fairness is needed to drive adoption, especially in sensitive use cases (Li et al., 2024).

However, such data scarcity solutions (DSS) are often highly specific to use cases (Alzubaidi et al., 2023) and thus DSS often focuses on different characteristics, such as privacy, computational requirements, and communication for federated learning (Murshed et al., 2022); data type and data composition for data augmentation (Bayer et al., 2023); data privacy and integration for data sharing (Fang et al., 2017). A systematic comparison of these attributes between solution types and use cases is missing. Consequently, there is a clear need for a structure that can effectively help practitioners compare solutions and find the one most suitable for their use case while considering fairness. This leads us to the following research questions:

RQ1: *Which characteristics can be used to describe DSS conceptually and which archetypes of DSS can be distinguished based on these characteristics?*

RQ2: *What are the implications of these solutions for fairness and bias in credit interest rate predictions?*

For **RQ1**, we aim to develop a DSS taxonomy based on academic literature, data science competition contributions, GitHub repositories, and startup companies, and use it to derive archetypical DSS. Understanding the different DSS characteristics is the basis for ensuring that these solutions do not inadvertently perpetuate or exacerbate biases, particularly in sensitive applications like consumer credit. Thus, for **RQ2**, we assess the impact of the identified DSS archetypes on fairness using consumer credit interest rate prediction as our application use case. While our literature review focuses on data scarcity solutions and their various characteristics, it is important to note that the implications for fairness and bias, particularly in the context of consumer credit interest rate predictions, have not been thoroughly explored in the existing literature. Our study addresses this gap by assessing the impact of different DSS archetypes on fairness, providing new insights that extend beyond the current body of research.

URI: https://hdl.handle.net/10125/109672
978-0-9981331-8-8

We first review prior research on data scarcity and fairness in ML, introduce our methodology for constructing and clustering our taxonomy, and analyze the fairness of these methods. Finally, we present our results and discuss the implications.

Research Background

Data Scarcity Solutions

While there is no single definition of DSS in the existing literature, Bansal et al. (2022) define them as "methods that deal with the issue of the limited dataset[s]". Yu et al. (2022) refine this by differentiating three types: where insufficient samples are available, the data is too low dimensional, or both. Thus, we consider all methods that help ML models deal with insufficient data as DSS. Drawing from prior literature reviews (Li et al 2020; Bansal et al. 20222; Alzubaidi et al 2023), different broader solution categories for data scarcity were identified: *1) Data augmentation* - This technique is a "regularization scheme that artificially inflates the data set by using label-preserving transformations to add more invariant examples" (Taylor & Nitschke, 2018). The rise of generative deep-learning models has resulted in advancements in the field (Bansal et al., 2022). *2) Transfer learning* – The transfer of knowledge from existing models is employed in domains where large amounts of similar public data sets are available, such as text and image processing (Gruetzemacher & Paradice, 2022). *3) Human Expertise Integration* - Integrating humans to address data scarcity is also viable, but finding a representation of domain knowledge that can be incorporated into the model is challenging and costly (M. M. Li, 2023; C. Yin et al., 2019).

Outside the literature review classifications, *Collaborative Solutions* are effective when sufficient data is available but distributed across multiple entities. Sharing data between institutions or creating a joint model on this data can overcome the data scarcity problem for individual data subset owners (Fang et al., 2017). However, in multiple domains, privacy is a critical concern when sharing data, and the literature focuses on how distributed data can be used cooperatively without revealing sensitive information. Two approaches exist in this regard. The first is to share encrypted or synthetic data to ensure privacy (Aloufi et al., 2022; Karst et al., 2023, 2024). Secondly, models can be trained in a distributed fashion so that data stays within the boundaries of each data provider (J. Zhang et al., 2022). Another potential solution not covered by the reviews is the idea of increasing *Labeling Efficiency*, e.g. by making it worthwhile to label data manually (Anahideh et al., 2022; M. M. Li et al., 2024).

Fairness in Machine Learning

Recently, a variety of research focusing on evaluating fairness, and building bias-free models for ML (Caton & Haas, 2024; Chouldechova & Roth, 2018; Pessach & Shmueli, 2022) with two categories of metrics: *Individual Fairness Metrics* - These measure whether similar individuals are treated equally but they heavily depend on the similarity metric, often making them impractical (Chouldechova & Roth, 2018). *Group Fairness Metrics* - These evaluate the similarity between outcomes for members of different sensitive groups but only for the average group member (Dwork et al., 2012).

In addition to the metrics, three method types enhance fairness along ML processes: *Pre-processing Methods* - These alter training data to make it fairer by reassigning labels (Luong et al., 2011) or changing feature representation (Backurs et al., 2019). This can increase fairness for any model type but may reduce explainability and complicate performance evaluation.

In-processing Methods - These modify the ML algorithm to account for fairness during training, often by adding constraints to the optimization problem (Kamishima et al., 2012). While popular, these methods are limited to specific models and predominantly address binary classification, leaving other tasks insufficiently covered (Pessach & Shmueli, 2023).

Post-processing Methods - These modify the output scores of the classifier to make decisions fairer, for example by creating separate decision thresholds per group (Hardt et al., 2016). Although applicable to any classification algorithm, these methods typically yield inferior results in comparison (Woodworth et al., 2017). Beyond this, few papers currently tackle fairness in non-classification tasks, understanding why individually fair elements (e.g., data) can produce unfair results. (Caton & Haas, 2024; Chouldechova & Roth, 2018). Additionally, research into modifying algorithms and training procedures to handle limited data has shown that these adjustments can significantly impact algorithm fairness, often more so than changing the algorithm itself (Friedler et al., 2019). Reasons span including biased data or model weights (Seth & Pai, 2024), overstating the presence of certain samples (Chakraborty et al., 2021) or engineering features revealing sensitive characteristics (De-Arteaga et al., 2019). Currently, no paper exists, comparing the effect of fairness between different DSS'.

Research Methodology

To analyze the impact of DSS on fairness, a two-step research process was chosen, firstly archetypical DSS

were identified. Then, we apply DSS to the context of credit origination, evaluating their impact on fairness.

Taxonomy Development & Clustering

Phase 1 - Set up the Database: We created a database of publications and practical solutions for taxonomy-building by conducting a structured literature review following Webster and Watson (2002). We used the search string: "Data Scarcity" OR (("Data Augmentation" OR "Data Sharing" OR "Synthetic Data" OR "Active Learning" OR "Federated Learning" OR "Transfer Learning") AND "Machine Learning"). The search focused on top IS journals, the AIS Basket of Eight, ICIS, ECIS, HICSS conferences, and the ACM Computing Surveys journal. Through forward and backward searches, we identified 439 unique articles. We then applied inclusion criteria (a) explicit discussion of data scarcity, (b) presenting a model for data scarcity, and (c) discussing architectures for limited data, reducing the number to 78 papers.

We extended our database with real-world examples from three leading online platforms. From Kaggle.com, we selected the top 100 competitions (by contribution) and solutions (by ratings), resulting in 58 data scarcity-relevant contributions. From GitHub.com, using literature review keywords, we found 32 repositories with at least 500 stars. From Crunchbase.com, we identified 95 data scarcity startups and included 37 based on operational status and website content. Our final database thus comprises 209 DSS samples.

Phase 2- Taxonomy Development: We follow the taxonomy development method by Nickerson et al. (2013) and its extension by Kundisch et al. (2022), which provides a rigorous process for iteratively constructing and evaluating taxonomies from theoretical and often overlooked practical empirical evidence. In line with our methodology, we specify the phenomenon (data scarcity), the target user group (practitioners in machine learning), and the taxonomy's purpose (identifying, classifying, and analyzing). We first establish meta characteristics, which serve as the initial taxonomy structure (Nickerson et al., 2013). Given that our taxonomy aims to capture the features of solutions for data scarcity, with a specific emphasis on how they are integrated, how additional knowledge is leveraged, and how they interact and modify data and models, we defined them as our meta-characteristics. They were further refined during 5 iterations until no changes to the dimensions occurred, indicating conceptual saturation (an overview of the iterative taxonomy development can be found in the online appendix).

Phase 3 - Taxonomy Clustering and Archetype Identification: The third phase targeted the empirical identification of DSS archetypes from the taxonomy by conducting a cluster analysis on our sample DSS (Kundisch et al., 2022). The objective of cluster analysis is to form groups of objects whereby objects in the same group are as similar as possible and objects in different groups are as dissimilar as possible (Kaufman & Rousseeuw, 2009). We did this by utilizing the well-established k-means clustering algorithm and tested cluster sizes from 3 to 15 (Wu, 2012). The elbow method using distortion was used to identify the optimal number of clusters (Kodinariya & Makwana, 2016), which can be seen in Figure 2. Next, we confirmed our selection of the optimal number of clusters, by analyzing the silhouette score for multiple numbers of clusters (for more details see the online appendix). This analysis resulted in the identification of 6 clusters, which were consistent with prior literature. Furthermore, based on the examples within each cluster a clear dominant attribute combination could be identified, making them interpretable and thus, qualitatively confirming the validity of the identified number of clusters.

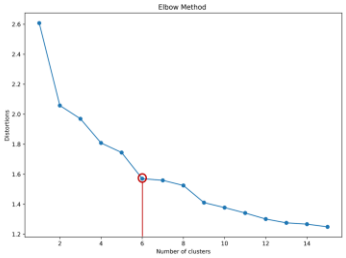

Figure 1: Number of Cluster Selection using the Elbow Method

Fairness Evaluation

Historically, credit allocation has been biased against non-white lenders (Cheng et al., 2015), thus this study explores the implications of fairness in setting interest rates for private loans. To do this we needed a dataset containing loan application data, applicant-sensitive group information, and charged interest rates. Since no such dataset exists publicly, we combined information from multiple sources to create one with the required properties. We used LendingClub's dataset of credit application data and interest rates as it is not only diverse due to LendingClub's position as one of the largest peer-to-peer lending platforms but also publicly available, ensuring the replicability of our study. Next, this data was combined with US census data on ethnic distribution by area code. To assign ethnic labels to individual loan applications, we matched loan applications based on zip codes with the ethnic distribution in the area. We assigned an ethnicity label of "white" or "non-white" if the respective group constituted more than 75% of the population in that area.

Finally, we split the dataset into multiple subsets based on the US Bureau of Economic Analysis regions and retained only those subsets with more than 100 observations for each subgroup. This ensured that each data subset owner had limited data access, but collaborators with similar data existed. This process resulted in the following data subsets (see Table 1):

Region	# observations	Average interest rate	Percentage of non-white borrowers
Far West	66526	13,18%	3,01%
Great Lakes	122166	13,24%	1,88%
Mideast	97433	13,44%	15,09%

Table 1: Descriptive statistics

Before setting up our evaluation, we first need to define measures for quantifying performance and fairness. For measuring regression performance, we used the well-established R^2 score (Chicco et al., 2021). However, measuring fairness is more complex, as it is highly context-sensitive and guidance for non-classification scenarios is lacking. Thus, we developed our own method for evaluating fairness in our scenario. Given the limited individual information, assessing individual fairness was not feasible, so we opted for a measure assessing group fairness. We chose the difference in interest rates between ethnicities (white vs. non-white) as our fairness measure due to its ability to highlight disparities between groups and its easy interpretability. While this measure does not provide absolute conclusions about fairness, it allows us to compare the fairness of different DSS approaches to each other, the baseline, and the ground truth/real data.

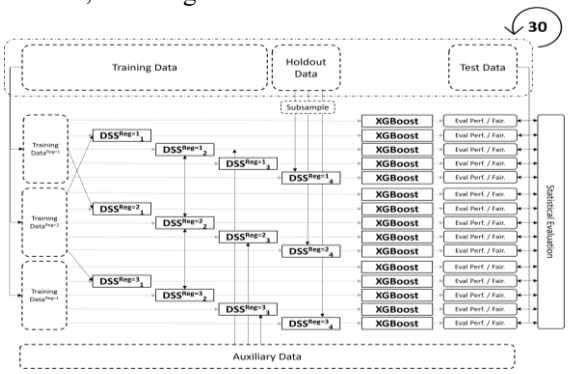

Figure 2: Training and evaluation setup

Building on the evaluation procedures proposed by Raschka (2020) we extend the recommended 5x2cv paired t-test. This resulted in the experimental setup, displayed in Figure 3, in which the dataset was split into different subsets (train, holdout, test) which were randomly created using bootstrapping (n= 30).
In each iteration we apply all DSS per region, utilizing data from other regions and external sources where necessary. For each DSS and the baseline, an XGBoost Regressor, the quasi-standard algorithm in the domain (Odegua, 2020), is trained and its predictions on the test dataset are used to evaluate performance and fairness. The accumulated data of all iterations is then employed to calculate individual t-tests to test for a pairwise positive/non-zero difference between the DSS measure and the measure for the real data or ML baseline, thus allowing us to conclude if its utilization has a significant positive impact on fairness or performance. The complete implementation with more details is openly accessible in the online appendix.

Results

Taxonomy Development & Clustering

Our final taxonomy encompasses ten dimensions, each featuring two to six characteristics. We visualize the taxonomy as a morphological box to illustrate the relationships between the dimensions and characteristics (see Figure 4). Afterward, we describe the taxonomies dimensions in detail.

Meta Dimension	Dimension	Characteristic					Exclusive
Integration	Level	Data	Model				No
	Architecture	Centralized	Decentralized	Distributed			Yes
Knowledge	Human Contribution	Environment	Explicit Knowledge	Feedback	Selection	None	Yes
	Exchange	None	Unilateral	Multilateral			Yes
	Type	Model Weights	Similar Labelled Data	Related Labelled Data	Unlabelled Data	Other	None	No
Data	Privacy	Data-modifying approaches	Data encrypting approaches	Data minimizing approaches	Data confining approaches	None		Yes
	Type	Text	Image	Sequence	Tabular	Other		No
	Augmentation Dimension	Increase samples	Decrease samples	Increase features	Decrease features	None		Yes
	Augmentation Type	Model-based	Simulation-based	Trans-formation	Combination	Selection	None	Yes
Model	Algorithm Type	Supervised ML	Semi-supervised ML	Unsupervised ML	Reinforcement Learning	Independent		No
	Augmentation	Knowledge Integration	Loss Function	Training Procedure	Model Architecture	None		Yes
	Pre-Training	Supervised	Unsupervised	None				Yes

Figure 3: Taxonomy of DSS for Machine Learning

Integration - Level: In multi-modal learning, the timing of data integration is critical. Early fusion combines data before model input, while late fusion integrates within the model (Ramachandram & Taylor, 2017). Applying this to DSS, we differentiate between methods that modify model input independently of the machine learning model and those that adjust the model itself, thus restricting algorithm compatibility (Feng, 2022; Song et al., 2023).
Integration – Architecture: DSS employing external private data uses various architectures (J. Zhang et al., 2022). Drawing from federated learning, we categorize them as centralized, decentralized, or distributed (Verbraeken et al., 2021). Centralized systems store and process data in one location, whereas decentralized systems keep data local, aggregating only model information centrally. Distributed systems are similar to decentralized systems but eliminate the need for central infrastructure, facilitating direct data exchange and increasing privacy (Erler et al., 2022).
Knowledge - Human Contribution: Human expertise addresses data scarcity (Bansal et al., 2022). We categorize how domain expertise integrates into data or models: explicit knowledge representations (Bui, 2019;

Choi et al., 2017), humanly created priors (Ma et al., 2018), feedback to models (Anahideh et al., 2022; Saito et al., 2014), and selecting relevant features.

Knowledge – Exchange: This dimension involves acquiring external knowledge. Unilateral exchange uses publicly available data and pre-trained models (Brickley et al., 2019; Han et al., 2021). Multilateral exchanges share data or train models jointly, especially when privacy is a concern (Nguyen et al., 2021; X. Yin et al., 2022).

Knowledge – Type: External knowledge integration often uses model weights from related tasks (Han et al., 2021; Weiss et al., 2016). Other methods include labeled data sharing (DuMont Schütte et al., 2021; X.-B. Li & Qin, 2013), using related domain data (J. Wang & Zhou, 2020); (Dvornik et al., 2021) or applying expert knowledge to limit the solution space (Ma et al., 2018) or by applying data transformations (Roe et al., 2020), with the latter subsumed as diverse methods as "Other" in the taxonomy.

Data - Privacy: We adapted Meurisch & Mühlhäuser's (2022) four AI-level data protection approaches for DSS. Data-modifying sanitizes user data, data-encrypting ensures confidentiality, data-minimizing reduces required personal data, and data-confining keeps data local and separated.

Data – Type: Different algorithms suit different data types. For instance, image data augmentation includes rotating and flipping (Perez & Wang, 2017) while text augmentation uses round-trip translation (Bayer et al., 2023). Multi-modal approaches combine data types (Baltrusaitis et al., 2019).

Data – Augmentation Dimension: Yu et al. (2022) identify sample scarcity, feature scarcity, and their combination. Solutions either expand or reduce the sample space (e.g.: synthetic data (Brophy et al., 2023; Cai et al., 2021) vs. undersampling (Xu-Ying Liu et al., 2009)) or the feature space (feature engineering (Butcher & Smith, 2020) vs. feature selection (J. Li et al., 2018)). Additionally, data transformation techniques like sample weighting (Shu et al., 2019) or kernel transformations (Y. Liu et al., 2023) can alter data without changing its dimensionality.

Data – Augmentation Type: Different ways exist to augment data to improve performance in data-scarce environments. Based on Bansal et al. (2022) and Shorten & Khoshgoftaar (2019), five augmentation types have emerged. Model-based techniques generate new data instances or features (Figueira & Vaz, 2022; Khosravian et al., 2021). Simulation-based methods create data through expert-informed simulations (Mastorakis et al., 2018). Feature engineering transforms existing data to improve its utility (Butcher & Smith, 2020; Roe et al., 2020). Further, combining data with external sources (Erler et al., 2022; Fang et al., 2017) as well as selecting features or samples to reduce noise (J. Li et al., 2018) have been shown to be successful augmentation types in data-scarce environments.

Model – Algorithm Type: We identified four key algorithm types prevalent in DSS: supervised learning (mapping inputs to outputs), unsupervised learning (pattern recognition without predefined targets), semi-supervised learning (combining labeled and unlabeled data), and reinforcement learning (decision-making based on observations) (Alzubaidi et al., 2023; Taiwo, 2010).

Model – Augmentation: Various modifications enhance model performance on small datasets. Bansal et al. (2022) discuss integrating new knowledge via transfer learning and adjusting loss functions for scarce data. Li et al. (2020) extend this by highlighting multi-task and meta-learning as alternative knowledge integration methods. Few-shot learning has recently gained popularity for incorporating contextual data. Other literature suggests introducing an unsupervised step, such as self-supervised learning, to the training procedure (Farhat et al., 2023; Schiappa et al., 2022). Additionally, adjusting model architecture is crucial for data-scarce scenarios. Common adjustments include using ensemble methods (Tüysüzoğlu & Yaslan, 2018; P. Wang et al., 2019) and advanced techniques like knowledge codistillation (Hinton et al., 2015; Ni et al., 2022).

After integrating each of the 209 cases into the taxonomy we ran the clustering scheme described in the methodology section to identify archetypical DSS. After determining the optimal number of clusters, the following six clusters emerged each covering 16 to 88 examples:

Data Extension (DE, 42.11%): The Data Extension cluster encompasses methods that enhance existing datasets through either expanding data points or modifying features. Solutions in this cluster can be divided into two sub-categories. The first involves adding new data points, such as through synthetic data generation, simulation-based data creation, or data augmentation. The second focuses on feature modification, including techniques like feature engineering and the creation of embeddings. These approaches address data scarcity at the data level and are typically implemented in centralized, local environments, which can enhance data privacy by keeping data modifications within secure confines.

Pre-Training (PT, 20.57%): This cluster focuses on leveraging models pre-trained on different tasks or related datasets as a foundation for new model training. Often, these pre-trained models are based on unlabelled data or datasets available within the company, providing a rich source of transferable knowledge. This approach

integrates existing insights into the new model, classifying it as a model-based DSS. The process is usually carried out locally, ensuring data privacy by avoiding external data exchanges. However, the approach's success hinges on having access to similar or related pre-training datasets, which can restrict its usability in scenarios lacking appropriate data sources.

Public Data Inclusion (PD, 12.92%): Public data inclusion enriches models by integrating external resources at both the data and model levels. At the data level, incorporating public datasets can expand the feature space, enhancing the model's input diversity. At the model level, utilizing publicly available pre-trained weights offers a superior starting point compared to random initialization, accelerating model training and improving performance. Despite the requirement for suitable public datasets or pre-trained models, which makes these solutions highly task-specific, public data inclusion remains a prevalent method for infusing internal models with external knowledge.

Data Sharing (DS, 8.61%): Data sharing significantly enhances the volume of data accessible to a model through the exchange of data between external parties (often competitors). However, practical challenges such as data privacy concerns and the complexities of establishing effective data exchange systems often pose significant obstacles. Solutions within this cluster aim to mitigate these issues by enabling secure data sharing through encryption or anonymization techniques. Additionally, they offer architectures and systems to facilitate data sharing, either through centralized platforms or distributed networks.

Federated Learning (FL, 8.13%): Approaches within this cluster can be assigned to the federated learning category and tackle data scarcity by enabling the cooperative training of supervised machine learning models across multiple distributed data sources without sharing the actual data. Instead of exchanging raw data, these approaches share model weights, allowing for distributed training while preserving data privacy. Typically implemented using a "Master – Client" architecture, a centralized server manages the coordination of training and the aggregation of model updates from various clients. This method significantly enhances privacy through data minimization, offering a secure alternative to traditional data-sharing techniques by ensuring that only model parameters are exchanged, not the data itself.

Active Learning & Pseudo Labelling (AL, 7.66%): The goal of approaches in this cluster is to boost the performance of supervised ML models by transforming unlabeled datasets into valuable sources of labeled training data. This process is achieved through two primary techniques: pseudo-labeling and active learning. Pseudo-labeling generates labels by using an auxiliary model to predict them from unlabeled data. In contrast, active learning optimizes the labeling process by selecting the most informative data points, which are then manually annotated, significantly improving labeling efficiency. These methods stand out by simultaneously modifying both the data and the model while maintaining human involvement in the loop (active learning).

Fairness and Performance Evaluation

To start our evaluation, first, specific DSS solutions for each cluster need to be defined. This was done using the most commonly mentioned solution within each cluster applicable to our scenario, resulting in the following selection:

Data Extension (DE): DE, the largest cluster, encompasses various solutions that extend data either by rows or columns. For our experiment, we chose polynomial feature generation and feature selection for column-wise extension (DE1). Conversely, row-wise extension is exemplified by synthetic data generation using CTGAN (DE2).

Pre-Training (PT): As the dataset available is limited and we cannot easily produce other data that would be available to a consumer credit company, pre-training utilized data from a randomly selected region. Then the model was initially trained on this external data and then fine-tuned to adapt to the specific regional data (PT1).

Public Data Inclusion (PD): Given the absence of suitable public model weights for credit interest rate prediction, we incorporated publicly available datasets. The first dataset (PD1) was uncorrelated with the sensitive class (average correlation: -0.001) and included broad economic indicators like GDP growth, unemployment rate, and interest rate. The second dataset (PD2), which showed a correlation with the sensitive class (correlation: -0.233), contained the housing price index for a specific region. This, not only allows us to analyze the benefit of public data in general but also for public data which might be problematic concerning the sensitive class.

Data Sharing (DS): To assess the impact of data sharing, the training dataset was supplemented with data from a randomly chosen second region before training (DS1).

Federated Learning (FL): In federated learning, a shared model was trained using the training datasets from multiple regions. This shared model was then evaluated on the test sets of the individual regions (FL1).

Active Learning (AL): To simulate AL the holdout set was used in combination with an active learning procedure to label the most informative data points of

this set (in total 1% of the total data) in an iterative manner (AL1) and include them into model training.
These solutions were applied to the bootstrapped dataset following the methodology outlined earlier. Each algorithm and bootstrap sample yielded two scores: performance (measured by R^2) and fairness (measured by the predefined fairness metric). We then compared the performance of various DSS methods against a baseline model without any DSS, as shown in Figure 5.

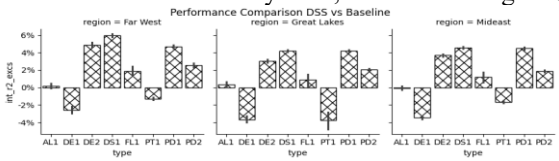

Figure 4: Performance comparison of DSS relative to the baseline model

As illustrated in Figure 5, most DSS applications enhanced the model's performance across different regions, as expected. Only DE1 and PT1 resulted in a performance decline. Since DSS would only be practically used if they improve performance, those with inferior outcomes were excluded from further analysis. We then examined the fairness of the predictions generated by the DSS. Across all regions, the use of DSS improved fairness by reducing the disparity in predicted interest rates between classes compared to the disparities observed in the actual data (see Figure 6).

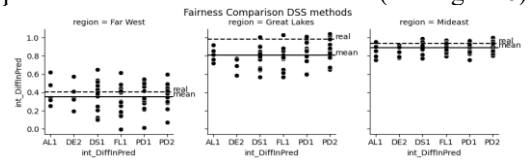

Figure 5: Fairness comparison of different DSS

Statistical testing of the observed differences confirms that the improvement in fairness due to DSS is significant and consistent across all algorithms. This leads to the conclusion that DSS generally enhances both performance and fairness compared to the actual dataset (see details in Table 2).

	AL1	DE2	DS1	FL1	PD1	PD2
real (H_1: μ > 0)	0.00000***	0.01187*	0.00000***	0.00000***	0.00000***	0,00024***
baseline (H_1: μ > 0)	0.61898	0.00483***	0.11710	0.00000***	0.70216	0.99920
baseline (H_1: μ != 0)	0.76203	0.00965**	0.23421	0.00000***	0.59566	0.00158**

Statistical Significance: * p ≤ 0.05 ** p ≤ 0.01 *** p ≤ 0.001

Table 2: Statistical comparison of fairness

Subsequently, we compared the fairness of algorithms using DSS to a baseline algorithm without DSS with two tests: one to verify performance improvement and another to assess whether there was no significant difference in fairness between the baseline and DSS-enhanced algorithms. As shown in Table 2, FL1 and DE2 exhibited increased fairness compared to the baseline. For other solutions, except PD2, there was no significant difference in fairness. Notably, PD2, which utilized data highly correlated with the sensitive group, resulted in decreased fairness compared to the baseline algorithm. Overall, our results suggest that DSS can be strategically employed to achieve better predictive performance and fairness. Only a minimal negative correlation between fairness and performance (corr: -0.054), was found thus, providing evidence that DSS can increase performance simultaneously (perform better than without DSS), and the increase in performance is only slightly reduced by increased fairness. Furthermore, looking more closely, we see that in particular algorithms incorporating external data (PD1, DS1), can be helpful to increase performance and significantly increase fairness. To an extend this holds for synthetic data generation (DE2) too, making it a viable approach when external data is not available. However, our research also reveals that the selection of DSS methods needs to be carefully analyzed and the effect might be highly context-sensitive.

Discussion

This study contributes to IS research by examining data scarcity through a sociotechnical lens, evaluating both technical characteristics and their societal impact, particularly on fairness in machine learning.
Given the increasing need for data, keeping pace with developments is challenging. To address this, we provide a guiding framework for DSS, resulting in a detailed taxonomy that serves as a robust framework for future fair design-oriented research, promoting comparability and transferability of findings and preventing redundant research efforts. Our taxonomy incorporates real-world applications, enriching the theoretical framework and making it more applicable to industry challenges, extending previous research by Li (2020), Bansal (2022), and Alzubaidi (2023). Further, we extended the basic classification of DSS (data augmentation, transfer knowledge, and cost-sensitive learning) with dimensions that provide a more comprehensive description of DSS (collaborative approaches). Furthermore, we are the first to compare DSS beyond ML algorithm performance, analyzing their impact on fairness. This expands existing research, primarily focused on algorithmic fairness (Caton & Haas, 2024; Chouldechova & Roth, 2018), and sheds light on the trade-offs between fairness and performance. We demonstrate that DSS can simultaneously improve performance and fairness, providing actionable insights for both researchers and practitioners in developing equitable and effective DSS. However, this research is also limited as the fairness of the archetypical DSS' was only evaluated on a single dataset and only a limited number of approaches per domain could be tested. Further research is needed to

achieve a more rigorous understanding of DSS on fairness across tasks, ML algorithms, and domains.

Conclusion

This study addresses the challenge of data scarcity in machine learning by developing a comprehensive taxonomy of DSS and analyzing their impact on performance and fairness in consumer credit interest rate prediction. This provides practitioners not only with a taxonomy to better understand DSS and map out their use cases but also provides guidance to them for selecting a specific solution category by creating archetypes mapped to the Taxonomy. Furthermore, this paper does not only raise awareness about the potential fairness implications of such solutions but also offers a framework on how to evaluate their fairness, exemplified for the use case of credit risk prediction. Lastly, we showed the benefit of selected DSS for performance and fairness, thus providing additional insights for DSS selection by practitioners. Further research is needed to validate the generalizability of our insights across domains and create a more practitioner-friendly taxonomy-based decision support system to identify the optimal DSS given a use case.

References

Aloufi, A., Hu, P., Song, Y., & Lauter, K. (2022). Computing Blindfolded on Data Homomorphically Encrypted under Multiple Keys: A Survey. *ACM Computing Surveys*, *54*(9), 1–37.

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., Duan, Y., Abdullah, A., Farhan, L., Lu, Y., Gupta, A., Albu, F., Abbosh, A., & Gu, Y. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(1), 46.

Anahideh, H., Asudeh, A., & Thirumuruganathan, S. (2022). Fair active learning. *Expert Systems with Applications*, *199*, 116981.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., & Wagner, T. (2019). Scalable Fair Clustering. *Proceedings of the 36th International Conference on Machine Learning*, 405–413.

Badr, E. (2022). Images in Space and Time: Real Big Data in Healthcare. *ACM Computing Surveys*, *54*(6), 1–38.

Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443.

Bansal, Ms. A., Sharma, Dr. R., & Kathuria, Dr. M. (2022). A Systematic Review on Data Scarcity Problem in Deep Learning: Solution and Applications. *ACM Computing Surveys*, *54*(10s), 208:1-208:29.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics*, *143*(1), 30–56.

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2023). A Survey on Data Augmentation for Text Classification. *ACM Computing Surveys*, *55*(7), 1–39.

Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: A systematic literature review. *Journal of Banking and Financial Technology*, *4*(1), 111–138.

Brickley, D., Burgess, M., & Noy, N. (2019). Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. *The World Wide Web Conference*, 1365–1375. WWW '19: The Web Conference.

Brophy, E., Wang, Z., She, Q., & Ward, T. (2023). Generative Adversarial Networks in Time Series: A Systematic Literature Review. *ACM Computing Surveys*, *55*(10), 1–31.

Bui, T. (2019). *Combining Enterprise Knowledge Graph and News Sentiment Analysis for Stock Price Volatility Prediction*. HICSS.

Butcher, B., & Smith, B. J. (2020). Feature Engineering and Selection: A Practical Approach for Predictive Models. *The American Statistician*, *74*(3), 308–309.

Cai, Z., Xiong, Z., Xu, H., Wang, P., Li, W., & Pan, Y. (2021). Generative Adversarial Networks: A Survey Toward Private and Secure Applications. *ACM Computing Surveys*, *54*(6), 132:1-132:38.

Caton, S., & Haas, C. (2024). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, *56*(7), 1–38.

Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: Why? how? what to do? *Proceedings of the 29th ACM European Software Engineering Conference*, 429–440.

Cheng, P., Lin, Z., & Liu, Y. (2015). Racial Discrepancy in Mortgage Interest Rates. *The Journal of Real Estate Finance and Economics*, *51*(1), 101–120.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., & Sun, J. (2017). GRAM: Graph-based Attention Model for Healthcare Representation Learning. *Proceedings of the 23rd ACM SIGKDD*, 787–795.

Chouldechova, A., & Roth, A. (2018). *The Frontiers of Fairness in Machine Learning* (arXiv:1810.08810). arXiv.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.

Dobridge, C. L. (2016). For Better and for Worse? Effects of Access to High-Cost Consumer Credit. *Finance and Economics Discussion Series*, *2016*(056), 1–30.

DuMont Schütte, A., Hetzel, J., Gatidis, S., Hepp, T., Dietz, B., Bauer, S., & Schwab, P. (2021). Overcoming barriers to data sharing with medical image generation: A

comprehensive evaluation. *Npj Digital Medicine*, *4*(1), Article 1.

Dvornik, N., Mairal, J., & Schmid, C. (2021). On the Importance of Visual Context for Data Augmentation in Scene Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(6), 2014–2028.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Erler, C., Schinle, M., Dietrich, M., & Stork, W. (2022). Decision model to design a blockchain-based system for storing sensitive health data. *ECIS 2022 Research Papers*.

Fang, R., Pouyanfar, S., Yang, Y., Chen, S.-C., & Iyengar, S. S. (2017). Computational Health Informatics in the Big Data Age: A Survey. *ACM Computing Surveys*, *49*(1), 1–36.

Farhat, M., Chaabouni-Chouayakh, H., & Ben-Hamadou, A. (2023). Self-supervised endoscopic image key-points matching. *Expert Systems with Applications*, *213*, 118696.

Feng, S. (2022). Vertical federated learning-based feature selection with non-overlapping sample utilization. *Expert Systems with Applications*, *208*, 118097.

Figueira, A., & Vaz, B. (2022). Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, *10*(15). Scopus.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–338.

Gruetzemacher, R., & Paradice, D. (2022). Deep Transfer Learning & Beyond: Transformer Language Models in Information Systems Research. *ACM Computing Surveys*, *54*(10s), 1–35.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., … Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225–250.

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, *29*.

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the Knowledge in a Neural Network* (arXiv:1503.02531). arXiv.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. In P. A. Flach, T. De Bie, & N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 35–50). Springer.

Karst, F. S., Li, M., & Leimeister, J. M. (2023). *Synthesizing Training Data with Generative Adversarial Networks: Towards the Design of a Data-Sharing Ecosystem Platform for Fraud Detection*.

Karst, F. S., Li, M., & Leimeister, J. M. (2024). *FinDEx: A Synthetic Data Sharing Platform for Financial Fraud Detection*.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley & Sons.

Khosravian, A., Amirkhani, A., Kashiani, H., & Masih-Tehrani, M. (2021). Generalizing state-of-the-art object detectors for autonomous vehicles in unseen environments. *Expert Systems with Applications*, *183*, 115417.

Kodinariya, T. M., & Makwana, P. R. (2016). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies*.

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An Update for Taxonomy Designers: Methodological Guidance from Information Systems Research. *Business & Information Systems Engineering*, *64*(4), 421–439.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection: A Data Perspective. *ACM Computing Surveys*, *50*(6), 1–45.

Li, M. M. (2023). *Theorizing a Service Structure*.

Li, M. M., Reinhard, P., Peters, C., Oeste-Reiss, S., & Leimeister, J. M. (2024). A Value Co-Creation Perspective on Data Labeling in Hybrid Intelligence Systems: A Design Study. *Information Systems*, *120*, 102311.

Li, X.-B., & Qin, J. (2013). A Framework for Privacy-Preserving Medical Document Sharing. *Healthcare Information Systems*.

Li, Y., Gooma, A., Li, X. (2024) Responsible Blockchain: STEADI Principles and the Actor-Network Theory-based Development Methodology (ANT-RDM) (arXiv: 2409.06179). arXiv.

Li, Z., Yao, H., & Ma, F. (2020). Learning with Small Data. *Proceedings of the 13th International Conference on Web Search and Data Mining*, 884–887.

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics).

Liu, Y., Tantithamthavorn, C., Li, L., & Liu, Y. (2023). Deep Learning for Android Malware Defenses: A Systematic Literature Review. *ACM Computing Surveys*, *55*(8), 1–36.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). K-NN as an implementation of situation testing for discrimination discovery and prevention. *Proceedings of the 17th ACM SIGKDD,* 502–510.

Ma, F., Gao, J., Suo, Q., You, Q., Zhou, J., & Zhang, A. (2018). Risk Prediction on Electronic Health Records with Prior Medical Knowledge. *Proceedings of the 24th ACM SIGKDD*, 1910–1919.

Mastorakis, G., Ellis, T., & Makris, D. (2018). Fall detection without people: A simulation approach tackling video data scarcity. *Expert Systems with Applications*, *112*, 125–137.

Meurisch, C., & Mühlhäuser, M. (2022). Data Protection in AI Services: A Survey. *ACM Computing Surveys*, *54*(2), 1–38.

Murshed, M. G. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., & Hussain, F. (2022). Machine

Learning at the Network Edge: A Survey. *ACM Computing Surveys*, *54*(8), 1–37.

Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O. A., & Hwang, W.-J. (2021). *Federated Learning for Smart Healthcare: A Survey* (arXiv:2111.08834). arXiv.

Ni, X., Shen, X., & Zhao, H. (2022). Federated optimization via knowledge codistillation. *Expert Systems with Applications*, *191*, 116310.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, *22*(3), 336–359.

Odegua, R. (2020). *Predicting Bank Loan Default with Extreme Gradient Boosting* (arXiv:2002.02011). arXiv.

Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *arXiv*.

Pessach, D., & Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), 51:1-51:44.

Pessach, D., & Shmueli, E. (2023). A Review on Fairness in Machine Learning. *ACM Computing Surveys*, *55*(3), 1–44.

Ramachandram, D., & Taylor, G. W. (2017). Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine*, *34*(6), 96–108. IEEE Signal Processing Magazine.

Raschka, S. (2020). *Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning* (arXiv:1811.12808). arXiv.

Roe, K. D., Jawa, V., Zhang, X., Chute, C. G., Epstein, J. A., Matelsky, J., Shpitser, I., & Taylor, C. O. (2020). Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. *PLOS ONE*, *15*(4), e0231300.

Saito, P. T. M., de Rezende, P. J., Falcão, A. X., Suzuki, C. T. N., & Gomes, J. F. (2014). An active learning paradigm based on a priori data reduction and organization. *Expert Systems with Applications*, *41*(14), 6086–6097.

Schiappa, M. C., Rawat, Y. S., & Shah, M. (2022). Self-Supervised Learning for Videos: A Survey. *ACM Computing Surveys*, 3577925.

Seth, P., & Pai, A. K. (2024). Does the Fairness of Your Pre-Training Hold Up? Examining the Influence of Pre-Training Techniques on Skin Tone Bias in Skin Lesion Classification. *2024 IEEE/CVF WACVW*, 580–587.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., & Meng, D. (2019). Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting. *Advances in Neural Information Processing Systems*, *32*.

Solis, A. (2017). Credit Access and College Enrollment. *Journal of Political Economy*, *125*(2), 562–622. https://doi.org/10.1086/690829

Song, Y., Wang, T., Cai, P., Mondal, S. K., & Sahoo, J. P. (2023). A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities. *ACM Computing Surveys*.

Taiwo, A. (2010). *Types of Machine Learning Algorithms* (Y. Zhang, Ed.). InTech.

Taylor, L., & Nitschke, G. (2018). Improving Deep Learning with Generic Data Augmentation. *2018 IEEE Symposium Series on Computational Intelligence*, 1542–1547.

Tüysüzoğlu, G., & Yaslan, Y. (2018). Sparse coding based classifier ensembles in supervised and active learning scenarios for data classification. *Expert Systems with Applications*, *91*, 364–373.

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2021). A Survey on Distributed Machine Learning. *ACM Computing Surveys*, *53*(2), 1–33.

Wang, J., & Zhou, Z.-H. (2020). Differentially Private Learning with Small Public Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), Article 04.

Wang, P., Li, Y., & Reddy, C. K. (2019). Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys*, *51*(6), 1–36.

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., & Srebro, N. (2017). *Learning Non-Discriminatory Predictors* (arXiv:1702.06081). arXiv.

Wu, J. (2012). Cluster Analysis and K-means Clustering: An Introduction. In J. Wu (Ed.), *Advances in K-means Clustering: A Data Mining Thinking* (pp. 1–16). Springer.

Wydick, W. B. (1999). Credit access, human capital and class structure mobility. *The Journal of Development Studies*.

Xu-Ying Liu, Jianxin Wu, & Zhi-Hua Zhou. (2009). Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550.

Yin, C., Zhao, R., Qian, B., Lv, X., & Zhang, P. (2019). Domain Knowledge Guided Deep Learning with Electronic Health Records. *2019 IEEE International Conference on Data Mining*, 738–747.

Yin, X., Zhu, Y., & Hu, J. (2022). A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions. *ACM Computing Surveys*, *54*(6), 1–36.

Yu, L., Zhang, X., & Yin, H. (2022). An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity. *Expert Systems with Applications*, *202*, 117363.

Zhang, J., Qu, Z., Chen, C., Wang, H., Zhan, Y., Ye, B., & Guo, S. (2022). Edge Learning: The Enabling Technology for Distributed Big Data Analytics in the Edge. *ACM Computing Surveys*, *54*(7), 1–36.

Zhang, Y., & Long, Q. (2021). Assessing Fairness in the Presence of Missing Data. *Advances in Neural Information Processing Systems*, *34*, 16007–16019.