The Journal of Computational Science and Engineering.
ISSN: 2583-9055

Analysis of data acquisition methods for Computational Learning

Usha Muniraju*1, SWETHA N2, Kavya H3 , Nameera Zuha4, Madhushree MG 5, Pratiksha 6

Department of Computer Science Engineering,
East West Institute of Technology
Visvesvaraya Technological University,
Belgaum-50018
Swethan@gmail.com, kavyareddyh03@gmail.com, zuhanameera@gmail.com, madhushreemg8@gmail.com, pratikshahosamani18@gmail.com

Keywords	Abstract
Data Evaluation, data lake, computational learning, Data sources, Data preprocessing, Data sampling, Data labeling, Data annotation, Data quality, Data augmentation, Data imbalance, Data selection, Data storage, Data retrieval	The analysis of data acquisition has significantly transformed the field of computational learning. One crucial aspect of this transformation is the process of data acquisition, which plays a vital role in the field of computational learning models. This study paper provides an essential overview of data acquisition methods in the framework of Big Data-AI integration. We first discuss the importance of data acquisition in computational learning and highlight the issues and openings presented by Big Data. Next, we present a detailed analysis of various data acquisition techniques, including traditional methods, crowdsourcing, and data augmentation. We also examine the effect of data quality, quantity, and diversity on the throughput of computational learning models. Furthermore, we discuss ethical considerations and privacy issues related to data acquisition. Hence, we deduce with future research paths and recommendations for enhancing data acquisition practices in the era of AI integration

Corresponding author: usharaj.m@gmail.com

INTRODUCTION

In recent years, the integration of Big Data and Artificial Intelligence (AI) has revolutionized the field of computational learning, enabling significant advancements in various applications such as image recognition, natural language processing, and predictive analytics. At the core of this transformation lies the procedure of data acquisition, which plays a fundamental role. The measurements and variety of data collected directly affect the throughput and authenticity of these models.

Data acquisition in the notion of computational learning has evolved rapidly with the arrival of Big Data technologies and AI algorithms. Traditional methods of data acquisition, such as manual data preprocessing and surveys, have been augmented and in some cases replaced by more automated and scalable approaches, including web scraping, sensor networks, and crowdsourcing platforms. These innovations have not only facilitated the collection of large volumes of data but also enabled the extraction of valuable insights from various sources such as social media, IoT devices, and online platforms.

Despite the numerous benefits of Big Data-AI integration in computational learning, several challenges and considerations arise in the framework of data gathering. Ensuring the quality and accuracy, addressing privacy and ethical concerns, and handling the throughput and efficiency of data acquisition processes are some of the key obstacles faced by scholars in this field. Moreover, the increasing complexity and diversity of data sources pose additional challenges in terms of data acquisition, preprocessing, and analysis.

This study paper helps to provide you an overview of data acquisition methods with respect to Big Data-AI integration for computational learning. We discuss the importance of data acquisition in computational learning and highlight the complications and scope presented by Big Data. We present adetailed analysis of various data acquisition techniques, including traditional methods, crowdsourcing, and data augmentation. Furthermore, we examine the affect of data quality, quantity, and diversity on the outcome of computational learning models. We also discuss ethical considerations and privacy issues related to data acquisition in the era of Big Data and AI integration.

Overall, this study paper helps to provide valuable insights and recommendations for enhancing data acquisition practices in the field of computational learning, with a focus on the integration of Big Data and AI technologies.

RESEARCH LANDSCAPE OF DATA ACQUISITION FOR COMPUTATIONAL LEARNING

Data acquisition is a critical aspect of computational learning, influencing the quality, diversity, and quantity of data used to train and develop models. This research landscape explores the various features of data acquisition, highlighting contributions from both the computational learning and data management communities. From the perspective of computational learning, data acquisition involves sourcing, processing, and preparing datasets for grounding and examining. Methods such as manual data entry and surveys have been supplemented by automated techniques including web scraping, sensor networks, and crowdsourcing. Computational learning researchers emphasize the importance of data quality, quantity, and diversity in improving model performance and generalization. Alternatively, the data management community brings a wealth of knowledge and techniques to the field of data acquisition. Data management researchers focus on efficient data storage, retrieval, and processing techniques, which are crucial for handling large-scale datasets in computational learning applications. They also focus on issues related to data integration, preprocessing, and cleaning, ensuring that datasets are suitable for computational learning tasks. The intersection of computational learning and data management is marked by several key research areas. For example, data management methods such as indexing, compression, and data summarization can influence the efficiency and scalability of data acquisition for computational learning. Additionally, data privacy, security, and ethical considerations are paramount in both communities, highlighting the value of responsible data acquisition practices.

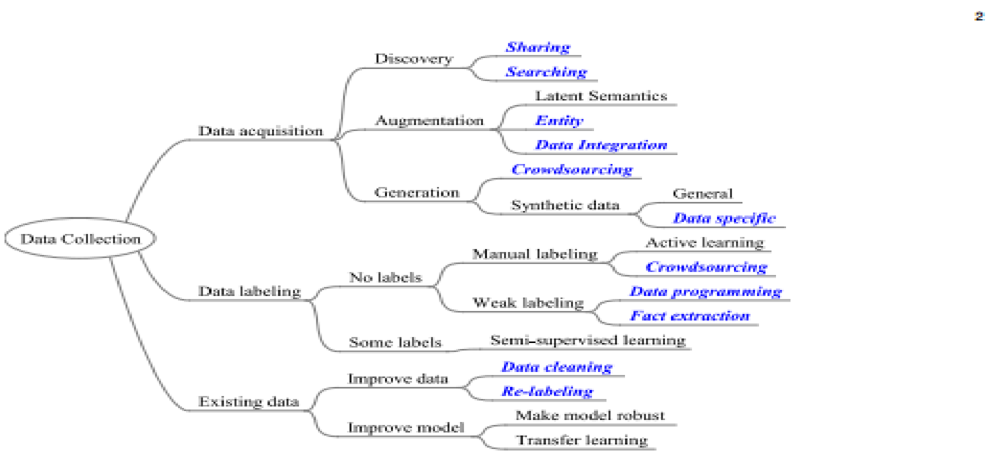

Fig. 1: A high level research landscape of data collection for machine learning. The topics that are at least partially contributed by the data management community are highlighted using blue italic text. Hence, to fully understand the research landscape, one needs to look at the literature from the viewpoints of both the machine learning and data management communities.

To fully understand the research landscape of data acquisition for computational learning, it is required to include the contributions from both the computational learning and data management communities. While computational learning researchers focus on algorithmic advancements and model performance, data management researchers provide valuable insights into data storage, processing, and handling methods that are mandatory for enabling large-scale computational learning applications.

In conclusion, the research landscape of data acquisition for computational learning is multifaceted, with contributions from both the computational learning and data management communities. By integrating perspectives from both fields, researchers can gain a extensive grip of the challenges and liberty in data acquisition for computational learning, leading to advancements in both theory and practice.

The decision structural outline for data acquisition is a valuable tool that helps guide researchers and practitioners through the process of acquiring, labeling, and improving data for computational learning models. The flow chart begins with a fundamental question: "Does Sally have enough data?" This initial step is crucial, as the availability and sufficiency of data directly impact the success of computational learning projects. If the answer to this question is "Yes," the flow chart suggests proceeding to the next step, which is to consider techniques for improving existing data or models. This may involve data augmentation techniques such as duplication, noise addition, or perturbation to enhance the diversity and quality of the dataset.

If the answer to the initial question is "No," the flow chart provides a series of follow-up questions to guide the selection of appropriate techniques for acquiring data. These questions address key considerations such as the availability of labeled data, the feasibility of collecting more data, and the possibility of using alternative data sources. For example, if there is less no of named data, the flow chart suggests exploring techniques such as self-learning or crowdsourcing for data labeling. Self-learning includes repeatedly training a model on a small part of labeled data and then using that model to label more data, gradually expanding the labeled dataset. Crowdsourcing, on the other hand, involves outsourcing data labeling charge to a huge number of workers through online platforms.It is necessary to note that structural outline provides a high-level outline and fails to cover all the details discussed in the survey. For instance, the survey discusses in detail the combination of self-learning and crowdsourcing techniques for data labeling, in addition to the challenges associated with assessing the sufficiency of labels for self-learning. Furthermore, the flow chart acknowledges that some questions, such as "Enough labels for self-learning?" may not have straightforward answers and may require a deep analyzation of the application and data. In such cases, researchers and practitioners are encouraged to delve into the specifics of their particular use case to illuminate decisions. Lastly, the structural outline in the fig.2 highlights that there are techniques specific to distinct types of data, such as images and text, which are detailed in the body of the paper. This emphasizes the importance of considering the characteristics of the data when selecting data acquisition techniques, as distinct types of data may require different approaches.

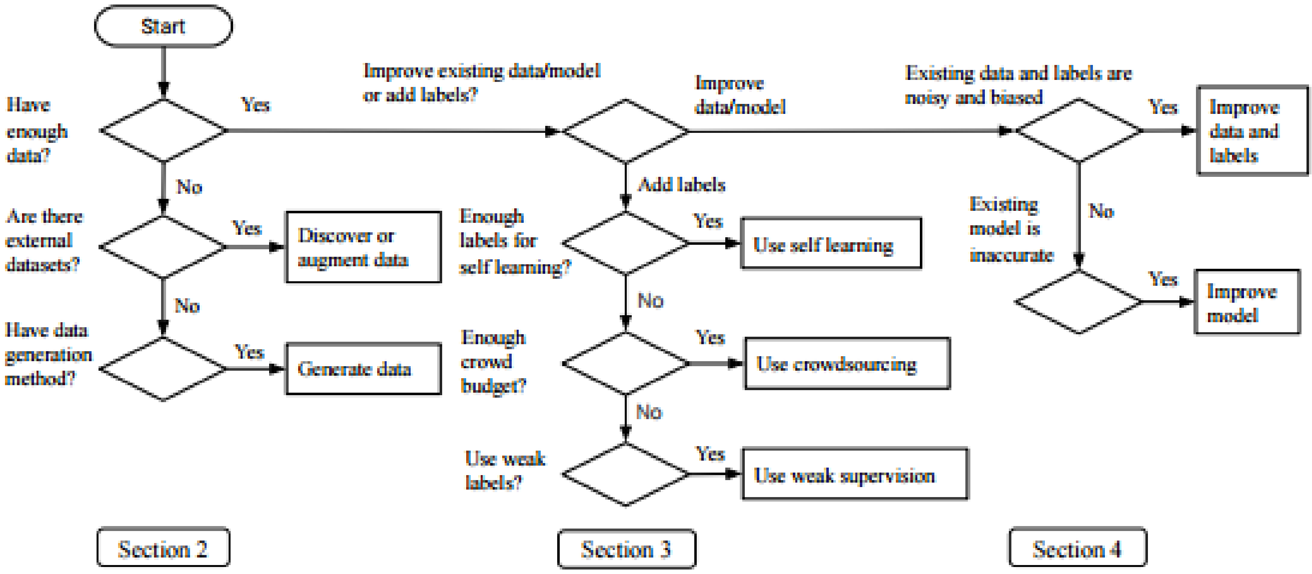

Fig. 2: A decision flow chart for data collection. From the top left, Sally can start by asking whether she has enough data. The following questions lead to specific techniques that can be used for acquiring data, labeling data, or improving existing data or models. This flow chart does not cover all the details in this survey. For example, data labeling techniques like self learning and crowdsourcing can be performed together as described in Section 3.2.1. Also, some questions (e.g., "Enough labels for self learning?") are not easy to answer and may require an in-depth understanding of the application and data. There are also techniques specific to the data type (images and text), which we detail in the body of the paper.

The running example of data acquisition in a smart factory setting illustrates a common challenge faced in computational learning applications: the availability of sufficient and high-quality training data. In this example, the smart factory produces various images of product components, which need to be classified as either normal or defective using a convolutional neural network (CNN) model.

The first step in this process is to collect a dataset of images representing both normal and defective product components. This dataset serves as the training data for the CNN model. However, in many cases, especially in specialized or niche applications like defect detection in manufacturing, finding enough data for training can be challenging.

One approach to address this challenge is data augmentation. Data augmentation techniques involve creating new training examples by applying transformations such as rotation, flipping, scaling, and cropping to existing images. These augmented images can help increase the diversity of the dataset and improve the robustness of the model.

Another approach is to use transfer learning. Transfer learning involves using a pre-trained CNN model (e.g., trained on a large dataset like ImageNet) as a starting point and fine-tuning it on the limited dataset of product component images. This approach leverages the knowledge learned by the pre-trained model and can be effective in situations where there is limited training data available.

Additionally, active learning can be employed to selectively label the most informative examples for training. In this approach, the model is initially trained on a small labeled dataset, and then it iteratively selects the most uncertain or informative examples for labeling by a human expert. This process can help prioritize the labeling of data that is most beneficial for improving the model's performance.

Overall, the example of data acquisition in a smart factory highlights the importance of innovative approaches to address the challenge of limited training data in computational learning applications. By leveraging techniques such as data augmentation, transfer learning, and active learning, it is possible to overcome the data scarcity problem and develop effective computational learning models for defect detection and other manufacturing applications.

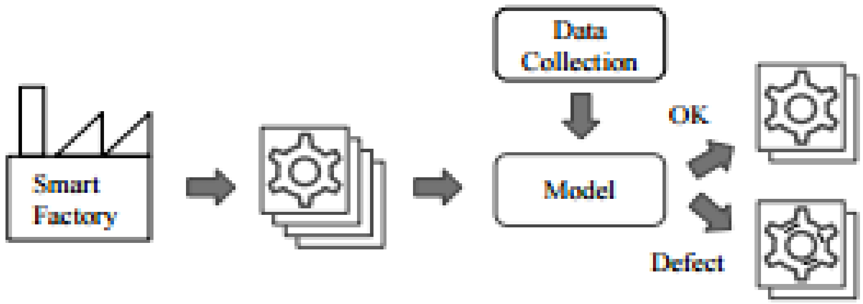

Fig. 3: A running example for data collection. A smart factory may produce various images of product components, which are classified as normal or defective by a convolutional neural network model. Unfortunately, with an application this specific, it is often difficult to find enough data for training the model.

LITERATURE SURVEY

An overview of this studyis provided in the tabular format below, providing a comprehensive overview of relevant research works. The table encompasses crucial details such as the name of the study, author(s), publication year, research objectives, and key advantages and disadvantages identifier

Title	Authors	Year	Objectives	Advantages	Disadvantages
Leveraging Network Data Analytics Function andComputational learning for Data acquisition, ResourceOptimization, Security and Privacyin 6G Networks	PANAGIOTIS K. GKONIS 1,NIKOLAOS NOMIKOS 2, IEEE), PANAGIOTIS TRAKADAS 2, LAMBROS SARAKIS 1, GEORGE XYLOURIS 3, XAVI MASIP-BRUIN4, JOSEP MARTRAT	2024	1. Developing a Framework: Createa comprehensive framework that integrates network data analytics functions and computational learning algorithms for data acquisition, resource optimization, security, and privacy enhancement in 6G networks. 2. Enhancing Data acquisition: Propose novel techniques to improve the efficiency and accuracy of data acquisition in 6G networks using	1.ImprovedNetwork Performance: By leveraging network data analytics and computational learning, 6G networks can achieve higher levels of performance, including faster data acquisition, more efficient resource allocation, and enhanced security measures. 2.EnhancedUserExperience: The use of data analytics and computational learning in 6G networks can lead to	1.Complexity:Implementing network data analytics and computational learning in 6G networks can add complexity to the network architecture, requiring specialized expertise and potentially increasing the risk of system failures or errors. 2. Resource Intensive: Data analytics and computational learning algorithms can be

			network data analytics and computational learning approaches.	improved user experience by optimizing network resources based on user behavior and preferences.	computationally intensive, requiring significant resources such as processing power and memory, which could lead to increased costs for network operators.
Computational learning Model Generation withCopula-Based Synthetic Dataset for LocalDifferentially Private Numerical Data	QYUICHI SEI 1,2, (Member, IEEE), J. ANDREW ONESIMU , AND AKIHIKO OHSUGA 1(Member, IEEE)	2022	1 Developing a Copula-Based Synthetic Dataset: Create a method to generate synthetic datasets using copulas, which preserve the statistical properties of the original data while ensuring local differential privacy for numerical data. 2.Computational learning Model Generation: Delve into the training of computational learningmodels. using copula-based synthetic datasets to achieve comparable performance to models trained on original data,while preserving privacy.	1.Privacy Preservation: The use of copula-based synthetic datasets ensures local differential privacy for numerical data, protecting certain information while allowing for meaningful analysis. 2.Data Utility: The synthetic datasets generated using copulas preserve the statistical properties of the original data, enabling computational learning models trained on these datasets to achieve comparable performance to models trained on original data.	1.Loss of Information: The process of generating synthetic datasets using copulas may lead to some loss of information compared to the source data, which could impact the throughput of computational learning models trained on these datasets. 2.Complexity: Implementing copula-based synthetic dataset generation and local differential privacy mechanisms may add complexity to the data preprocessing and model training pipeline, requiring additional computational resources and expe rtise.
Title	**Authors**	**Year**	**Objectives**	**Advantages**	**Disadvantages**

Advancing Aviation Safety Through Computational learning and Psychophysi ologicalData :A Systematic Review	IBRAHIM ALRESHI DI 1,2,3, IRENE MOULITS AS 1,2, AND KARL W. JENKINS1	2023	1.Identification of Trends and Patterns: Identify trends and patterns in the application of computational learning and psychophysiological data analysis techniques in aviation safety research. 2.Evaluation of Methodologies: Evaluate the methodologies used in existing studies, including data acquisition methods, computational learning algorithms, and psychophysiological data analysis techniques.	1.EnhancedSafetyMeasures: The systematic review can provide insights into how computational learning and psychophysiological data analysis can enhance existing safety measures in aviation, leading to a safer aviation environment. 2.ImprovedRisk Prediction: By synthesizing existing research, the review can identify patterns and trends that improve the prediction of safety-critical events in aviation, allowing for proactive risk management.	1.Data Quality Concerns: Many studies in the field may suffer from data quality issues, such as small sample sizes, incomplete data, or biased datasets, which could affect the reliability of the findings. 2.Algorithmic Limitations: Computational learning algorithms used in aviation safety research may have limitations in terms of accuracy, interpretability, or scalability, which could impact their effectiveness inreal-world applic ations.
Prioritization of Mobile IoT Data Transmission Based on Data Importance Extracted From Computation al learning Model	YUICHI INAGAKI , RYOICHI SHINKUM A , TAKEHIR O SATO , AND EIJI OKI	2019	1.Development of a Computational learning Model: Develop a computational learning model capable of extracting data importance from mobile IoT devices to prioritize data transmission based on its significance. 2.Data Importance Metrics: Define metrics for measuring the importance of mobile IoT data, considering factors such as data relevance, timeliness,	1. Improved Efficiency: Prioritizing mobile IoT data transmission based on data importance can lead to improved efficiency in data delivery, ensuring that critical data is transmitted in a timely manner while reducing unnecessary data transmission. 2.Enhanced Resource Utilization: By prioritizing data transmission,	1.Complexity: Implementing a computational learning model to extract data importance and prioritize data transmission adds complexity to the mobile IoT network architecture, requiring specialized expertise and potentially increasing the risk of system failures or errors. 2.Resource

			and impact on decision-making processes.	mobile IoT devices can optimize their use of network resources, such as bandwidth and energy, leading to improved overall network performance	Intensive: Computational learning algorithms used for data importance extraction can be computationally intensive, requiring significant resources such as processing power and memory, which could lead to increased costs for IoT device manufacturers or network operators 1

Title	Authors	Year	Objectives	Advantages	Disadvantages
Time-Series Data Classification and Analysis Associated With Computational learning Algorithms for Cognitive Perception and Phenomenon	TAIKYEONG JEONG , (Senior Member, IEEE)	2020	1.Methodological Framework Development: Propose a robust methodological framework for handling time-series data in the context of cognitive perception and phenomenon. This framework should encompass data preprocessing, feature extraction, model selection, and evaluation methodologies tailored to the intricacies of time-series data. 2.Algorithm Selection and Optimization: Investigate and select appropriate computational learning algorithms for time-series data classification and analysis, considering their suitability for cognitive perception tasks. Additionally, optimize these algorithms to enhance their performance and efficiency in handling complex temporal patterns inherent in cognitive data.	1.Enhanced Accuracy: Computational learning algorithms applied to time-series data can significantly improve classification accuracy compared to traditional methods. This is crucial in domains where precision is paramount, such as medical diagnosis or financial forecasting. 2.Real-Time Insights: By leveraging computational learning techniques, real-time insights can be gleaned from time-series data, allowing for timely decision-making and intervention. This is particularly beneficial in applications like anomaly detection in network traffic or predicting equipment failures in industrial settings.	1.Complexity of Implementation: Time-series data analysis coupled with computational learning algorithms can be intricate to implement, requiring a deep understanding of both domains. Researchers might face challenges in accurately applying these methodologies, leading to potential errors or misinterpretations. 2.Data Preprocessing Overhead: Time-series data often require extensive preprocessing to handle missing values, outliers, noise, and temporal misalignments. This preprocessing overhead can be significant, potentially consuming a substantial amount of computational resources and time.

ANonlinear Regression Application via Machine Learning Techniques for Geomagnetic Data Reconstruction Processing	Huan Liu , Member, IEEE, Zheng Liu, Senior Member, IEEE, Shuo Liu, Student Member, IEEE, Yihao Liu, Junchi Bin, Fang Shi, and Haobin Dong	2019	1.Developing Accurate Predictive Models: One objective could be to develop and validate computational learning-based predictive models capable of accurately reconstructing geomagnetic data. This involves training models that can capture the complex nonlinear relationships inherent in geomagnetic phenomena. 2.Exploring Nonlinear Regression Techniques: Another objective could be to explore and compare various nonlinear regression techniques within the computational learning domain. This involves experimenting with algorithms such as support vector machines (SVM), random forests, neural networks, and Gaussian processes to identify the most suitable approach for geomagnetic data reconstruction.	1.Improved Accuracy: Computational learning techniques, particularly nonlinear regression models, can offer improved accuracy in reconstructing geomagnetic data compared to traditional methods. These techniques can capture complex relationships and patterns present in the data, leading to more accurate predictions. 2.Flexibility and Adaptability: Computational learning models can adapt to various types of geomagnetic data and environmental conditions. They can handle nonlinear relationships and complex interactions between different variables, providing a more flexible framework for data reconstruction processing.	1.Limited Generalizability: Computational learning models trained on specific geomagnetic datasets may have limited generalizability to different geographical locations or time periods. This could restrict the applicability of the proposed techniques to broader contexts or real-world scenarios. 2.Data Quality and Reliability: Geomagnetic data can be susceptible to various sources of noise, artifacts, and biases, which may adversely affect the performance of computational learning algorithms. Ensuring the quality and reliability of the input data is crucial for obtaining accurate reconstruction results.
Title	**Authors**	**Year**	**Objectives**	**Advantages**	**Disadvantages**

Lessons from archives: Strategies for collecting sociocultural data in computational learning	Eun Seo Jo, Timnit Gebru	2020	1.Highlight the Importance of Ethical Data acquisition: Emphasize the significance of ethical considerations in gathering sociocultural data for computational learning applications, drawing from the insights provided in the referenced paper. 2.Discuss Methodological Strategies: Outline and analyze the methodological strategies proposed in the referenced paper for collecting sociocultural data, including archival research techniques and community engagement approach es.	1. BroaderContext: Incorporating insights from the referenced paper can provide a broader contextual understanding of sociocultural data acquisition in computational learning. This helps readers situate the specific research within a larger framework, making it more relevant and insightful. 2.Methodological Guidance: The referenced paper likely offers methodological guidance and best practices for collecting sociocultural data, which can enhance the rigor and validity of the research presented in the IEEE conference paper. This can include strategies for data acquisition, handling bias, and ensuring inclusivity.	1.Copyright Issues: Reproducing or paraphrasing significant portions of another paper without proper citation and permission from the authors can lead to copyright infringement issues. It's essential to ensure that proper credit is given to the original authors and that any reused content adheres to fair use principles. 2.Divergent Focus: The content of "Lessons from archives" may not align perfectly with the focus or theme of the IEEE conference paper. Introducing material that diverges from the main topic could confuse readers and dilute the clarity and coherence of the paper's argument.

A survey on data acquisition for computational learning: a big data-ai integration perspective	Yuji Roh, Geon Heo, Steven Euijong Whang	2019	1.Provide a comprehensive overview of the current landscape of data acquisition methods for computational learning, with a specific focus on integrating big data and AI technologies. 2.Identify key challenges and opportunities in data acquisition for computational learning within the context of big data and AI integration.	1.Comprehensive Literature Review: Integrating content from the survey paper into an IEEE conference paper can enhance the literature review section by providing a comprehensive overview of data acquisition methods for computational learning within the context of big data and AI integration. This can demonstrate a deep understanding of existing research in the field. 2.Methodological Insights: The survey paper likely discusses various methodologies and techniques used in data acquisition for computational learning. Incorporating these insights into the conference paper can enrich the methodological framework, providing a more robust approach for collecting and processing data for the research being presented.	1.Plagiarism Concerns: Directly incorporating content from another paper without proper citation and acknowledgment could lead to accusations of plagiarism, which is a serious ethical violation in academia. 2.Copyright Issues: Reproducing content from a copyrighted paper without obtaining permission from the copyright holder (usually the publisher) can result in legal consequences.

Title	Authors	Year	Objectives	Advantages	Disadvantages
Computational learning model towards evaluating data gathering methods in manufacturing and mechanical engineering	Mahyar Amini, Koosha Sharifani, Ali Rahmani	2023	1.Introduce the Proposed Computational learning Model: Present the architecture and design of the computational learning model tailored for evaluating data gathering methods in manufacturing and mechanical engineering. Detail the underlying algorithms, techniques, and methodologies employed in the model's development. 2.Data Preparation and Feature Engineering: Describe the process of data preparation and feature engineering, including data acquisition methods, data preprocessing techniques, and feature selection strategies. Emphasize the importance of representative datasets and feature engineering in enhancing the model's performance and generalizability.	1.Enhanced Methodology: The paper's methodology for evaluating data gathering methods using computational learning can enrich the methodology section of the IEEE conference paper. This could provide a more comprehensive approach to data acquisition and analysis, especially for researchers in manufacturing and mechanical engineering fields. 2.Validation and Comparison: If the IEEE conference paper aims to propose or evaluate data gathering methods, incorporating the results from the mentioned paper can serve as a validation or comparison. This adds credibility to the proposed methods by showing how they perform relative to existing approaches.	1.Copyright Issues: Reproducing content from another publication without proper permission or acknowledgment could lead to copyright infringement issues, potentially resulting in legal consequences or rejection of the conference paper. 2.Plagiarism Concerns: Simply copying text or ideas from another source without proper citation or attribution is considered academic misconduct and can damage the credibility of the authors and their wo rk.

The Science of Data acquisition: Insights from Surveys can Improve Computational learning Models	Stephanie Eckman, Barbara Plank, Frauke Kreuter	2024	1.Survey Design: Develop guidelines for designing surveys that collect data relevant to computational learning model improvement, considering factors such as question types, response formats, and sample sizes. 2. Data Preprocessing: Investigate methods for preprocessing survey data to ensure its quality and compatibility with computational learning algorithms, such as handling missing values, encoding categorical variables, and removing outliers.	1.EnhancedModel Performance: By incorporating insights from surveys, computational learning models can be trained on more relevant and informative data, leading to improved performance in prediction and classification tasks. 2.ImprovedData Understanding: Surveys can provide valuable insights into the underlying patterns and relationships in the data, helping to identify important features and variables that can enhance computational learning models.	1.Bias in Survey Responses: Survey responses may be biased due to factors such as respondent demographics, survey wording, or response format, which could lead to biased insights and potentially biased computational learning models. 2.Limited Generalizability: Survey data may have limited generalizability to the broader population or context, as it is based on a specific sample of respondents and may not capture the full range of variability in the dat a.

METHODOLOGY

Data Acquisition Techniques: Utilize data discovery methods to find relevant datasets for training computational learning models. Consider data augmentation strategies to enhance existing datasets with external data
Data Labeling: Explore options like crowdsourcing, self-learning, or weak supervision for labeling data when needed Improving Existing Data: Focus on techniques for enhancing data quality through cleaning, especially with a computational learning application perspective
Decision Flow Chart: Develop a decision flow chart that guides the selection of data acquisition techniques, considering aspects like data acquisition, labeling, and improving existing data
Interdisciplinary Approach: Acknowledge that data acquisition techniques stem from various disciplines like computational learning, natural language processing, computer vision, and data management. Understanding this interdisciplinary landscape is crucial for informed decision-making
Future Research Challenges: Identify and address future research challenges in data acquisition for computational learning to contribute to the advancement of the field

CONCLUSION

In this survey, we have explored the critical role of data acquisition in the integration of Big Data and AI for computational learning. We have highlighted the challenges and opportunities in data acquisition, emphasizing the importance of high-quality, diverse, and labeled datasets. Additionally, we have discussed various techniques and strategies for effective data acquisition, including data augmentation, federated learning, and differential privacy.
Our survey underscores the need for a holistic approach to data acquisition that considers not only the technical aspects but also the ethical and legal implications. As AI continues to advance, the quality and quantity of data will play an increasingly significant role in the success of computational learning models.
By understanding the key trends and challenges in data acquisition, researchers and practitioners can develop more robust and effective data acquisition strategies, leading to improved AI systems and applications. We hope that this survey serves as a valuable resource for the research community and contributes to the ongoing advancement of AI technologies.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Computational learning Research, vol. 3, pp. 993–1022,2003.
[2] A. Kumar, J. Naughton, J. M. Patel, and X. Zhu, "To join or notto join? Thinking twice about joins before feature selection," inSIGMOD, 2016, pp. 19–34.
[3] V. Shah, A. Kumar, and X. Zhu, "Are key-foreign key joins safe toavoid when learning high-capacity classifiers?" PVLDB, vol. 11,no. 3, pp. 366–379, Nov. 2017.
[4] M. Stonebraker and I. F. Ilyas, "Data integration: The currentstatus and the way forward," IEEE Data Eng. Bull., vol. 41, no. 2, pp. 3–9, 2018.
[5] A. Doan, A. Y. Halevy, and Z. G. Ives, Principles of Data Integration. Morgan Kaufmann, 2012.
[6] S. Li, L. Chen, and A. Kumar, "Enabling and optimizing non-linear feature interactions in factorized linear algebra," in SIGMOD, 2019, pp. 1571–1588.
[7] L. Chen, A. Kumar, J. F. Naughton, and J. M. Patel, "Towardslinear algebra over normalized data," PVLDB, vol. 10, no. 11, pp.1214–1225, 2017.
[8] A. Kumar, J. F. Naughton, J. M. Patel, and X. Zhu, "To join or notto join? Thinking twice about joins before featureselection," inSIGMOD, 2016, pp. 19–34.
[9] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller, "Turkit:Human computation algorithms on mechanical turk," in UIST,2010, pp. 57–66.
[10] D. W. Barowy, C. Curtsinger, E. D. Berger, and A. McGregor, "Automan: A platform for integrating human-based and digitalcomputation," in OOPSLA, 2012, pp. 639–654.
[11] S. Ahmad, A. Battle, Z. Malkani, and S. Kamvar, "The jabberwocky programming environment for structured social

computing," in UIST, 2011, pp. 53–64.

[12] H. Park, R. Pang, A. G. Parameswaran, H. Garcia-Molina, N. Polyzotis, and J. Widom, "Deco: A system for declarative crowdsourcing," PVLDB, vol. 5, no. 12, pp. 1990–1993, 2012.

[13] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin,"Crowddb: Answering queries with crowdsourcing," in SIGMOD, 2011, pp. 61–72.

[14] A. Marcus, E. Wu, S. Madden, and R. C. Miller, "Crowdsourceddatabases: Query processing with people," in CIDR, 2011, pp.211–214.

[15] R. Boim, O. Greenshpan, T. Milo, S. Novgorodov, N. Polyzotis,and W. C. Tan, "Asking the right questions in crowd datasourcing," in ICDE, 2012, pp. 1261–1264.

[16] Prasad N. Achyutha, Sushovan Chaudhury, Subhas Chandra Bose, Rajnish Kler, Jyoti Surve, Karthikeyan Kaliyaperumal, "User Classification and Stock Market-Based Recommendation Engine Based on Machine Learning and Twitter Analysis", Mathematical Problems in Engineering, vol. 2022, Article ID 4644855, 9 pages, 2022. https://doi.org/10.1155/2022/4644855