Financial Big data Visualization: A Machine Learning Perspective

Alice Xiaodan Dong
University of Technology Sydney
Sydney, Australia
xiaodan.dong@uts.edu.au

Weidong Huang
University of Technology Sydney
Sydney, Australia
weidong.huang@uts.edu.au

Jitong Wang
University of Technology Sydney
Sydney, Australia
jitong.wang@student.uts.edu.au

ABSTRACT

In today's technology-driven environment, the exponential growth of big data underscores the importance of visualizing and analyzing it to derive actionable insights. This need spans across industrial sectors, with particular importance in the financial industry. While numerous modern models and algorithms have been developed, and utilized in diverse applications, it is crucial to classify these methodologies for users to identify the most suitable ones. In this paper, we embark on a selective review to streamline the classification of financial big data visualization methodologies from a machine learning perspective and explore the latest trends. We categorize techniques based on two key elements: the modeling stage and the nature of big data. The analytical stage divides methods into three phases: pre-model building, during-model building, and post-model building. Additionally, the characteristics of big data play an important role in shaping methodologies. We delve into three primary types of big data—structured, semi-structured, and unstructured and identify the most popular financial data types within each category in current society. We also discuss and highlight some research opportunities that we hope could be useful for visual analytics researchers.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Visualization techniques**.

KEYWORDS

Machine Learning, Visualization, Telematics, Big Data

ACM Reference Format:
Alice Xiaodan Dong, Weidong Huang, and Jitong Wang. 2024. Financial Big data Visualization: A Machine Learning Perspective. In *The 17th International Symposium on Visual Information Communication and Interaction (VINCI 2024), December 11–13, 2024, Hsinchu, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3678698.3678702

1 INTRODUCTION

Big Data is a critical asset in the competitive market of the digital economy. With advancing technology, the availability and capture of big data are increasing exponentially. Applications utilizing big data, artificial intelligence, and machine learning result in decisions and business processes that are specifically tailored to each person's wants and expectations, enhancing the expansion and effectiveness of commercial operations [27]. Particularly in the financial sector, big data wields a profound influence on the economic system. Moreover, the integration of machine learning and visualization techniques is rapidly gaining traction, shaping contemporary trends in data analysis and decision-making processes. In the area of financial data and technology, machine learning and deep leanring exert a profound and transformative influence. Its algorithms have significantly impacted various facets of finance, spanning portfolio management, risk mitigation, and customer acquisition. Visualization in this area is increasingly in demand because business stakeholders are eager to understand these algorithms and models. Visualization of financial information is particularly important for insiders and stockholders for effective communication and collaboration and to close the gap between financial experts and non-experts [9].

An essential utilization of machine learning visualization within finance lies in the prediction of market trends. Conventional market analysis models have faced difficulties in managing the intricacies and fluctuations inherent in financial markets. However, leveraging machine learning visualization facilitates the identification and explication of intricate patterns within financial data. Fraud detection within financial institutions has been revolutionized by machine learning visualization. In the past, spotting fraudulent transactions was slow and prone to errors. But now, with machine learning's ability to recognize patterns, we can quickly and accurately identify anomalies, safeguarding both businesses and consumers from financial losses [28]. The recent advancements and trends in financial big data visualization analytics necessitate a diverse set of objectives, prompting research across various application areas. To this end, our review encourages a multidisciplinary approach focused on achieving intelligibility and transparency goals. The following sub-sections present the overall review of Financial Big Data visualization and its applications from a machine learning or deep learning perspective.

2 RELATED SURVEYS AND GUIDELINES

In recent years, surveys and review papers have suggested research directions and highlighted challenges in data visual analytics and deep learning. Mohseni et al. [31] summarized evaluation methods and provided recommendations for various design goals in Explainable AI research. Yuan et al. [54] built a taxonomy to include three first-level categories: techniques before model building, techniques during modeling building, and techniques after model building in machine learning visualization. Cockcroft et al. [10] summarized the research opportunities for the use of 'big data' in accounting and finance. We have summarized the most recent surveys on financial big data analytics from a machine learning perspective. These surveys span multiple disciplines, so we have organized them into the following areas.

VINCI 2024, December 11–13, 2024, Hsinchu, Taiwan

ACM ISBN 979-8-4007-0967-8/24/12
https://doi.org/10.1145/3678698.3678702

- Machine Learning or deep learning visualization Surveys. The interpretation of machine learning models is a hot topic in the information visualization community. Research shows that insights from these models can enhance predictions and increase the trustworthiness of results. Consequently, numerous extensive survey articles have been published recently to summarize the growing body of original research papers on this subject. Chatzimparmpas et al. [7] presented a meta-analysis that includes both an overview and an in-depth examination of existing surveys on the interpretation of machine learning models. Jing et al. [22] provided an extensive review of deep learning-based self-supervised methods for general visual feature learning from images and videos.
- Big Data Analytics Surveys. With the advent of Big Data, researchers have developed new value chain models known as Data Value Chains to facilitate data-driven organizations. These models have evolved into Big Data Value Chains to address challenges such as high volume, velocity, and variety. Big Data Value Chains outline the flow of data within organizations that leverage Big Data to extract valuable insights [13]. Jahani et al. [21] reviewed the methodology for the field of Data Science and Big Data Analytics (DS & BDA) in Supply Chain and Logistics (SC & L) to classify existing DS & BDA models and techniques, structure their practical application areas, and identify research gaps and potential future research directions.
- Financial Analytics Surveys. Trends in financial surveys indicate a growing use of analytic technology to enhance productivity and enable more agile decision-making across enterprises [1]. A recent survey [1] suggested that along with investing in new technology, almost all organizations (91%) are investing or planning to invest in big data solutions that specifically support finance functions. In addition to investing in big data solutions, there are also reviews focused on specific financial industries, such as banking. Nobanee et al. [17] provided a review of existing literature on big data applications in banking, utilizing a bibliometric analysis approach.

While all the surveys and guidelines in this section contribute valuable insights to big data visualization research, to the best of our knowledge, there is no existing comprehensive review or framework specifically addressing financial big data visualization from a machine learning or deep learning perspective. This paper aims to fill in this gap.

3 REVIEW METHOD

To ensure thorough coverage of the literature, we adopted a systematic methodology for our search. Firstly, we identified keywords and phrases that encompass the field of financial big data visualization from a machine learning perspective, e.g., “big data”, “visualization”, “visual analytics”, “machine learning”, and “financial”. Subsequently, we utilized a logical AND combination of these keywords to query digital libraries and conference proceedings. Our initial search was conducted on abstracts from the IEEE Xplore and ACM Digital Libraries over the past ten years, from 2014 to 2024. We found 52 papers in IEEE and 46 papers in ACM.

To cover broader articles, We searched the same keywords using web of science. The final list was limited to papers from 2014 to 2024. The result gives 109 articles. After excluding the irrelevant articles and articles searched by other libraries above, we then used 32 to identify the research attributes such as data size and visualisation components in those papers. Finally, we selected 25 representative papers for the reference table. Some papers contain all the relevant keywords. For instance, Usman et al. [50] proposed a Poisson mixture model to predict future insurance claim frequencies using telematics data. Although this work is contextually relevant to our review, its emphasis is on prediction rather than visualization. Therefore, it will be excluded from this review.

4 CATEGORIZATION METHOD

After thoroughly reviewing the literature, we found that the design goals and types of data are the primary factors influencing the design and purpose of visualization methodologies. Therefore, we propose categorizing the papers in the field of big data visual analytics based on their design goals and the types of big data they address. In this paper, we examine the most popular types of data within each classification. We aim to use this approach to better understand big data visualization methodologies, and to identify the most effective strategies for different scenarios.

- **Design Goals** The first element in our categorization is the design goal for financial big data visualization. Visual analytics techniques for machine learning are classified into three groups by the corresponding analysis stage: techniques before, during, and after model building [54]. In financial institutions, key areas using machine learning include risk management, compliance, and customer trust. In risk management, algorithms are essential for risk assessment and decision-making. Transparent methodologies ensure these algorithms are fair, accurate, and regulatory-compliant. Pre-model building analysis sets clear guidelines, aligning algorithms with risk management strategies. In compliance and regulation, regulatory bodies require transparency in algorithms for credit scoring, loan approvals, and investments. After thoroughly reviewing and analyzing the research design goals in the financial big data visualization area, we used the three key categories organized by the process of model building: pre-model building, during model building, and post-model building. These categories capture and organize the breadth of design goals in the field of financial big data.
  - Pre-model building
    - Data preprocessing visualization
    - Algorithmic Transparency
  - During model building
    - Model Visualization and Inspection
  - Post-model building
    - Model Diagnosis and Interpretation
- **Types of Big Data** The type of big data influences the methodology review by determining the tools, techniques, and frameworks needed to efficiently process, analyze, and derive insights from the data. We review the current types of data commonly associated with big data in the financial

sector. Big data has many characteristics and consists of structured, unstructured, and semi-structured data formats [38]. Structured data refers to information that is highly organized and easily searchable within databases due to its pre-defined format. Unstructured data, on the other hand, lacks a predefined format or organizational structure. This type of data is often textual content, such as emails and social media posts. By employing this classification, we identify the most popular data types within each category. Despite the variety of data types, the emerging most popular big data categories most relevant to financial institutions today can be classified into the following five types:

- Structured data - Stock market data. Stock market data includes key metrics such as stock prices and interest rates. This data is considered a principal component of financial big data due to its extensive use in financial analysis, predictive modeling, and strategic decision-making [18]. The large volume and high velocity of stock market data make it an essential resource for investors, analysts, and financial institutions seeking to derive actionable insights and optimize financial outcomes.
- Structured data - Transactional data. With the development of data-storage technology, banks have accumulated a large amount of payment and transactional data. These data are collected constantly and are more reliable and flexible than financial statements[24]. Saxena et al. [40] established a network comprising 1.6 million nodes, derived from banking transactions of Rabobank users, aimed at enhancing the early detection of suspicious and anomalous user activities.
- Semi-structured data - Internet mouse-tracking data. Banks are starting to track user behavior on the internet and mobile apps, resulting in the generation of extensive mouse-clicking data. This data is commonly used to identify fraudulent activities and analyzing web traffic. Pageviews are the most widely used metric for analyzing web traffic across various sectors, including blogs, social media, and technology. The geometric data generated by mouse tracking is vast and qualifies as big data [34].
- Semi-structured data - Telematic data. Telematics data has emerged as a source of big data as it has volume, variety, velocity and veracity [14]. Car telematics is a large and growing business sector aiming to collect mobility-related data to develop a wide range of services for both individual citizens and companies. The data collected through car telematics typically includes vehicle movement traces, gathered via an ad hoc device installed in the vehicle. Telematics time series data captures a multitude of variables every second while driving. Unlike traditional covariate information, telematics car driving data provides direct insights into driving habits and styles, making it useful for claims prediction in financial institutions[15]. It is increasingly utilized in finance for various purposes, including risk assessment where insurance companies use it to accurately assess risk profiles. It also plays a crucial role in fraud detection, for instance, analyzing transaction locations and timing to identify unusual or suspicious activities.
- Unstructured data - text data. Unstructured textual data have been increasing rapidly in the finance industry [6]. It is generally used for auditing, customer satisfaction and marketing purposes. Text mining utilizes various techniques to process and interpret unstructured text data. These methods transform large volumes of text to uncover key facts, relationships, and patterns. Advances in both statistical learning and symbolic AI have greatly enhanced text-mining methods [45].

To help summarize our characterization and provide illustrative examples from the literature, Table 1 presents an example of cross-referencing the literature, focusing on the aspects of design goals and types of big data. In Section 5, we review research focusing on design goals, detailing six goals organized by user groups. Following that, Section 6 reviews the types of data and methods used.

5 DESIGN GOALS

Design goals are crucial for categorizing research methods in data visualization because they define the objectives and criteria for evaluating various visualization techniques. These goals guide the selection of appropriate methods throughout different phases of model building, from pre-model building to post-model building.

5.1 Pre-model building

- **DG1:Data Preprocessing Visualization** Before the model building phase, big data preprocessing visualization plays a crucial role in preparing the data for effective and accurate modeling. Equal-Height Treemaps[52] was introduced to add additional visual information in a multi-variate treemap. This method aids in visual mapping of more than two data variables and visualizes the hierarchy and relationships between multiple variables. Dimensionality reduction is a common goal when working with big data. Manifold learning algorithm and Laplacian Eigenmaps (LE) were employed to extract the intrinsic manifold structure embedding in the financial system [19]. Its goal is to decrease the dataset's dimensionality while maintaining its inherent structure intact. This facilitates the visualization and comprehension of relationships among various financial instruments or variables. Roemsri et al. [39] proposed a web-based application leveraging data virtualization techniques to analyze and visualize distributed denial-of-service attack data in the Internet of Things network traffic using 3D plots. This framework provides a comprehensive visualization of the nature and characteristics of distributed denial-of-service (DDoS) attacks in Internet of Things (IoT) networks.
- **DG2:Algorithmic Transparency** Visualization plays a crucial role in enhancing algorithmic transparency in machine learning by providing clear and interpretable insights into how models make decisions. Shahoud et al. [41] introduced a new generic microservice-based framework for realizing the concept of meta-learning in Big Data environments. Meta-learning supports non-expert users by recommending

Table 1: Summary of Categorization

Papers	Design Goals				Types of Big Data				
Category	DG1: Data preprocessing visualization	DG2: Algorithmic Transparency	DG3 and 4: Model Visualization and Inspection	DG5 and 6: Model Diagnosis and Interpretation	Structured data - Stock market data	Structured data - Transactional data	Semi-structured data - Internet mouse-tracking data	Semi-structured data - Telematic data	Unstructured data - Text data
Pang et al. [33]	*	*	*	*	*				
Shi et al. [43]		*	*	*	*				*
Yu et al. [53]	*	*	*	*		*			
Madhukar et al. [36]	*	*	*	*		*			
Liu et al. [26]	*		*	*		*			
Leporowski et al.[25]	*		*	*	*				
Cheng et al. [32]	*	*	*	*	*				
Demestichas et al. [11]	*		*	*		*			
Chen et al.[8]	*	*	*	*	*				
Huang et al.[20]	*	*	*	*		*			
Tuarob et al.[49]	*	*		*	*				
Birogul et al.[4]		*	*	*	*				
Shen et al. [42]		*	*	*	*				
Purnama and Usagawa[35]	*						*		
Gao et al.[16]	*	*						*	
Raveh et al.[37]				*					*
Takama et al.[47]			*	*					*
Huang[19]		*	*		*				
Leite et al.[3]			*			*			
Macas et al. [29]	*	*				*			
Song et al.[44]	*		*	*		*			
Meyer et al.[30]	*			*			*		
Roel et al. [51]	*		*	*				*	
Boylan et al.[5]		*						*	
Suh et al. [46]		*	*	*					*

learning algorithms based on meta-features computed from a given dataset. The proposed generic framework makes use of a Big Data software stack, container visualization, and a microservice architecture for a fully manageable and highly scalable solution to provide algorithmic transparency. This framework is useful for various applications in finance, including economic forecasting and future loss prediction. Furthermore, 3D bubble plots are proposed to help visualize how neural network models work [12]. This method visualizes the frequency and size of each neuron in a network model, enabling non-technical business users to understand it.

5.2 During model building

- **DG 3 and 4: Model Visualization and Inspection** During the model-building phase, model visualization and inspection tools help understanding how the model behaves concerning the input data. This includes identifying which features are most influential and how they interact. 3D bubble plots were proposed to help inspect whether a neural network and Bayesian optimization algorithm work properly by analyzing the direction of each neuron and the steps in Bayesian [12]. With the pervasive use of big data techniques, companies increasingly rely on algorithms to select individuals who meet certain criteria. However, this process often involves bias. A framework named FairSight was proposed to implement a visual analytics system to support the analysis of fairness in ranking problems [2]. The machine learning pipeline in FairSight is divided into three phases (data, model, and outcome), and bias is measured at both individual and group levels using various metrics for model diagnosis purposes.

5.3 Post-model building

- **DG 5 and 6:Model Diagnosis and Interpretation** Post-model building, visualization plays a key role in diagnosing and interpreting the model's performance and behavior. Suh et al. [46] introduced AnchorViz, an interactive visualization tool designed for exploring semantic data and uncovering concepts in machine learning. This framework helps users identify more prediction errors for model diagnosis purposes compared to stratified random and uncertainty sampling methods after model building. An interactive visualisation tool called EMILE-UI [23] was applied for users to evaluate the provided explanations of an image-based classification task, specifically those provided by saliency maps. This tool visualizes the relationship between the machine learning model and its explanation of input images, making it easier to interpret saliency maps and understand how the model makes predictions. Tian et al. [48] proposed the use of machine learning algorithms to analyze and visualize risk indicators in enterprises, aiming to predict their likelihood of fraud. Specifically, K-Nearest Neighbors (KNN) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms were utilized to visualize distinct predictions for various risk types.

6 TYPES OF BIG DATA

The types of big data are important to the review and categorization of research methods for data visualization because different types of data have distinct characteristics and requirements.Currently, financial institutions recognize four major types of big data, namely Stock market data, Transactional data, Internet mouse-tracking data, and Telematic data. By understanding these categories, researchers can effectively choose and classify visualization methods tailored to address the distinct challenges and capitalize on the opportunities associated with each data type.

6.1 Structured data - Stock Market data

Stock market data is a prime example of structured big data in finance. Major stock exchanges like NYSE, NASDAQ, and others provide extensive structured data on stock prices, trading volumes, and other financial metrics. Visualization helps with transforming complex data into accessible, interpretable, and actionable insights. Pang et al.[33] compared the accuracy of two neural network model predicting stock market, including long short-term memory (LSTM) with automatic encoder and embedded layer. 3D charts are used to illustrate the pattern of stock vector as input data of deep learning models. A visual system, with an interface combining stock price charts, impacted texts and dense chart interpreting the relationships with keywords, can provide more transparency to users to understand stock prediction from neural network models that using texts such as news and social media information[43]. Leporowski et al.[25] proposed an innovation visualisation method to present time-series stock data in one two-dimension chart to be easily used to compare various deep learning models.

Combining multiple sources of information is another area that receives great attention in visualization. Cheng et al.[32] combined both social media information and stock market information into neural network model to predict stock market and use bar and line charts to present the trend and insights to users in an interface. Chen et al.[8] propose a method to convert time-series data into 2D images which used to train convolutional neural networks (CNNs) for stock market prediction, which is using visual to support deep learning modelling. The paper [49] proposed DAViS, an interactive web application for capturing dynamic stock market trends with a unified solution for real-time data collection, analysis, and visualization using line charts and word clouds for timely decision-making. Birogul et al.[4] proposed a methodology YOLO to predict stock market purely by utilizing the timeseries 2D Candlestick charts to assist to make decisions of buy or sell.

6.2 Structured data - Transactional data

Transactions, as geo-referenced movements of money, can be visualized as origin–destination (OD) flow maps [3]. The analysis of financial transactions can be an overwhelming task for bank and fraud analysts [29]. By combining computational strategies with visual cognitive intelligence, visual analytics can facilitate the analysis of transactional data and enhance the representation of transactions over time [29]. Madhukar et al.[36] present a deep learning algorithm that combines dimensionality reduction with an RNN classifier to detect credit card fraud. Scatter plot is used to show fraud data by various pre-process methodologies. Macas et

al.[29] propose a visualization tool that aims to ease the analysis of banking transactions over time and the detection of the transactions' topology and of suspicious behavior. Several innovated graphs are designed specifically for interpreting banking transactions data, such as combination of bar chart and scatter plot, icon circled matrix projection and force directed graph.

Furthermore, a temporal graph neural network approach [44] was proposed to investigate individual fairness on dynamic graphs for the loan approval task on Transaction Networks. Yu et al.[53] present a long short-term memory model to identify credit card fraud and compared various model results. A dense graph by blue and red colors was used to represent the transaction input data by fraud or not. Deep learning model with a comprehensive flow chart is presented to predict and reduce the risk of overdue payments from repayment transactions[26]. Demestichas et al.[11] combined financial transactional data and other available data sources to build a system to detect abnormal and fraud activities from crime mornitoring perspective. The tool introduce a visualisation and interaction module, including a knowledge graph to assist users to understand the correlations. Huang et al.[20] propose a K-means clustering method to improve financial fraud detection from transaction data where 3D graph is used to interpret the clustering detection result.

6.3 Semi-structured data - Internet mouse-tracking data

Internet mouse-tracking data refers to the collection and analysis of users' mouse movements and clicks as they navigate a website or online application. The geometrical data generated by mouse tracking are extremely large, and qualify as big data [35]. As technology advances, more people are performing activities online, ranging from shopping to banking. Internet mouse-tracking data presents great potential for businesses by providing valuable insights into user behavior and preferences. By analyzing this data, businesses can enhance user experience, optimize website design, and tailor their services to better meet customer needs. The techniques of mouse tracking are divided into three types, namely default mouse tracking, whole page tracking, and Region of Interest tracking [35]. Default mouse tracking data can precisely visualize exact points of user interaction. The visualization of mouse tracking data can be achieved through a heatmap, which is based on the duration the mouse cursor stays at each point [35]. Mouse tracking is an important source of data in cognitive science [30]. Singular Value Decomposition (SVD) and Detrended Fluctuation Analysis (DFA) were utilized to study unconstrained mouse tracking data, aiming to predict and visualize user behavior[30]. Through the visualization and experimentation with a simple online game, they discovered that the time series of mouse movements reveals systematic differences between accurate and non-accurate players. This finding confirms the presence of high-level information in mouse trace data.

6.4 Semi-structured data - Telematic Data

When discussing big data in fleet management for financial institutes, telematics is the first topic that comes to mind. Telematics technology involves collecting, storing, and transmitting information between end users and vehicles through telecommunication devices. These potentially high dimensional telematics data, force pricing actuaries to change their current practice [51]. The application of big data in telematics enhances the utility of this data, providing deeper insights and more valuable use cases. Machine learning is the dominant method for analysis and visualization of in-vehicle telematics data [5]. Gao et al. [16] proposed two data-driven neural network approaches that process telematics car driving data to complement classical actuarial pricing with a driving behavior risk factor from telematics data. The individual telematics car driving data was compressed into a speed-acceleration (v-a) heatmap to visualize each driver's riskiness [16]. These v-a heatmaps illustrate how drivers accelerate and brake at different speeds. Specifically, a densely connected feed-forward neural network and a convolutional neural network were built to extract driver risk information from telematics car driving data, represented by v-a heatmaps.

6.5 Unstructured data - Text data

People worldwide are increasingly facing vast amounts of unstructured Big Data, predominantly in textual format. In many cases, people, such as evaluators and auditors, don't know what they are looking for in financial and text reports. An automated big data science technique based on text mining, machine learning, and data visualization was developed to help researchers and evaluation practitioners reveal trends, trajectories, and interrelations within textual information to support their evaluations [37]. The proposed system automatically extracts extensive descriptive terminology for a specific domain, identifies semantic connections between documents, visualizes the entire document repository as a graph of semantic connections, and directs the user to areas where the most interesting trends can be observed [37]. It has been shown that texts such as financial news and tweets on stock markets are useful in predicting stock price movements [43].

Furthermore, DeepClue [43], a visual interpretation system. was built to bridge text-based deep learning models and end users by visually interpreting the key factors learned in the stock price prediction model. This system aims to interpret stock price predictions and enhance investment decisions in the stock market. Takama et al. [47] proposed a Treemap-based visualization for supporting cluster analysis of multi-dimensional data. The visualization method applies Fuzzy c-Means to target data and visualizes the result based on the highest and the second-highest membership values with Treemap. A prototype interface is implemented and its effectiveness is investigated through a user experiment on a news articles dataset.

7 RESEARCH OPPORTUNITIES

Although the hybrid fields of financial big data, visual analytics, and machine learning have achieved promising results in both academia and real-world applications, several long-term research challenges remain. In this paper, we categories these methods and explore potential research opportunities in this area. While great progress has been made in leveraging machine learning visualization for structured and unstructured data, There has been relatively little focus on leveraging machine learning to visualize semi-structured data across all three modeling stages in the financial domain. This highlights a potential area for future research.

While many existing studies in this field primarily focus on prediction methodologies [50] and utilize visualization to aid machine learning for prediction, there is comparatively less emphasis on employing machine learning to facilitate visualization for semi-structured financial data. The two most recent popular semi-structured data types in finanical institutes, internet mouse tracking and telematics data, contain a wealth of information. For example, telematics data encompasses speed, acceleration, and spatial information such as geographic coordinates. Utilizing machine learning visualization can render this complex data more intuitive and accessible, aiding both understanding and risk management efforts in financial institutions. Integrating visualization of telematics data from a machine-learning perspective offers great advantages for banks and insurance companies. It enables better classification of customer risk, enhances driving safety, and contributes to the development of self-driving vehicles. In summary, exploring machine learning techniques for visualizing semi-structured financial data represents a great area of research opportunity.

Furthermore, there is comparatively less attention given to visualization techniques that explore the fusion of structured and unstructured components within semi-structured data in the financial domain. Integrating and combining various data types within semi-structured big data could enhance the performance and outcomes of machine learning visualization models or systems. For example, by combining unstructured elements of telematics data, such as video and images, with structured elements like vehicle metrics and driver information, organizations can gain a comprehensive understanding of the risks involved. This integrated approach enables more accurate risk prediction and allows for proactive measures to mitigate potential hazards.

8 CONCLUSION

This paper selectively reviews recent progress and developments in financial big data visual analytics from a machine learning perspective. The techniques are classified based on two elements: the analytical stage and the types of big data. The analytical stage categorizes methods into three groups: pre-model building, during-model building, and post-model building. The types of big data also influence the methodology. We review three main types of big data, namely structured, semi-structured, and unstructured, and identify the popular data types in each category in today's society. Each category is detailed with a set of representative works. By selectively analyzing existing financial big data visual analytics research from a machine learning perspective, we suggest two areas for future research. These include enhancing machine learning visualization for semi-structured data and introducing fusions between the structured and unstructured elements within semi-structured data to improve overall model performance and outcomes. We hope this survey will be useful for researchers in this area.

REFERENCES

[1] 2023. New Survey Finds Financial Leaders Investing in Analytics, AI and Machine Learning Tools as They Navigate Economic Uncertainty in 2023. http://ezproxy.lib.uts.edu.au/login?url=https://www.proquest.com/wire-feeds/new-survey-finds-financial-leaders-investing/docview/2811185598/se-2 Name - OneStream Software; Copyright - Copyright Business Wire 2023; Last updated - 2023-11-28.

[2] Y. Ahn and H Shibata. 2019. airSight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 1086–1095.

[3] Leite RA Dustdar S Miksch S Sorger J. Arleo A, Tsigkanos C. 2023. Visual Exploration of Financial Data with Incremental Domain Knowledge. *Comput Graph Forum* 42, 01 (2023), 101–116. https://doi.org/10.1111/cgf.14723

[4] Serdar Birogul, Günay Temür, and Utku Köse. 2020. YOLO Object Recognition Algorithm and "Buy-Sell Decision" Model over 2D Candlestick Charts. *IEEE Access* PP (05 2020), 1–1. https://doi.org/10.1109/ACCESS.2020.2994282

[5] James Boylan, Denny Meyer, and Won Sun Chen. 2024. A systematic review of the use of in-vehicle telematics in monitoring driving behaviours. *Accident Analysis I& Prevention* 199 (2024), 107519. https://doi.org/10.1016/j.aap.2024.107519

[6] Lewis C and Young S. 2019. Fad or future? Automated analysis of financial text and its implications for corporate reporting. In *2019 Accounting and Business Research*, Vol. 49. 616–618. https://doi.org/10.1080/00014788.2019.1611731

[7] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. 2020. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization* 19, 3 (2020), 207–233. https://doi.org/10.1177/1473871620904671 arXiv:https://doi.org/10.1177/1473871620904671

[8] Jou-Fan Chen, Wei-Lun Chen, Chun-Ping Huang, Szu-Hao Huang, and An-Pin Chen. 2016. Financial Time-Series Data Analysis Using Deep Convolutional Neural Networks. In *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. 87–92. https://doi.org/10.1109/CCBD.2016.027

[9] M. Chy and O. Buadi. 2023. Role of Data Visualization in Finance. *American Journal of Industrial and Business Management* 13 (2023). https://doi.org/0.4236/ajibm.2023.138047

[10] I& Russell M. Cockcroft, S. 2018. Big data opportunities for accounting and finance practice and research. *Australian Accounting Review* 28 (2018), 323. https://doi.org/10.1111/auar.12218

[11] Konstantinos Demestichas, Nikolaos Peppes, Theodoros Alexakis, and Evgenia Adamopoulou. 2021. An Advanced Abnormal Behavior Detection Engine Embedding Autoencoders for the Investigation of Financial Transactions. *Information* 12, 1 (2021). https://doi.org/10.3390/info12010034

[12] Xiaodan Dong, Weidong Huang, and Jitong Wang. 2024. Business-Centric Modelling and Visualization for Retail Promotion. I-DO 2024 conference., Taipei, Taiwang. https://www.ido2024-conferences.ntunhs.edu.tw/program

[13] Abou Z. Faroukhi, Alaoui I. El, Gahi Youssef, and Amine Aouatif. 2020. Big data monetization throughout Big Data Value Chain: a comprehensive review. *Journal of Big Data* 7, 1 (01 2020). http://ezproxy.lib.uts.edu.au/login?url=https://www.proquest.com/scholarly-journals/big-data-monetization-throughout-value-chain/docview/2343302848/se-2

[14] Beate Franke, Jean-François Plante, Ribana Roscher, En-Shiun Annie Lee, Cathal Smyth, Armin Hatefi, Fuqi Chen, Einat Gil, Alexander Schwing, Alessandro Selvitella, Michael M. Hoffman, Roger Grosse, Dieter Hendricks, and Nancy Reid. 2016. Statistical Inference, Learning and Models in Big Data. *International Statistical Review / Revue Internationale de Statistique* 84, 3 (2016), 371–389. http://www.jstor.org/stable/44162504

[15] Guangyuan Gao, Shengwang Meng, and Mario V. Wüthrich. 2022. What can we learn from telematics car driving data: A survey. *Insurance: Mathematics and Economics* 104 (2022), 185–199. https://doi.org/10.1016/j.insmatheco.2022.02.004

[16] Wang H. Gao, G. and M.V. Wüthrich. 2022. Boosting Poisson regression models with telematics car driving data. *The Journal of Machine Learning* 111 (2022), 243–272. https://doi.org/10.1007/s10994-021-05957-0

[17] Mona Al Dhanhani Maitha Al Neyadi1 Sultan Al Qubaisi Haitham Nobanee, Mehroz Nida Dilshad and Saeed Al Shamsi. 2021. Big Data Applications the Banking Sector: A Bibliometric Analysis Approach. *Sage Open* 11 (2021). https://doi.org/10.1177/21582440211067234

[18] Popp J. Hasan, M.M. and J. Oláh. 2020. Current landscape and influence of big data on finance. *Journal of Big Data* 21 (2020). https://doi.org/10.1186/s40537-020-00291-z

[19] Yan Huang. 2020. Manifold Learning for Financial Market Visualization. In *Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence* (Chengdu, China) *(ICMAI '20)*. Association for Computing Machinery, New York, NY, USA, 239–243. https://doi.org/10.1145/3395260.3395297

[20] Zengyi Huang, Haotian Zheng, Chen Li, and Chang Che. 2024. Application of Machine Learning-Based K-Means Clustering for Financial Fraud Detection. *Academic Journal of Science and Technology* 10, 1 (2024), 33–39.

[21] Jain R. Jahani, H. and D. Ivanov. 2023. Data science and big data analytics: a systematic review of methodologies used in the supply chain and logistics research. *Annals of Operations Research* 7 (2023). https://doi.org/10.1007/s10479-023-05390-7

[22] Longlong Jing and Yingli Tian. 2021. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2021), 4037–4058. https://doi.org/10.1109/TPAMI.2020.2992393

[23] Md Abdul Kadir, Abdulrahman Mohamed Selim, Michael Barz, and Daniel Sonntag. 2023. A User Interface for Explaining Machine Learning Model Explanations. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (<conf-loc>, <city>Sydney</city>, <state>NSW</state>, <country>Australia</country>, </conf-loc>) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 59–63. https://doi.org/10.1145/3581754.3584131

[24] Gang Kou, Yong Xu, Yi Peng, Feng Shen, Yang Chen, Kun Chang, and Shaomin Kou. 2021. Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. *Decision Support Systems* 140 (2021), 113429. https://doi.org/10.1016/j.dss.2020.113429

[25] Blazej Leporowski and Alexandros Iosifidis. 2021. Visualising Deep Network's Time-Series Representations. *CoRR* abs/2103.07176 (2021). arXiv:2103.07176 https://arxiv.org/abs/2103.07176

[26] Bin Liu, Zhexi Zhang, Junchi Yan, Ning Zhang, Hongyuan Zha, Guofu Li, Yanting Li, and Quan Yu. 2020. A Deep Learning Approach with Feature Derivation and Selection for Overdue Repayment Forecasting. *Applied Sciences* 10, 23 (2020). https://doi.org/10.3390/app10238491

[27] SUNITA MALL, TUSHAR RANJAN PANIGRAHI, and SUSHMA VERMA. 2023. BIBLIOMETRIC ANALYSIS ON BIG DATA APPLICATIONS IN INSURANCE SECTOR: PAST, PRESENT, AND FUTURE RESEARCH DIRECTIONS. *Journal of Financial Management, Markets and Institutions* 11, 01 (2023), 2330001. https://doi.org/10.1142/S2282717X23300015 arXiv:https://doi.org/10.1142/S2282717X23300015

[28] Abdel Latif Marazqah Btoush, Eyad, Xujuan Zhou, Raj Gururajan, Ka C. Chan, Rohan Genrich, and Prema Sankaran. 2023. A systematic review of literature on credit card cyber fraud detection using machine and deep learning. *PeerJ.Computer science* 9 (2023), 1. http://ezproxy.lib.uts.edu.au/login?url=https://www.proquest.com/scholarly-journals/systematic-review-literature-on-credit-card-cyber/docview/2828773923/se-2 Date created - 2023-06-22; Date revised - 2024-02-02; SuppNotes - Conflict of Interest: The authors declare that they have no competing interests. Cited By: Syst Rev. 2016 Dec 5;5(1):210 27919275] Comput Intell Neurosci. 2020 Feb 8;2020:6503459 32089669] Ann Oper Res. 2021 Jun 8;:1-23 34121790; Last updated - 2024-02-08.

[29] Catarina Maçãs, Evgheni Polisciuc, and Penousal Machado. 2020. VaBank: Visual Analytics for Banking Transactions. In *2020 24th International Conference Information Visualisation (IV)*. 336–343. https://doi.org/10.1109/IV51561.2020.00062

[30] Kim A. D. Spivey M. I& Yoshimi J. Meyer, T. 2023. A new approach to analyzing continuous mouse tracking data. *Behavior Research Methods* (2023). https://doi.org/10.3758/s13428-023-02210-5

[31] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4, Article 24 (sep 2021), 45 pages. https://doi.org/10.1145/3387166

[32] Huihui Ni, Shuting Wang, and Peng Cheng. 2021. A hybrid approach for stock trend prediction based on tweets embedding and historical prices. *World Wide Web* 24 (05 2021), 1–20. https://doi.org/10.1007/s11280-021-00880-9

[33] Xiongwen Pang, Yanqiang Zhou, Pan Wang, Weiwei Lin, and Victor Chang. 2020. An innovative neural network approach for stock market prediction. *The Journal of Supercomputing* 76 (2020), 2098–2118.

[34] F. Purnama and T. Usagawa. 2020. Using real-time online preprocessed mouse tracking for lower storage and transmission costs. *Journal of Big Data* 7 (2020). https://doi.org/10.1186/s40537-020-00304-x

[35] F. Purnama and T. Usagawa. 2020. Using real-time online preprocessed mouse tracking for lower storage and transmission costs. *Journal of Big Data* 27, 7 (2020). https://doi.org/10.1186/s40537-020-00304-x

[36] G. Madhukar Rao and K. Srinivas. 2022. RNN-BD: an approach for fraud visualisation and detection using deep learning. *International Journal of Computational Science and Engineering* 25, 2 (2022), 166–173. https://doi.org/10.1504/IJCSE.2022.122212 arXiv:https://www.inderscienceonline.com/doi/pdf/10.1504/IJCSE.2022.122212

[37] Ofek Y. Bekkerman R. I& Cohen H. Raveh, E. 2020. Applying Big Data visualization to detect trends in 30 years of performance reports. *Evaluation* 26, 04 (2020), 516–540. https://doi-org.ezproxy.lib.uts.edu.au/10.1177/1356389020905322

[38] R Rawat and R Yadav. 2021. Big Data: Big Data Analysis, Issues and Challenges and Technologies. *IOP Conference Series: Materials Science and Engineering* 1022, 1 (jan 2021), 012014. https://doi.org/10.1088/1757-899X/1022/1/012014

[39] Phornsawan Roemsri and Tommy Dang. 2024. Visualization of Attack Behavior Data in IoT Network Traffic. In *Proceedings of the 2023 5th International Conference on Big-Data Service and Intelligent Computation* (<conf-loc>, <city>Singapore</city>, <country>Singapore</country>, </conf-loc>) *(BDSIC '23)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3633624.3633625

[40] Akrati Saxena, Yulong Pei, Jan Veldsink, van I. Werner, George Fletcher, and Mykola Pechenizkiy. 2021. The Banking Transactions Dataset and its Comparative Analysis with Scale-free Networks. http://ezproxy.lib.uts.edu.au/login?url=https://www.proquest.com/working-papers/banking-transactions-dataset-comparative-analysis/docview/2575659859/se-2 Copyright - © 2021. This work is published under http://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2022-08-17.

[41] Shadi Shahoud, Hatem Khalloof, Moritz Winter, Clemens Duepmeier, and Veit Hagenmeyer. 2020. A Meta Learning Approach for Automating Model Selection in Big Data Environments using Microservice and Container Virtualization Technologies. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems* (Virtual Event, United Arab Emirates) *(MEDES '20)*. Association for Computing Machinery, New York, NY, USA, 84–91. https://doi.org/10.1145/3415958.3433072

[42] Jingyi Shen and M. Shafiq. 2020. Short-term stock market price trend prediction using a comprehensive deep learning system. *Journal of Big Data* 7 (08 2020). https://doi.org/10.1186/s40537-020-00333-6

[43] Lei Shi, Zhiyang Teng, Le Wang, Yue Zhang, and Alexander Binder. 2019. DeepClue: Visual Interpretation of Text-Based Deep Stock Prediction. *IEEE Transactions on Knowledge and Data Engineering* 31, 6 (2019), 1094–1108. https://doi.org/10.1109/TKDE.2018.2854193

[44] Zixing Song, Yuji Zhang, and Irwin King. 2023. Towards Fair Financial Services for All: A Temporal GNN Approach for Individual Fairness on Transaction Networks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (<conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 2331–2341. https://doi.org/10.1145/3583780.3615091

[45] Axel J. Soto, Ryan Kiros, Vlado Keselj, and Evangelos Milios. 2016. Machine learning meets visualization for extracting insights from text data. *AI Matters* 2, 2 (jan 2016), 15–17. https://doi.org/10.1145/2847557.2847560

[46] Jina Suh, Soroush Ghorashi, Gonzalo Ramos, Nan-Chen Chen, Steven Drucker, Johan Verwey, and Patrice Simard. 2019. AnchorViz: Facilitating Semantic Data Exploration and Concept Discovery for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 10, 1, Article 7 (aug 2019), 38 pages. https://doi.org/10.1145/3241379

[47] Tanaka Yuna Takama Yasufumi and Mori Yoshiyuki. 2021. Treemap-Based Cluster Visualization and its Application to Text Data Analysis. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 25, 4 (2021), 498–507. https://doi.org/10.20965/jaciii.2021.p0498

[48] Maohong Tian, Jian Liang, Dequan Zhang, Xintong Zhang, Zuo Wang, and Hualin Li. 2024. Detection of Financial Fraudulent Activities with Machine Learning:A Case Study of Detecting Potential Tax and Invoice Fraud. In *Proceedings of the 2023 7th International Conference on Computer Science and Artificial Intelligence* (<conf-loc>, <city>Beijing</city>, <country>China</country>, </conf-loc>) *(CSAI '23)*. Association for Computing Machinery, New York, NY, USA, 33–39. https://doi.org/10.1145/3638584.3638669

[49] Suppawong Tuarob, Poom Wettayakorn, Ponpat Phetchai, Siripong Traivijitkhun, Sunghoon Lim, Thanapon Noraset, and Tipajin Thaipisutikul. 2021. DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction. *Financial Innovation* 7 (2021), 1–32. https://api.semanticscholar.org/CorpusID:235763404

[50] J Udi M Wang Y Usman. F, Chan and Dong. A. [n. d.]. Claim Prediction and Premium Pricing for Telematics Auto-Insurance Data Using Poisson Regression with Lasso Regularisations. *Journal of Risk and Insurance* ([n. d.]). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4501573

[51] Roel Verbelen, Katrien Antonio, and Gerda Claeskens. 2018. Unravelling the Predictive Power of Telematics Data in Car Insurance Pricing. *Journal of the Royal Statistical Society Series C: Applied Statistics* 67, 5 (04 2018), 1275–1304. https://doi.org/10.1111/rssc.12283 arXiv:https://academic.oup.com/jrsssc/article-pdf/67/5/1275/49336700/rssc12283-sup-0001-appendixa-d.pdf

[52] Kent Wittenburg and Teng-Yok Lee. 2018. Equal-height treemaps for multivariate data. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces* (Castiglione della Pescaia, Grosseto, Italy) *(AVI '18)*. Association for Computing Machinery, New York, NY, USA, Article 88, 3 pages. https://doi.org/10.1145/3206505.3206591

[53] Yu Xie, Guanjun Liu, Chungang Yan, Changjun Jiang, Mengchu Zhou, and Maozhen Li. 2024. Learning Transactional Behavioral Representations for Credit Card Fraud Detection. *IEEE Transactions on Neural Networks and Learning Systems* 35, 4 (2024), 5735–5748. https://doi.org/10.1109/TNNLS.2022.3208967

[54] Jun Yuan, Changjian Chen, Yang Weikai, Liu Mengchen, Xia Jiazhi, and Shixia Liu. 2021. A survey of visual analytics techniques for machine learning. *Computational Visual Media* 7, 1 (03 2021), 3–36. http://ezproxy.lib.uts.edu.au/login?url=https://www.proquest.com/scholarly-journals/survey-visual-analytics-techniques-machine/docview/2500913694/se-2 Copyright - © The Author(s) 2020. This work is published under http://creativecommons.org/licenses/by/4.0/ (the "License"). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2023-11-22.