EXPLORE
Jurnal Sistem Informasi dan Telematika
(Telekomunikasi, Multimedia dan Informatika)

Utilizing Machine Learning in Big Data Visualization: A Systematic Review

Ade Kurniawan, Aldi Rosyid, Ade Rudi Masa'id, Wendi Usino
Jurusan Ilmu Komputer, Fakultas Teknologi Informasi
Universitas Budi Luhur
Jakarta, Indonesia
ade.curniawan@gmail.com, aldirosyid@gmail.com
aderudimasaid1@gmail.com, wendi.usino@budiluhur.ac.id

Abstract- This systematic review examines the existing literature on data visualization approaches for big data, aiming to identify key challenges, emerging trends, and promising solutions. The review highlights the growing importance of data visualization in extracting meaningful insights from large and complex datasets. It underscores the unique challenges posed by the scale and multidimensionality of big data, emphasizing the need for innovative visualization techniques. Through a systematic search and analysis of peer-reviewed articles published between 2023 and 2024, focusing on the integration of machine learning in big data visualization, this paper explores various data visualization methods currently employed in big data analysis, discussing their strengths, limitations, and potential areas for future development. Findings reveal that machine learning and deep learning significantly enhance the effectiveness of big data visualization. Specifically, techniques like knowledge graph embedding with Convolutional Neural Networks and hybrid models combining t-SNE, PCA, and QR Decomposition show promise in handling complex medical and high-dimensional datasets respectively. The review also identifies a need for more standardized evaluation frameworks and further exploration of emerging technologies like virtual and augmented reality in this field. This review concludes by emphasizing the importance of continued research to improve the accessibility and effectiveness of big data visualization for a wide range of applications.

Keyword: Data Visualization, Automation, Machine Learning, Data Visualization Machine Learning

Abstrak - Tinjauan sistematis ini mengkaji literatur yang ada mengenai pendekatan visualisasi data untuk *big data*, dengan tujuan mengidentifikasi tantangan utama, tren yang muncul, dan solusi yang menjanjikan. Tinjauan ini menyoroti semakin pentingnya visualisasi data dalam mengekstraksi wawasan yang bermakna dari kumpulan data yang besar dan kompleks. Ini menggarisbawahi tantangan unik yang ditimbulkan oleh skala dan multidimensionalitas *big data*, menekankan perlunya teknik visualisasi yang inovatif. Melalui pencarian dan analisis sistematis artikel tinjauan sejawat yang diterbitkan antara tahun 2023 dan 2024, dengan fokus pada integrasi pembelajaran mesin dalam visualisasi *big data*, makalah ini mengeksplorasi berbagai metode visualisasi data yang saat ini digunakan dalam analisis *big data*, membahas kekuatan, keterbatasan, dan potensi area untuk pengembangan di masa depan. Temuan menunjukkan bahwa pembelajaran mesin dan pembelajaran mendalam secara signifikan meningkatkan efektivitas visualisasi *big data*. Secara khusus, teknik seperti penyematan grafik pengetahuan (*knowledge graph embedding*) dengan *Convolutional Neural Networks* dan model hibrida yang menggabungkan t-SNE, PCA, dan Dekomposisi QR menunjukkan potensi dalam menangani *dataset* medis yang kompleks dan *dataset* berdimensi tinggi secara berturut-turut. Tinjauan ini juga mengidentifikasi kebutuhan akan kerangka kerja evaluasi yang lebih terstandardisasi dan eksplorasi lebih lanjut terhadap teknologi yang muncul seperti realitas virtual dan *augmented reality* di bidang ini. Tinjauan ini diakhiri dengan menekankan pentingnya penelitian berkelanjutan di bidang ini untuk meningkatkan aksesibilitas dan efektivitas visualisasi *big data* untuk berbagai aplikasi.

Keyword: Data Visualization, Automation, Machine Learning, Data Visualization Machine Learning

1. Introduction

Data visualization has emerged as a crucial tool across diverse disciplines, from scientific inquiry to business intelligence. For instance, in scientific research, complex genomic data can be visualized to identify patterns in disease progression, while in business, interactive dashboards illustrate sales trends and customer behavior for strategic decision-making. The exponential escalation in data volume and complexity has rendered the capacity to effectively interpret and convey insights an indispensable skill across a broad spectrum of domains

[1]. This systematic review aims to synthesize the existing literature on data visualization approaches, examining the key trends, methodologies, and applications of this critical field. Effective data visualization enables the extraction of meaningful insights from vast, complex datasets, transforming raw information into impactful narratives that can inform critical decision-making, drive innovation, and foster deeper understanding across a range of contexts [2]. Given the sheer scale and multidimensional nature of big data, visualizing such mammoth datasets poses unique challenges that require innovative, tailored approaches. Researchers have long grappled with the complexities of big data visualization, exploring new and innovative techniques to make sense of the vast troves of information that define the contemporary data landscape [3][4]. These complexities often manifest as information overload due to the immense volume, inconsistent data quality stemming from diverse sources leading to inaccuracies, and the difficulty in representing intricate data relationships with numerous interdependencies. Furthermore, the high velocity of real-time data processing demands continuous updates and significant computational resources, while inherent perceptual and cognitive limitations of the human visual system present hurdles like over-plotting. The variety of data types and the critical need to ensure data trustworthiness (veracity) add further layers of difficulty. In summary, the challenges of big data visualization primarily stem from the overwhelming volume, diverse and often inconsistent nature of the data, coupled with the need to accurately represent complex relationships and handle real-time processing within human perceptual limits.

2. Methodology

This systematic review examines the current state of the literature on data visualization methods for big data, with the goal of identifying the key challenges, emerging trends, and promising solutions in this field.

A. Research Objectives

The primary objective of this systematic review is to provide a comprehensive analysis of the existing literature on data visualization approaches for big data, with a specific focus on those integrating machine learning techniques. This review aims to systematically assess the strengths, limitations, and potential areas for future development within this critical and evolving field, thereby offering a structured overview of the current state-of-the-art and identifying emerging research directions.

B. Research Question

The key research question addressed in this review is:

RQ1: What are the current data visualization approaches, including those leveraging machine learning, used for big data analysis, what are the unique challenges associated with big data visualization, and what are the associated strengths, limitations, and emerging trends in this field?

RQ2: How can these data visualization approaches be further improved or adapted to better address the unique challenges of big data?

RQ3: What are the potential future directions and areas of development in the field of big data visualization?

C. Quality Assessment

Each included study was assessed for methodological quality and relevance to the research questions. Relevance was determined by rigorous application of predefined inclusion and exclusion criteria, focusing on peer-reviewed English-language journal articles published between January 2023 and 2024 that specifically leveraged machine learning for big data visualization. While a standardized evaluation framework was utilized for methodological quality assessment, this review specifically focused on studies that met these criteria to ensure their direct applicability to the research objectives.

D. Research Method

This systematic review examines the current state of the literature on data visualization methods for big data, with the goal of identifying the key challenges, emerging trends, and promising solutions in this field. This review was conducted following a systematic literature review (SLR) framework, with phases adapted from Tranfield et al. [5] and adhering to principles of the PRISMA 2020 statement, to ensure a comprehensive and rigorous analysis of the existing body of knowledge.

Fig. 1. Research Flow

Table 1. Review Phases: adapted from Tranfield et al. [5]

Phase	Phase Description
Planning	• Keywords and search strings:("Machine Learning" OR "ML" OR "Automation Data Visualization") AND ("Big Data Visualization" OR "Visualization Method" OR "Visualization Analytics") • This source will use IEEE database only • Limiting the scope to titles, abstracts, and keywords • Focusing on publications from 2023 to 2024 • Including only English-language • Peer-reviewed journal articles.
Implementation	• Conducting a systematic search of the specified databases using the defined keywords and inclusion criteria

EXPLORE: ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v17i1.3818

	• Refining the found publications b removing any duplicate entries • Screening the remaining publications against the established inclusion and exclusion criteria
Reporting	The findings are reported in two main sections: • Descriptive analysis • Thematic analysis

Table 2. Inclusion and Exclusion Criteria

Inclusion criteria	Exclusion criteria
• Peer-reviewed journ articles and research • English Language • Published from January 2023 to 202 • Leveraging machine learning for data visualization in the paper topic • ML in the big data visualization context	• Book, conference papers, • theses • Systematic Literature review and bibliometric analysis studies • ML is not main topic in the paper • ML outside of big data visualization scope

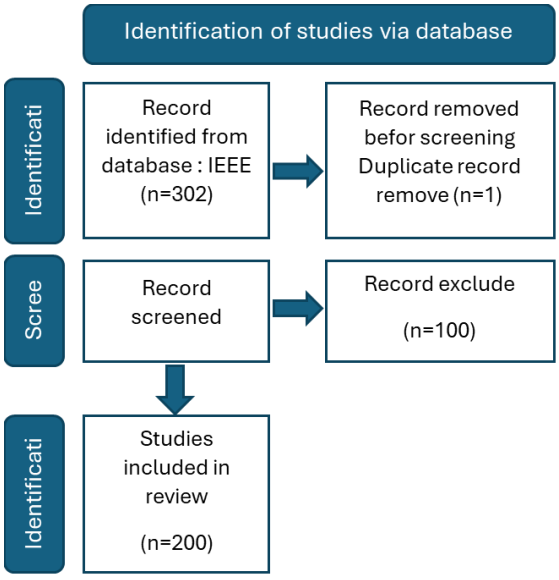

Fig. 2. Records Screening Phases: adapted from page et al. [6]

3. Result and Discussion

A. Results

T-SNE	Integrated Modelling	Google Studio	Neural Network	Graph Decoder
9	19	2	164	8

Utilizes a deep learning-based approach for medical data visualization. Specifically, it leverages a combination of: **Knowledge graph embedding Convolutional Neural Networks.** By integrating these techniques, the authors aim to create a visualization algorithm capable of handling the complexities of medical big data and supporting more effective disease diagnosis. Tested their proposed visualization algorithm on various medical text datasets. They found that their method achieved a sensitivity of 0.94 and an accuracy of 0.87. Essentially, this means the algorithm demonstrated strong performance in correctly identifying and classifying diseases based on the provided textual descriptions. The authors concluded that combining knowledge graphs and deep learning is a promising approach for improving the initial diagnosis of diseases. [7]

Proposes a hybrid model using a combination of methods for data visualization: **t-SNE (t-Distributed Stochastic Neighbor Embedding):** A machine learning algorithm for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map. **PCA:** A linear algebra technique used for dimensionality reduction, often employed as a preprocessing step before applying t-SNE. **QR Decomposition Algorithm:** A numerical linear algebra method used to factorize a matrix into an orthogonal matrix (Q) and an upper triangular matrix (R). In this paper, it's specifically used to determine eigenvalues and eigenvectors within the PCA process. Based on combining this method in summary, the authors aim to leverage the strengths of each to create a more effective visualization tool for complex datasets. They claim that their hybrid model, combining PCA, t-SNE, and QR decomposition, leads to more effective data visualization than using PCA and t-SNE alone. They support this by: **Visualizations:** Presenting figures (though without specific details here) showing the output of their model, suggesting improved clarity in how data clusters are represented. **Model Accuracy Metrics:** They provide results for training accuracy, testing accuracy, MSE, and AUC scores, indicating good performance of their model. This research introduces an innovative hybrid model designed to enhance the visualization of complex high-dimensional data, integrating **t-SNE (t-Distributed Stochastic Neighbor Embedding)**, **PCA (Principal Component Analysis)**, and the **QR Decomposition Algorithm**. t-SNE, a powerful non-linear machine learning algorithm, is utilized for its capacity to effectively map high-dimensional data points into a lower-dimensional space (two or three dimensions), thereby preserving crucial local data structures and revealing hidden clusters. PCA, a foundational linear algebra technique, serves as a crucial preprocessing step for dimensionality reduction, helping to mitigate noise and simplify data complexity before t-SNE application. Uniquely, the QR Decomposition Algorithm is specifically employed within the PCA process to efficiently determine the eigenvalues and eigenvectors, which are fundamental for identifying the principal components that capture the maximum variance in the data. By combining these three methods, the authors

aim to leverage their respective strengths: PCA's efficiency in initial dimension reduction, QR Decomposition's precision in PCA's eigen-calculations, and t-SNE's prowess in visualizing intricate local structures. The primary claim is that this proposed hybrid model yields more effective data visualizations compared to using PCA and t-SNE alone. This assertion is supported by several lines of evidence, including presented figures that visually suggest improved clarity in how data clusters are represented, a range of promising accuracy metrics (training accuracy, testing accuracy, MSE, and AUC scores) indicating strong model performance, and a detailed analysis of precision, recall, F1-score, and overall accuracy derived from a confusion matrix, which collectively imply the model's effectiveness in classification. Furthermore, the study employs a two-sample t-test for hypothesis testing to statistically compare their approach against a standard PCA-t-SNE method. However, despite these various results pointing towards the hybrid model's efficacy, a direct and quantitatively measurable performance comparison with existing standalone PCA-t-SNE models is not explicitly provided, as the reported metrics for the hybrid model are not consistently contrasted with baseline results. Similarly, while hypothesis testing is mentioned, the specific details of its outcomes, such as p-values, effect sizes, or confidence intervals, are not furnished, which hinders a comprehensive assessment of the statistical significance of the claimed improvements. To robustly validate the claimed superiority of the hybrid model, the paper would significantly benefit from a clear, tabular, or graphical presentation directly comparing all key performance metrics (including accuracy, MSE, AUC, precision, recall, and F1-score) between the hybrid model and a well-defined baseline (e.g., standalone PCA-t-SNE), coupled with a more detailed quantitative interpretation of the hypothesis testing results. [8]

Utilizes several methods to achieve its research objectives. Some of the key methods mentioned include:

Data Mining: The authors use a data mining model that combines **network decomposition** and **symptom combination**. This helps them analyze a large dataset of biomedical literature and extract relevant information about coronary heart disease. **sMIL Algorithm:** The paper employs a **sparse multi-instance learning algorithm (sMIL)** to identify potential genetic relationships within the collected literature abstracts. This method is chosen for its ability to work with limited labeled data compared to traditional supervised learning approaches. **Dictionary Matching and Tokenization:** These techniques are used for identifying gene entities within the text. **Network Analysis:** The paper mentions using network analysis techniques to study the complex relationships between genes. By combining these methods, the authors aim to achieve a more accurate and comprehensive understanding of the genetic basis of coronary heart disease. While the authors don't present a singular, definitive result, we can summarize the key contributions and potential implications of their work: **The paper proposes a novel method for analyzing large datasets of biomedical literature related to classic and famous prescriptions.** This method combines data mining, network analysis, and visualization techniques to uncover potential genetic relationships associated with complex diseases like coronary heart disease. **The authors demonstrate the feasibility of their approach by applying it to a real-world dataset.** They showcase the ability to identify relevant genes, categorize symptoms, and visualize gene networks, offering valuable insights into the genetic underpinnings of coronary heart disease. **The paper lays the groundwork for future research in the field.** By introducing this intelligent visualization method, the authors pave the way for more comprehensive and accurate analysis of complex medical data, potentially leading to new discoveries and treatment strategies. Essentially, the paper doesn't offer definitive answers but presents a promising new approach for analyzing complex medical data, with the potential to advance our understanding and treatment of diseases.[9]

a case study where the authors aimed to solve a sales problem for a virtual corporation. They meticulously recorded their process, which involved: **Data Collection Expert Interviews Idea Building and Hypothesis Generation Google Data Studio Visualization Analysis and Interpretation** recommendations to the company. Essentially, they combined data analysis, expert knowledge, and data visualization techniques within Google Data Studio to understand the sales problem and suggest data-driven solutions. Overall, the paper concludes that combining big data analysis techniques with visualization tools like Google Data Studio is a powerful approach for solving business problems, particularly in sales.[10]

Developed a method for automatically generating annotations on scatterplots.

Their approach involves: **Identifying Common Annotation Strategies:** They studied how people annotate scatterplots to understand common practices. **Problem Formulation:** They formulated the annotation process as a Markov Decision Process, which provides a framework for decision-making in situations where outcomes are partly random and partly under the control of a decision-maker. **Annotation Model:** They created a model that makes annotation decisions based on data insights and chart observations. This model analyzes the data and decides which annotations to add, such as lines for correlations, circles for clusters, or arrows for movement.

Step-by-Step Generation: The model generates annotations step-by-step, aiming to convey a clear and understandable narrative of the data, especially for temporal changes. Essentially, they combined human insights with computational modeling to create an automated annotation system. It found that their automated annotation system could generate annotations comparable in quality to those created by human experts.

Here's a breakdown of their key results: **Human-Level Quality:** Their system, called Baseline 2, achieved

EXPLORE: ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v17i1.3818

similar ratings to human annotations in terms of quality and effectiveness. **Improved Understanding:** Participants found the automatically generated annotations helpful in understanding the data presented in the scatterplots. **Outperforming Baselines:** Their system significantly outperformed two baseline approaches: one using random annotations and another using a rule-based method. These results suggest that automating the annotation process can lead to more understandable scatterplots, potentially benefiting a wider audience in data interpretation. [11]
Focuses on analyzing and understanding several Dimension Reduction methods, but it also introduces a novel one: **Existing methods analyzed:** t-SNE, UMAP, TriMAP **New method proposed:** PaCMAP The authors use an empirical approach to decipher the workings of these methods, aiming to understand the impact of different design choices on their ability to preserve local and global structure. They then leverage these insights to design PaCMAP, which is specifically created to address the limitations of the existing methods. This paper delves into the challenges of Dimension Reduction techniques, particularly the trade-off between preserving local and global data structure. By analyzing existing methods like t-SNE, UMAP, and TriMAP, the authors identify key design principles for effective DR. They highlight the importance of carefully choosing loss functions and dynamically selecting graph components. These insights lead to the development of PaCMAP, a novel DR algorithm designed to overcome the limitations of its predecessors and achieve superior preservation of both local and global structure.[12]

B. Gaps and Future Research Direction

The review has identified several key gaps and opportunities for future research in the field of data visualization:

1) Lack of standardized evaluation frameworks: This is crucial for the scientific rigor and comparability of research. Without consistent methods to evaluate effectiveness, it's challenging to objectively assess and advance the field.
2) Underrepresentation of certain domains: This points to an opportunity for broader application and impact. Extending research to new areas can uncover unique challenges and requirements, leading to more versatile visualization solutions.
3) Exploration of emerging technologies: This highlights the need to innovate and leverage new technological capabilities (like VR/AR) to push the boundaries of how data is visualized and interacted with, potentially addressing current limitations and enhancing user experience.

4. Limitation

To ensure objectivity and transparency, it is important to acknowledge the potential limitations of this systematic review. This systematic review acknowledges several inherent limitations. Firstly, the limited database coverage, restricted to three major digital libraries, may have omitted other relevant publications, potentially impacting the comprehensiveness and generalizability of the findings. Secondly, the quality assessment of the included studies involved some level of subjectivity, despite the use of a standardized evaluation framework. Lastly, the narrow temporal scope, confined to publications within the past 5 years, inherently carries the risk of excluding important older, yet still relevant, studies.

Despite these limitations, this systematic review provides a valuable synthesis of the current state of the literature on data visualization approaches and their application to big data challenges.

In conclusion this systematic review has provided a comprehensive overview of the literature on utilizing machine learning in big data visualization. The analysis has revealed that the use of machine learning and deep learning can enhance the effectiveness of data visualization. Additionally, the findings highlight the critical role of data visualization in extracting meaningful insights from large, complex datasets, as well as the ongoing challenges and emerging opportunities in this rapidly evolving field. Finally, it is the combination of related research areas including visualization, data mining, and statistics that turns visual analytics into a promising field of research.

References

[1]. S. Mysore, M. Jasim, H. Song, S. Akbar, A. K. C. Randall and N. Mahyar, "How Data Scientists Review the Scholarly Literature".
[2]. S. M. Ali, N. Gupta, G. K. Nayak and R. K. Lenka, "Big data visualization: Tools and challenges".
[3]. G. Chawla, S. Bamal and R. Khatana, "Big Data Analytics for Data Visualization: Review of Techniques".
[4]. E. Y. Gorodov and V. Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data".
[5]. D. Tranfield, D. Denyer, and P. Smart, 'Towards a methodology for developing evidence-informed management knowledge by means of systematic review', British journal of management, vol. 14, no. 3, pp. 207–222, 2003.
[6]. M. J. Page et al., 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', International journal of surgery, vol. 88, p. 105906, 2021.
[7]. Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMAP, and PaCMAP for Data

EXPLORE: ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v17i1.3818

Visualization," *Journal of Machine Learning Research*, vol 22, pp. 1-73, 2021.

[8]. M. H. Allaymoun, M. Khaled, F. Saleh, and F. Merza "Data Visualization and Statistical Graphics in Big Data Analysis by Google Data Studio – Sales Case Study," *2022 IEEE Technology and Engineering Management Conference*, 2022, pp. X-X.

[9]. D. Shi, A. Oulasvirta, T. Weinkauf and N. Cao "Understanding and Automating Graphical Annotations on Animated Scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol 25, no. 1, pp. 821-830, Jan. 2019. doi 10.1109/TVCG.2018.2865000

[10]. M. Ali, J. Choudhary, and T. Kasbe, "A hybrid model for data visualization using linear algebra methods and machine learning algorithm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, pp. 463-475, Jan. 2024. doi: 10.11591/ijeecs.v33.i1.pp463-475

[11]. Y. Qiu and J. Lu, "A visualization algorithm for medical big data based on deep learning," *Measurement*, vol. 183, p. 109808, 2021.

[12]. G. Yan and B. Yu, "An Intelligent Visualization Method for Classic and Famous Prescriptions Based on Big Data," *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics*, Dhaka, Bangladesh, 2024, pp. 1-6.

EXPLORE: ISSN: 2087-2062, Online ISSN: 2686-181X / DOI: http://dx.doi.org/10.36448/jsit.v17i1.3818