Machine Learning Based Mobile Big Data Analytics: State-of-the-art Applications, Taxonomy, Challenges and Future Research Directions

C. I. Eke[1*], A. S. Al-Shamayleh[2], M. Phiri[3], K. Maswadi[4], D. K. Kwaghtyo[1], M. Mulenga[5], C. J. Iyidobi[6]

[1]Department of Computer Science, Faculty of Computing, Federal University of Lafia, Lafia, Nasarawa State, Nigeria.
[2]Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan.
[3]Department of Information Technology, School of Social Science and Technology, University of Lusaka, Lusaka, Zambia.
[4]Department of Management Information Systems, Jazan University, Jazan, Saudi Arabia.
[5]Business Studies Division, National Institute of Public Administration, Lusaka, Zambia.
[6]Department of Electrical and Electronic Engineering, Enugu State University of Science and Technology. Agbani, Enugu State, Nigeria.

ABSTRACT: Integrating machine learning with Mobile Big Data (MBD) offers unprecedented possibilities for deriving insights from the massive volumes of data produced by mobile devices, sensors, and services. Traditional data analytics methods struggle with scalability. This is due to the increasing volume, velocity, and variety of mobile data. Furthermore, unstructured data types, including text, images, and sensor readings present in mobile data, may introduce more challenges for traditional analytics techniques. Mobile Big Data analytics based on machine learning provide answers to these problems by utilizing sophisticated algorithms that can scale effectively to process enormous volumes of data in real time, handle unstructured data formats, and handle shifting patterns. This study presents a comprehensive review of state-of-the-art machine learning-based MBD analytics. This survey also considers the taxonomy of Mobile Big Data analytics based on various classification parameters, including data sources, data preprocessing and feature engineering, machine learning algorithms, evaluation metrics and methods, and model deployment and real-world applications. This review addresses the challenges faced by practitioners and researchers; sheds light on the state of machine learning-based MBD analytics today and suggests future research avenues to further the subject. Researchers and practitioners will find this survey report to be a useful resource for analysing and developing more advanced and effective machine learning-based Mobile Big Data analytics.

KEYWORDS: Mobile Big Data, Machine learning, State-of-the-art applications, Taxonomy, Challenges, Future research directions.

[Received Feb. 10, 2025; Revised June 29, 2025; Accepted July 16, 2025] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

The rapid growth of handheld devices combined with the rise in data acquisition from various sources, including services and sensors, has resulted in the creation of Mobile Big Data (MBD), a wealth of data with enormous analytical and insight possibilities. With over a billion cell phones in use today, a significant amount of data is generated daily. This creates commercial opportunities and has far-reaching effects on society and social interaction. With the increasing mobility like sensors carried by objects in motion automatically generate high volume of data as IoT devices evolve rapidly. These data volume, velocity, and variety are fast growing and will replace the existing standards for data analytics often used by businesses and academics. As a result, MBD has come to stay as part of our daily lives and is fast spreading. This exponential increase in the volume of data caused by the rise in network connectivity reflects the exponential growth observed in the Moore's Law for semiconductors (Fettweis & Alamouti, 2014). Studies and industry forecasts suggest that data volume may hit 660 zettabytes by 2030 (Fullerton et al. 2024). This support the concept of MBD which centres on the huge amount of data generated by mobile devices, that over powers the processing capacity of single machines (Guo *et al*., 2016).

In the current era, particularly with the widespread adoption of mobile devices like smartphones and IoT gadgets in the 4G and 5G eras (Agiwal *et al*., 2016; Alsheikh *et al*., 2016), MBD assumes and will continue to assume a crucial role, underscoring its increasing significance. Diverse datasets generated by different devices are demonstrating exponential growth patterns due to the rapid increase in computing devices (Li & Zhou, 2015). This makes big-data a global and essential strategic resource with diverse potential applications in several industries (Kučera, 2023). As the amount of data grows in this age of big data, various tools for data analysis are encountering significant difficulties. As such, MBD analytics is a highly

*Corresponding author: afolabiilesanmi@yahoo.com
doi: http://doi.org/10.63746/njtd.v22i4.3385

concentrated issue at the moment. The development of intricate mobile systems that facilitate a wide range of intellectually interactive services, such as social media, intelligent energy networks, smart homes, and medical treatment, is what makes MBD analytics so important (Alsheikh *et al*., 2016). Thus, the relationship involving MBD and ML has become more important, driving the design of innovative applications in several fields. MBD analysis involves extracting insights from terabyte or petabyte-scale data gathered from mobile users and wireless devices, either at the network or application level. The objective is to uncover previously unknown, latent, and significant patterns and knowledge using large-scale machine-learning techniques (Yazti & Krishnaswamy, 2014). Current MBD requirements emphasize a software-defined approach to enhance scalability and flexibility, anticipating a more intricate and interconnected M-Internet environment in the future (Zhao *et al*., 2021). To achieve this, MBD data centres must gather user statistics from millions of users and employ effective analytic methods to derive meaningful results. With the decreasing cost of data storage and the widespread availability of high-performance computing, machine learning has expanded beyond theoretical research to various big data application areas. However, there is still considerable progress needed for machine learning-based MBD analytics.

The topic of machine learning-based Mobile Big Data (MBD) analytics is developing at the nexus of cutting-edge analytics and the abundance of data produced by mobile devices, sensors, and services. Several Internet firms have integrated machine learning technology into their services, which range from searching the Internet (Pinargote-Ortega *et al*., 2024) to information filtering (Zhang *et al*., 2016) and recommendation systems (Zhang *et al*., 2015) on social media sites, e-commerce sites or rival distribution networks. This implies that traditional data analytics approaches struggle to manage the scale, speed, and complexity of Mobile Big Data. Mobile Big Data is characterised by high velocity, volume, and variety. As a result, traditional analytics frameworks often fail to deliver timely and scalable insights. This requires integrating ML techniques for robustness and adaptability. ML techniques fits in here due to their ability to autonomously learn patterns from vast datasets and adapt to the evolving data distributions properties in real-time, while discovering hidden patterns from structured or unstructured data sources (Alsheikh *et al*., 2016; Zhao *et al*., 2021).

Comparatively, ML algorithms work autonomously unlike the rule-based methods which are manual-based. ML approaches can adapt in real-time as data patterns changes, an essential strength needed for handling mobile data with high-velocity (Yazti & Krishnaswamy, 2014). Most especially, deep learning architectures handle diverse input types like text, images, sensor signals among others, uncover latent structures from unlabelled data, and perform scalable inference across distributed environments (Georgiadou *et al*., 2020; Rayan *et al*., 2022). These qualities make ML a necessity for extracting timely, context-aware, and actionable insights from MBD. Therefore, integrating ML with MBD enhances analytics that are ordinarily unattainable using the traditional methods (Cheng *et al*., 2017; Guo *et al*., 2016). Thillaigovindhan *et al*. (2024), further explored this trend dwelling on how machine learning algorithms are being used to address critical mobility and connectivity challenges in 5G environments. Consequently, the application of machine learning methods to MBD has shown outstanding opportunities for deriving knowledge and insights from this heterogeneous dataset, influencing applications like smart city planning, medical analytics, and personalised suggestions (Cheng *et al*., 2017). While machine learning-based methods are extensively utilized in MBD and demonstrate commendable performance in real data tests, there is a need for further development in the current methods. Recently, El-Sayed *et al*. (2025) stressed the impact of integrated data pipelines and scalable deployment mechanisms. This reinforces the need for an end-to-end approach in ML-based MBD analytics.

Historically, existing studies including (Chen *et al*., 2014; Hadi *et al*., 2018; Pouyanfar *et al*., 2018; Stergiou & Psannis, 2016) have been conducted on big data analytics. Specifically, Hadi *et al*. (2018) examined the impact of big data analytics on the architecture of data communication networks. The authors asserted that combining it with the traffic and control layers of the network may improve its intelligence, performance, and robustness. In a related study, Stergiou & Psannis (2016) combined two technologies (Mobile Cloud Computing (MCC) and IoT) with Big Data to investigate their shared characteristics and determine which benefits of MCC and IoT improve the utilisation of Big Data applications. Also, Pouyanfar *et al*. (2018), carried out a thorough review of the most recent works on big data analytics for multimedia. Describing current big data frameworks, their applications in multimedia analysis, the benefits and drawbacks of current techniques, and prospective future directions in the field of multimedia big data analytics. The study aimed to close the gap between the difficulties in multimedia and big data solutions.

In spite of the rising number of scholarly works on Mobile Big Data (MBD) analytics, existing works frequently lag in 3 critical aspects. First, most prior reviews concentrate on traditional data analytics approaches and do not adequately explore the transformative role of machine learning in MBD environments (Chen *et al*., 2014; Stergiou & Psannis, 2017). Second, they frequently lack a structured taxonomy that encompasses the full MBD pipeline from data sources to real-world deployment, particularly omitting key stages such as data pre-processing, feature engineering, model evaluation, and deployment (Iorzua *et al*., 2025; Pouyanfar *et al*., 2018). Third, the contextualisation of ML techniques across diverse mobile application domains (e.g., healthcare, finance, IoT) is either missing or superficially treated. These oversights negatively affect researchers and practitioners from gaining a nuanced understanding MBD analytics. This limitation is tackled in this study by providing an all-inclusive taxonomy and synthesis of current ML-based MBD analytics across various domains. Particularly helpful in guiding future research and implementation strategies.

The key contributions of this study are as follows:

i. A description of core technologies and traits of Mobile Big Data, including data velocity, volume, variety, volatility, and veracity is presented.

ii. This study provides a comprehensive survey of the State-of-the-art Applications of ML-based mobile Big Data Analytics.
iii. Identification of the impact of ML capabilities in various domains such as Mobile Recommender Systems, Location-Based Services and Geospatial Analytics, Mobile Health and Healthcare Analytics, Mobile Finance and Commerce, Mobile social media and Sentiment Analysis, Mobile Security and Anomaly Detection, Transportation and Mobility Analytics, Mobile Sensing and the Internet of Things (IoT).
iv. The taxonomy of Mobile Big Data analytics in relation to various classification parameters, including data sources, data preprocessing and feature engineering, machine learning algorithms, evaluation metrics and methods, and model deployment and real-world applications
v. Identification of current research gaps and recommendations for future lines of inquiry to address challenges in machine learning-based Mobile Big Data analytics.

The remainder of this study is outlined thus: Section II provides the characteristics of Mobile Big Data. In Section III, the current state-of-the-art applications of machine learning-based MBD analytics across various domains are described. Section IV presents a taxonomy of Mobile Big Data analytics techniques, methodologies, and frameworks. Challenges and Open Issues in ML-based MBD analytics are provided in Section V. In Section VI, a potential avenue for further research and development in ML-based MBD analytics is presented. Section VII concludes the paper.

II. MOBILE BIG DATA: BASIC TECHNOLOGIES AND CHARACTERISTICS.

A. Basic Mobile Big Data Technologies

The primary modern technologies that have led to the demanding era of Mobile Big Data (MBD) include expansive and fast mobile networks, portability, and crowdsourcing (Alsheikh *et al*., 2016; Dimou *et al*., 2022). The proliferation of mobile devices alongside the development of rapid mobile networks, such as WiFi and cellular networks, results in the emergence of substantial and progressively more competitive mobile data traffic (Gooderham *et al*., 2022). Each of these technological advancements plays a significant role in shaping the characteristics of MBD as described below:

i. **Large-scale and high-speed mobile networks**

The core elements of Mobile Big Data (MBD) ecosystems are large-scale, fast mobile networks, as demonstrated by technologies such as Wi-Fi and cellular networks (Chen *et al*., 2014). The number and speed of mobile data traffic have significantly increased as a result of the widespread use of mobile devices and the development of these networks (Mahdi *et al*., 2021). The rapid increase in data traffic poses several obstacles as well as opportunities for machine learning and data-driven analytics (MBD) by facilitating the real-time acquisition of large datasets from many sources. However, to efficiently manage and analyse this enormous and fiercely competitive mobile data traffic, advanced tools and approaches are needed to extract insightful information and speed up decision-making processes in a variety of fields (Sree Ranjani, 2021). The proliferation of mobile devices and high-speed mobile networks (such as Wi-Fi and cellular networks) brings about substantial and progressively competitive mobile data traffic (Eke *et al*., 2023).

ii. **Portability**

Every handheld device has the freedom to shift autonomously across various locations, rendering Mobile Big Data (MBD) non-stationary (characterized by volatility) (Chen *et al*., 2014). Because of mobility, the amount of time that gathered data is useful to make decisions can be fairly short. MBD analysis has to be performed often to handle an inflow of freshly acquired data items (Taleb *et al*., 2021). Mobile devices' portability allows them to collect data from a variety of settings and scenarios, enriching the datasets utilized for analysis and decision-making in a variety of fields.

iii. **Crowdsourcing**

In Mobile Big Data (MBD), crowdsourcing has become a prominent trend that enables widespread sensing and large-scale data gathering from several participating users (Sree Ranjani, 2021). Crowdsensing is different from standard mobile sensing systems in that it collects data from people using their own mobile devices in different places (Kolonia & Forsberg, 2020). This methodology facilitates the acquisition of heterogeneous and widely dispersed datasets, augmenting the depth and scope of information accessible for scrutiny in machine learning applications. To assure the quality and integrity of the data gathered, crowdsourcing in MBD requires strong methods for data validation, aggregation, and privacy protection, in addition to offering opportunities and problems.

B. Basic Mobile Big Data Characteristics

Mobile Big Data exhibits different characteristics inform including large-scale and high-speed massive heterogeneous datasets, volatility, veracity, and variety(Alsheikh *et al*., 2016; Anagnostopoulos *et al*., 2016). These characteristics are a result of the platforms that generate these data and the importance attached to them by organizations for decision-making. Figure 1 depicts the schematic diagram of the characteristics of Mobile Big Data analytics. The most enduring characteristics of MBD are:

i. **Volume**: The volume defines the large dataset generated by different platforms such as Wi-Fi, Mobile social networks, the Internet of Things, and cellular networks. Benseny et al. (2023) estimated multi-device daily volume could reach 1.6 PB per day with consumers devices broadband traffic being a dominant contributor.
ii. **Velocity**: Velocity describes the rate of data generation at a steady speed. This affects the latency of mobile devices as it increases the queue time and degrades quick decision-making. In this regard, extensive research has been done on how to leverage the power of mobile cloud (Aminzadeh *et al*., 2015).
iii. **Veracity**: Veracity entails the trustworthiness, accuracy, and quality of data collected from mobile devices like smartphones, sensors, and IoT-enabled devices (Hashem *et al*., 2015). Not only does the type of mobile data like video, images, text, and audio that matters, but the

structured, semi-structured, and unstructured formats in which they are generated are important. Additionally, the relatively short time lag necessary for decision-making further highlights the importance of veracity.

iv. **Volatility**: Volatility is simply dealing with the rate at which mobile data changes over time leading to loss in value or relevance with time. In mobile analytics, data is often generated in real-time to address this challenge. For instance, datasets from the Global Positioning Systems (GPS), can become out-of-date within minutes or hours.

v. **Value**: Value specifies the actual discovery of value from MBD. That is to say that Mobile Big Data value defines the services the analytics process provides the mobile users and revenue generation for mobile network operators. For instance, Mobile Big Data analytics for mobile marketing can be measured based on increased customer patronage, while geo-location data analysis will be on enhanced traffic management and reduced crime rate.

III. STATE-OF-THE-ART APPLICATIONS

Recent advancements in machine learning-based Mobile Big Data (MBD) analytics have significantly transformed various sectors. Key applications include personalised recommendation systems, predictive IoT maintenance, mobile health monitoring, GPS-based transportation optimisation, and targeted retail strategies. These innovations demonstrate the potential of ML-driven MBD analytics to support data-driven decision-making across healthcare, transport, e-commerce, and more. This section offers a comprehensive review of state-of-the-art applications across domains like mobile recommender systems, location-based services, mobile finance, security, and IoT, as categorised in Figure 1.

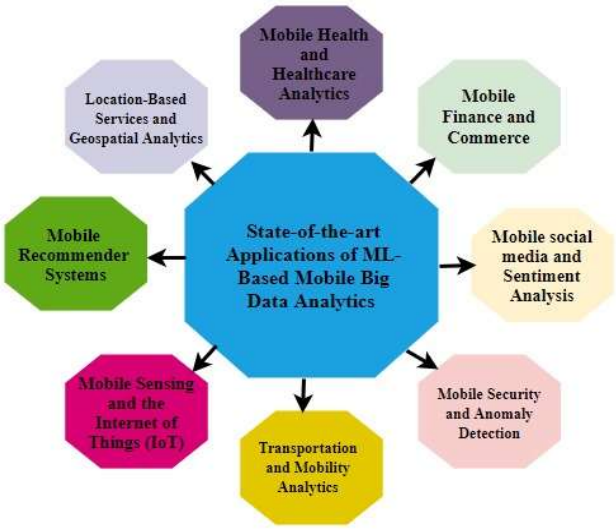

Figure 1: State-of-the-art applications of ML-based Mobile Big Data analytics

A. Mobile Recommender Systems

Mobile recommender systems are pivotal in boosting user experience as well as personalization in mobile applications. These systems analyse user behaviour, preferences, and historical data, providing tailored recommendations that align with individual user interests (Guruge *et al*., 2021). The following state-of-the-art applications of mobile recommender systems showcase advancements that significantly enhance the personalization, and efficiency of content recommendations. In line with the foregoing, (Chiu *et al*., 2021) presented a method comprising two essential components: (1) the use of unsupervised NLP techniques to analyse data given by users, and (2) the incorporation of a recommendation system utilizing Deep Learning (DL) to provide customized solutions to customers. This method shifts customers from passive service receivers to active data producers, participating in a cooperative method of creating value with the service vendors. The study focused on tourist recommendations, which validated the efficacy of the proposed technique. Principally, the study contributed by devising the personalised smart Product-Service System (PSS) method, establishing a mutually beneficial situation for all participants involved in the process. Ihnaini *et al*. (2021) presented a sophisticated healthcare recommendation system tailored for diabetes, integrating both deep machine learning and data fusion approaches. Through the use of data fusion, the system alleviates computational burdens, improving its accuracy in predicting diabetes. The implementation includes training an ensemble machine learning model specifically for diabetes prediction. The system undergoes evaluation using a well-known diabetes dataset and is compared with recent literature advancements. Remarkably, it attains an accuracy of 99.6 percent, surpassing current deep machine learning methods and highlighting its efficacy in diabetes prediction.

B. Location-Based Services and Geospatial Analytics

Location-based services (LBS) and Geospatial Analytics (GA) play a crucial role in mobile applications, providing a range of functionalities that leverage the user's geographical position. The integration of location-based services enhances user experiences, offers personalised content, and enables various practical applications (Schmidtke, 2020). Recent studies, such as Thillaigovindhan *et al*. (2024) experimented how ML models can enhance performance in 5G mobile networks. They investigated how reinforcement, supervised, and hybrid learning models tailored for seamless mobile connectivity, reinforcing the real-time geospatial relevance of ML in mobile applications. Consequently, Bellini *et al*. (2019) introduced the Snap4city platform, a system designed to obtain extensive datasets from different platforms. This infrastructure integrates the data into a semantic structure, aligning with the Km4City multi-ontology, and offers Location-Based Services (LBSs) for numerous users. Currently in operation across various geographical contexts, such as towns like Antwerp, Helsinki, Santiago de Compostela, and other Italian cities, the Snap4City platform covers whole geophysical zones including Sardinia, Finland, Belgium, Tuscany, and Garda Lake. The authors talk about preliminary findings for these implementations-based system validations. With respect to machine learning, Khan *et al*. (2023) employed four ML methods: the generalized linear model, LR, DL, and gradient-boosted trees. These models underwent design, testing, and evaluation using various

metrics, such as Receiver Operating Characteristics, Accuracy, Recall, Precision, F-score, and Sensitivity. The findings illustrate the high effectiveness of these ML techniques in perfectly categorizing data. DL models outperformed other approaches yielding an accuracy of 99 percent for both single and multiclass predictions.

C. Mobile Health and Healthcare Analytics

The healthcare sector has undergone transformations via mobile bigdata applications/analytics. This dynamic and impactful process brings about improvements in patient care, and overall patient management (Oyeniyi, 2024). ML-based Mobile Big Data analytics in mobile health (mHealth) and healthcare analytics are at the forefront of innovations aimed at improving patient outcomes, optimizing healthcare delivery, and enhancing overall healthcare efficiency (Liu & Wang, 2025). Nguyen *et al*. (2019) introduced a framework for handling Electronic Health Records (EHRs), utilizing a combination of blockchain and the decentralized Interplanetary File System (IPFS) through a mobile environment. The framework demonstrates the use of Ethereum in a data-distribution state in mobile applications hosted on Amazon. The outcome shows that the framework offered efficient and secure data sharing in a mobile environment, ensuring the protection of sensitive health information from potential threats. Leveraging big data analytics, Wang *et al*. (2018) introduced a transformation model. The model unveils the cause-and-effect connections among capabilities in big data analytics, IT-enabled transformation practices, dimensions of benefits, and business value. To validate its applicability, the model was tested in healthcare, where 3 significant pathways to value creation for healthcare organizations were identified, offering a practical insight as a decision-support system.

D. Mobile Finance and Commerce

In the financial and commercial sectors, MBD has pioneered transformations that redefines how businesses operate, supporting decision-making, and engaging customers (Famoti *et al*., 2025). Sarker (2019) affirmed that MBD have played a transformative role that enhance user experiences, optimize processes, and contribute to the evolution of the financial and commercial sectors. Shorfuzzaman *et al*. (2019) also introduced a MBD framework, employing big data analytics techniques to derive insights from datasets generated by mobile learners. Specifically, the authors put forth a hypothetical model for the adoption of mobile learning based on the indigenous extensive Technology Acceptance Model (TAM) (Buddha *et al*., 2024). Liu *et al*. (2020) suggested Cloud Laundry, an IoT-based online trade scheme, for large-scale washing services. This algorithm determines the best routes for transportation and reroutes logistic terminals using big data, and intelligent logistics management, including ML. While preserving client confidentiality and privacy, cloud laundry provides services that are visible, easy to use, and efficient. Higher capital turnover, liquidity, and profitability are features of this model that make it attractive to e-commerce firms and scholars.

E. Mobile social media and Sentiment Analysis

The impact of mobile data on social media analytics is profound, reshaping the way businesses, marketers, and researchers understand and engage with their audiences. The widespread use of mobile devices has significantly increased the volume, variety, and velocity of data generated on social media platforms (Alaei *et al*., 2019). Machine learning-based Mobile Big Data analytics in mobile social media and sentiment analysis have ushered in a new era of understanding user behaviour, sentiment, and trends on social platforms. These technologies provide valuable insights for businesses, marketers, and researchers (Guo *et al*., 2021). In another study, Guo (2022) proposed the Deep Learning Aided Semantic Text Analysis (DLSTA) method for Big-Data-based emotion identification. NLP approaches incorporate the syntactic and semantic aspects of the text, enhancing the effectiveness of AI-based systems. Experimentally, the proposed technique achieved an emotion detection rate of 97.22 percent and a categorisation accuracy of 98.02 percent on various baseline methods. As a further development, Jalil *et al*. (2022) examined the COVID-19 effect with a focus on Twitter users' attitudes toward the social media network. The suggested method uses a variety of feature sets and classifiers to assess the attitudes included in gathered tweets to classify them. A variety of parameters were used to assess the performance of the DL and ML models: accuracy, precision, recall, and F1-score. The method, when applied to publicly available datasets, such as COVISenti, COVIDSenti_A, COVIDSenti_B, and COVIDSenti_C demonstrated 96.66, 95.22, 94.33, and 93.88 percent accuracy, respectively.

F. Mobile Security and Anomaly Detection

Security is of paramount importance in mobile data analytics, given the delicate nature of the dataset involved, along with the possible dangers connected to its collection, storage, including analysis. Cutting-edge architectures of ML-based mobile Big-Data analytics in mobile security and anomaly detection are critical for identifying and mitigating potential threats and abnormal activities on mobile devices and networks (Habeeb *et al*., 2022). To offer an automated, adaptable, and scalable network defence system, Lam and Abbas (2020) leveraged SDS (Software Defined Security). Utilizing the most recent developments in machine learning, SDS will create a CNN and use Neural Architecture Search (NAS) to identify unusual network activity. A more proactive and comprehensive end-to-end defence for a 5G network can be achieved by integrating SDS with an intrusion detection system. This assumption has been tested by gathering and using a CNN to assess both normal and aberrant network traffic from a simulated environment. This approach has shown encouraging results, as the model has detected anomalous traffic having a 96.4 percent identification rate and benign traffic with 100 percent accuracy. This highlights the network flow analysis. Onah *et al*. (2021) presented a Genetic Algorithm Wrapper-Based feature selection and Naive Bayes for Anomaly Detection Model (GANBADM) in a Fog Environment. The study aimed at decreasing the time complexity while creating a more accurate model that can predict results using the Security Laboratory Knowledge Discovery Dataset (NSL-KDD). The evaluation results reveal that the model performs better overall,

with an accuracy of 99.73 percent and a false positive rate as low as 0.6 percent. The outcomes demonstrate that the suggested GANBADM technique outperforms comparable methods found in the literature.

G. Transportation and Mobility Analytics

Mobile data plays a pivotal role in shaping transportation and mobility solutions, contributing to enhanced efficiency, safety, and user experience. The integration of mobile technologies and data-driven insights is transforming how people navigate and utilize transportation services. Cutting-edge models for ML-based mobile Big-Data analytics in transportation and mobility analytics are transforming the way transportation networks function, increasing effectiveness, security, and user experience (Ahmad Jan *et al*., 2025). In another study, Malek *et al*. (2021) devised a scheme for speed forecasting leveraging the Long Short-Term Memory (LSTM) architecture. Using real-world data that represents urban itineraries, a traffic simulator's data is employed in the training of the LSTM scheme. The suggested schemes are designed for both univariate and multivariate situations, and how accurate they are at speedy predictions was assessed. According to simulation data, the multivariate model performs better in predictions, both short- and long-term, than the univariate scheme. Chen *et al*. (2022) presented a fog-based vehicular crowdsensing strategy in which automobiles can function as fog nodes and onboard sensors as sensing participants. It integrates computing job allocation and crowdsensing participant selection to simultaneously improve execution delay and cost. In addition, they introduce a heuristic technique for solving the suggested issue called random chemical reaction optimisation (RCRO). Lastly, they simulate real-world traffic to assess the suggested scheme and algorithm. The RCRO algorithm's performance and the suggested technique's superiority over the traditional way are demonstrated by numerical results.

H. Mobile Sensing and the Internet of Things (IoT)

The incorporation of portable sensing and the IoT in data analytics has significantly transformed the landscape of data collection, processing, and decision-making. This convergence of technologies allows for the seamless gathering of data from mobile devices and various IoT-connected sensors, providing rich datasets for analysis. Cutting-edge applications of ML-based Mobile Big Data analytics in mobile sensing and the IoT are reshaping various industries by providing insights, automating processes, and improving overall efficiency (Hemmati *et al*., 2024). In another study, Ren *et al*. (2023) proposed a Data Relay Mule-Based Collection System (DRMCS) to advance Quality of Service (QoS). First, the novelty of DRMCS is the data collection system that takes into account the pace at which sensing jobs are completed, redundant, and delayed. Furthermore, the Micro Mobile Data Centre (MMDC) was devised to address the challenge of linking a large number of self-learning tools to a data hub. Third, DRMCS proposed an MMDC selection approach grounded on the mimicked hardening process to increase data-gathering performance. DRMCS has a 78.6 percent higher sensing job completion rate than a typical vehicle network Opportunistic Communication Devoid of a Data Relay Mule (OCDRM). To summarise the above, Table 1 summarises key areas where ML-based MBD analytics are applied. It outlines the core contributions, AI methods used, data sources, advantages, and usage scenarios, including benefits and issues. This helps contextualise domain-specific challenges and highlights practical implementation concerns.

Table 1: State-of-the-Art Research Issues on Big Data Applications

Reference	Area of Application	Key Contributions	AI Approaches	Data Sources	Usage scenarios	Major Benefits	Problems / Issues
Ihnaini et al., 2021	Mobile Recommender Systems	Analysed user preferences to provide personalised recommendations across sectors.	Deep Learning, Regression, Classification, and Clustering	Traffic, utilities, and IoT sensor data	E-commerce, social media, finance and investment, travel and local services, entertainment, food delivery, health and fitness.	Better public safety, more effective use of resources, and better urban planning	Large amounts of data and the blending of various data
Barik et al., 2018	Location-Based Services (LBS) and Geospatial Analytics	Applied geospatial analytics to deliver real-time, location-aware services.	Deep learning, ensemble learning, regression, and classification	Logs of user activities and GPS data	Ride-sharing programmes, geofenced areas, and location-based ads	individualized user experiences, focused advertising, and instantaneous navigation	sparsity of data and privacy issues

	Mobile Health and Healthcare Analytics	Developed health forecasting models using mobile and wearable sensor data.	Deep Learning, Regression, and Classification	Both physiological and activity data	Monitoring of mental health, chronic illness, and fitness	Early health issue identification, individualized treatment, and ongoing observation	Regulation adherence and data privacy
Nguyen et al., 2019	Mobile Finance and Commerce	Used behavioural data to improve client targeting and in-store navigation.	Classification, Association Rule Learning, and Clustering	Data on transactions and user behaviour	Customer sentiment analysis, in-store navigation, and customized suggestions	Enhanced client retention, customized purchasing experiences, and focused advertising	Adaptive customer tastes and high-quality data
	Mobile Social Media and Sentiment Analysis	Extracted social behaviour insights from mobile interactions and sentiment patterns.	Deep Learning, Sentiment Analysis, Clustering, and Network Analysis	Data from social media and interactions	Identification of influencers, study of social trends, and discovery of communities	Understanding social behaviour, forecasting trends, and focusing on advertisements	Data noise and privacy
Liu et al., 2020	Mobile Security and Anomaly Detection	Introduced fraud detection techniques based on mobile transaction analysis.	Recognition, Categorisation, and In-depth Learning of Anomalies	User activity and transaction data	Payment gateways, online shopping, and mobile banking	Better security, fewer monetary losses, and instantaneous fraud detection	High prevalence of false positives and changing patterns of fraud
	Transportation and Mobility Analytics	Modelled urban mobility using real-time mobile sensor and traffic data.		Clustering, Deep Learning, and Time Series Analysis	A better travel experience, less traffic, and more effective public transportation	On-demand traffic tracking, optimized routes, and planning for public transportation	Enhanced travel experiences, less traffic, and more effective public transportation
Guo, 2022	Mobile Sensing and the Internet of Things (IoT)	Modelled urban mobility using real-time mobile sensor and traffic data.		Anomalous Detection, Deep Learning, and Time Series Analysis	Enhanced dependability, less downtime, and cost savings	Maintenance of vehicles, smart home gadgets, and industrial IoT	Enhanced dependability, less downtime, and cost savings

IV. TAXONOMY OF MOBILE BIG DATA ANALYTICS

This section presents the taxonomy of Mobile Big Data analytics based on various classification parameters, including data sources, data preprocessing and feature engineering, machine learning algorithms, evaluation

metrics and methods, and model deployment and real-world applications. Figure 2 presents a visual taxonomy that organises the MBD analytics pipeline, from data sources and preprocessing techniques to machine learning models and deployment methods.

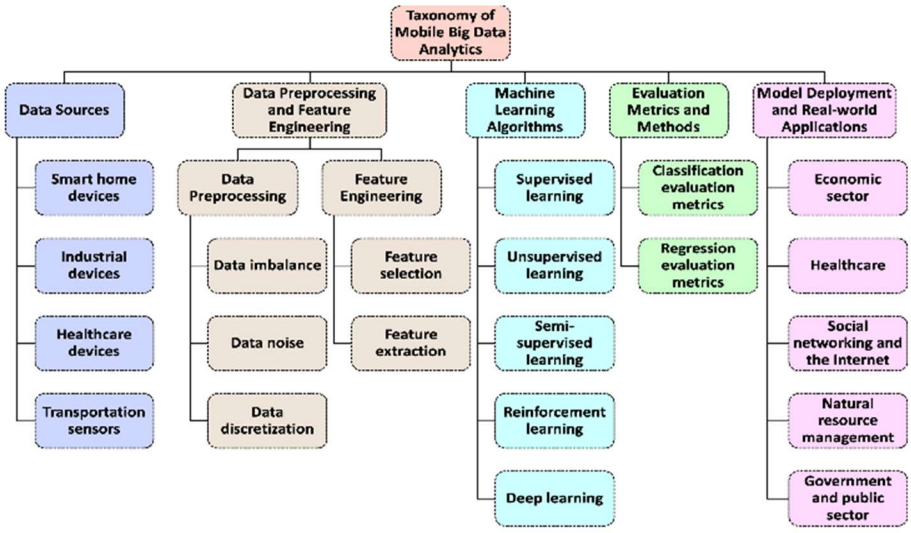

Figure 2: Taxonomy of Mobile Big Data Analytics

A. Data Sources

Data generated from IoT devices forms a substantial component of Mobile Big Data sources. Huge amounts of data are spread across several sources like IoT devices (Mohamed *et al*., 2020). IoT devices encompass a broad range of interconnected devices that are capable of collecting, exchanging, and utilizing data (Mphale *et al*., 2024). These devices comprise smart home devices, industrial devices, healthcare devices, and transportation sensors. Figure 3 summarises these device categories and highlights the typical assets monitored within each data source type.

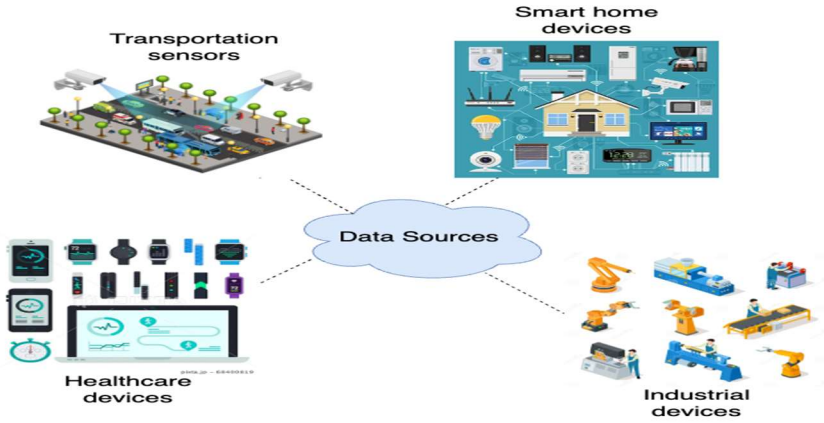

Figure 3: Data sources

Brief discussions of these device categories are presented thus:

i. Smart home devices: In smart homes, sensors embedded in thermostats, lighting systems, and appliances collect data to optimise energy usage and enhance living comfort. For instance, systems like Google Nest use real-time data from motion and temperature sensors to predict and automate climate control decisions. These are devices that are found in the house or living environments and comprise mainly everyday devices such as washing machines, toasters, smart TVs, refrigerators, door locks, security cameras, and smart hubs (Vardakis *et al*., 2024). Through the use of laptop computers, smartphones, and tablets, these smart home devices can be monitored and controlled anywhere in the world via a private network or the internet. Furthermore, smart home devices provide an ecosystem that is often integrated with virtual assistants for voice-activated control. These devices generate vast amounts of data based on the user's day-to-day usage of the appliances and are often used to improve personalization and provide an enhanced user experience based on the collected data.

ii. Industrial devices: These are devices that are used in the automation, safety, and data collection of industry-related activities. For instance, in the manufacturing industry, these tools enable producers to react quickly and

efficiently to modifications in production conditions and customer demands (Chukwunweike *et al*., 2024). Industry workers can identify early warning signs of product quality flaws as well as possible equipment failure by evaluating enormous volumes of data gathered through sensors and smart devices failure all in real time. Another industry that can be thought of as using smart technology is the agricultural industry. This is often referred to as smart agriculture, which encompasses the use of information technology to improve farm yield and productivity. Improving crop output and financial returns can be achieved, for instance, by conserving natural resources and optimizing water use. Smart farming also includes precision farming, which involves monitoring livestock for early illness identification, treatment, and nutrition adjustments (Kwaghtyo & Eke, 2022; Kwaghtyo *et al*., 2024). Typically, smart devices in the agricultural industry include livestock monitoring sensors, soil moisture sensors, irrigation controllers, greenhouse sensors, aerial drones, and atmospheric monitors.

iii. Healthcare devices: In the healthcare domain, wearable devices like Fitbit, Apple Watch, and continuous glucose monitors (CGMs) stream physiological data, enabling predictive analytics for health anomalies and chronic disease management. These devices contribute to personal health monitoring and population-level analytics in mobile health research. With the aid of these gadgets, medical professionals, including physicians, may monitor, diagnose, and treat patients from a distance via telehealth (Kiran *et al*., 2014). These devices have gained significant popularity around the world, especially in India. Implantable glucose monitoring systems, pacemakers, insulin pumps, ECGs, and infusion pumps are a few examples of these devices. Additionally, devices such as smart wearables have also gained much attention. With the help of intelligent sensors and actuators, these end-to-end connected devices can remotely link to a user's tablet, laptop, or smartphone (Raja & Chakraborty, 2023). These gadgets can be fastened to the body, placed on the wrist, or worn around the neck. While these devices are widely used in several sectors, they are more commonly used in the healthcare domain for weight loss monitoring, telemedicine purposes, and general fitness. Typical examples of such devices are smart watches, smart shoes, and smart glasses.

iv. Transportation sensors: In transportation, GPS trackers and RFID-based mobile sensors provide real-time geolocation and vehicle condition data, which are instrumental in fleet management, route optimisation, and delivery performance forecasting. These are often IoT-based sensors that develop a network of transportation that is greener, safer, and more effective (Tyagi & Sreenath, 2023). These sensors generate real-time traffic information from interconnected vehicles and roadside infrastructure. This, in turn, makes data collection, preparation, and decision-making more effective (Tyagi & Sreenath, 2023). Furthermore, the data generated from these devices improves the general quality of life by reducing traffic congestion, which translates into reduced fuel and electricity usage. Examples of such technologies include connected vehicles, traffic lights, dynamic traffic management systems, and parking guidance.

B. Data Preprocessing and Feature Engineering

The quality of knowledge extracted from data collected from the sources mentioned above depends heavily not only on the systems' architecture but also on how appropriate and high-quality the data is (Brown & Anderson, 2023). It is a well-known fact that unclean data leads to low-quality or poor knowledge gathering. Steps such as data pre-processing and feature engineering are key phases in preparing the data for further analysis and effective training of machine learning algorithms (Iorzua *et al*., 2025). The major steps in data pre-processing and feature engineering are data cleaning, transformation, and extraction methods tailored for Mobile Big Data environments.

1). Data Preprocessing

Data preprocessing is a well-known technique that is applied before any sort of analysis can be performed on the data (Brown & Anderson, 2023). The creation of preprocessing procedures that guarantee the data can be supported for efficient algorithm training takes up a significant amount of work before machine learning models are trained. Moreover, most of the assumptions held for traditional data sizes do not hold for large amounts of data. As a result, some preprocessing techniques are inapplicable. In most cases, raw data collected from various sources suffers from various challenges, such as data imbalance, noise, inconsistency, data redundancy, and labelling (for supervised and semi-supervised learning) (Huynh *et al*., 2022). The following are some of the techniques that are commonly used when handling the aforementioned in the context of big data:

i. Data imbalance*:* Data imbalance often poses a challenge for many classification tasks. In most conventional situations, the issue of data imbalance is usually addressed by traditional stratified random sampling techniques. Thus, in the case of big data, this can be a very time-consuming approach that may involve numerous repetitions of calculating error metrics and generating subsamples. Moreover, value-based sampling is among the unnamed subsets of data that traditional sampling approaches are unable to effectively sample. Sampling of data concurrently is necessary for big data. For example, (Su *et al*., 2014) proposed a parallel sampling algorithm that generates random samples from the initial dataset constructed from several dispersed index files. The size of the dataset and the procedures that are accessible can be used to determine the parallel level.

ii. Data noise*:* Data sparsity, outliers, incorrect values, and missing values are all cases that often lead to noise in the data. Conventional solutions to handling noisy data are usually inadequate in the case of big data, for instance, because manual procedures are not scalable, and they become independent. Moreover, methods like substitution by mean will ultimately forfeit the benefit of large data and fine granularity. Additionally, the process of removing or excluding data points to remedy noise in the data may not be a wise alternative, as some of the patterns may lie in these noisy data points. To alleviate some of these issues, big data

offers the solution of estimating missing values by simply correcting readings from malfunctioning sensors. Furthermore, to counteract the possible bias introduced into prediction by the collective influences of methods, a maximum entropy constraint is applied during the inference step, aligning predictions with the distribution of observed labels (Grandvalet & Bengio, 2004). Additionally, despite the persistence and potential aggravation of data sparsity in big data, the sheer volume of information presents unique opportunities for machine learning, as a diverse group of samples accumulated from various sources provides sufficient frequency to alleviate data sparsity.

iii. Data discretization*:* Using this method, a continuous domain can be divided into non-overlapping segments by converting qualitative features into other qualitative attributes. Some ML models, such as NB and DT, have been known to produce remarkable results when trained with discrete features (García *et al*., 2015). Data discretization also offers the added benefits of data simplification and readability. About Big Data, most conventional discretization methods are often inadequate, and as such, parallelization is used. For instance, utilizing the lowest description length concept to create a distributed entropy-minimizing discretizer as presented by Ramírez-Gallego *et al*. (2016). In another study by Zhang and Cheung (2014), after first being sorted according to the numerical feature values, the data is divided into pieces representing the original class features. Ultimately, these pieces, which are summed up by a percentage composition of different classes, are seen as super instances and are what discretization is aimed at. Nevertheless, though discretization presents significant benefits, this often comes at the price of significant data loss, which may be essential for model training. As such, a balance is needed when using this preprocessing technique (García *et al*., 2015).

2). Feature Engineering

Feature engineering in Mobile Big Data presents distinct challenges compared to traditional datasets. Traditional feature extraction typically deals with structured, batch-processed data, whereas MBD demands real-time processing of high-velocity, heterogeneous streams. Since the choice of data representation greatly affects the effectiveness of ML algorithms, feature engineering is regarded as one of the most important stages in the development process (Iorzua *et al*., 2025). Furthermore, due to storage and bandwidth constraints in mobile environments, feature selection must prioritise lightweight, interpretable features that enable efficient on-device processing (Emmoh *et al*., 2025). This necessitates the use of embedded feature selection methods or federated feature engineering frameworks to ensure scalability across devices.

Overall, feature engineering is also a labour-intensive task that may sometimes require prior domain knowledge and human ingenuity. To overcome the limitations of conventional feature engineering algorithms when dealing with big data, several alternatives have been proposed, including distributed feature selection (Theng & Bhoyar, 2024). Additionally, low-rank matrix approximation, such as the standard Nyström method, has also been used. Another approach involves representation learning to reduce dependence on feature engineering by learning a generic prior. For ultra-high-dimensional feature selection, an adaptive feature scaling scheme is employed, which iteratively activates feature groups and addresses multiple kernel learning sub-problems. Furthermore, there is a unified framework for feature selection based on spectral graph theory. This framework can generate algorithmic families for both supervised and unsupervised feature selection (Ding *et al*., 2024). To reduce data dimensions and volumes, techniques like Random Forest-Forward Selection Ranking, Random Forest-Backward Elimination Ranking, and a linguistic hedges neuro-fuzzy classifier with selected features are employed.

C. Machine Learning Algorithms

Numerous machine-learning techniques are currently in use and have been applied to typical data sizes (Phiri *et al*., 2023). However, given the exponential growth of data generated at every moment, conventional machine learning algorithms fail to perform optimally when trained on unprecedented volumes of data. As a result, numerous parallel ML models have been put forth to manage large amounts of data. These algorithms make use of MapReduce or its various variations. In general, there are several categories into which ML models can be categorized, including reinforcement learning, unsupervised learning, semi-supervised learning, supervised learning, and deep learning.

1). Supervised learning

Supervised learning (SL) refers to a case where annotated data is employed in training the ML model. The SL approach is aimed at learning a mapping function that shows how the feature and target variables are related. The SL approach can subsequently be grouped into two subcategories, namely, classification and regression. Classification is a form of supervised learning employed to predict a future pattern. As is the case with machine learning tasks, classification also involves two phases, model training and testing for validation (Eke *et al*., 2022). This technique finds utility in numerous application domains such as fraud detection, image and speech identification, and spam filtering. Conversely, regression is employed for prediction tasks and is distinctively used to forecast continuous target variables within a dataset. About big data, classification and regression algorithms are often adapted to the MapReduce framework. Examples of such algorithms include DT, RF, NB, SVM, and multiple linear regression.

2). Unsupervised learning

Unsupervised learning approach aims to learn the intrinsic representation or underlying patterns of a dataset to uncover hidden insights and relationships using unlabelled data. This approach is frequently employed for tasks such as dimensionality reduction, clustering, and generative techniques. Among these, clustering stands out as the most common form of unsupervised learning, finding use in pattern classification, document retrieval, image segmentation, and customer segmentation (Ezugwu *et al*., 2022). Utilizing the MapReduce architecture, some

traditional clustering techniques with huge data have been improved.

3). Semi-supervised learning

This type of learning integrates both unsupervised and supervised learning techniques. This method offers robustness and improved accuracy by using both labelled and unlabelled data to train the model. This method works effectively with large data sets where most of the data is unlabelled, and labelling would be extremely expensive. Co-training and active learning are two of this learning approach's most often utilized strategies. With training, two separate views are extracted from each example as part of a semi-supervised learning technique. It entails training distinct classifiers for every view, and then labelling the learned data with the best reliable predictions made by these classifiers on unlabelled data. Cleuziou (2024) presented a multi-view learning technique for MapReduce scenarios that is based on co-training. The authors suggested using data structures, including bidirectional reducers, label files, and mapping tables, to enhance the MapReduce processing process. On the other hand, active learning minimizes learning effort by having the learning algorithm dynamically query a source for particular occurrences. To improve the BP model's effectiveness in learning, Zhao *et al*. (2013) put forth the Punishing-Characterized Active Learning Back-Propagation (PCAL-BP) neural network technique for big data.

4). Reinforcement learning

Reinforcement learning entails a continuous learning via trial and error where actions that result in the highest rewards to enhance decision-making are identified. A good example of reinforcement learning include the Markov Decision Process often used in robotics and gaming applications. Bashabsheh *et al*. (2022) utilized the MapReduce framework to implement some techniques, including value iteration, policy evaluation, and policy iteration. This involved distributing computations, especially large matrix multiplications, across the framework for efficiency. The concept of parallelism was employed to accelerate large matrix factorisation, though the learning process itself was not parallelised. Also, methods like the Bayesian Reinforcement Learning techniques were proposed in the literature to concurrently reinforce learning where agents learn simultaneously to enhance convergence (Bashabsheh *et al*., 2022).

5). Deep learning (DL)

DL, an advanced machine learning method has exhibited potentials in addressing various challenges in diverse fields. It leverages its inherent learning mechanism characterized by nonlinear and multi-layered processing. Additionally, DL has demonstrated impressive outcomes in image processing, speech identification, and NLP, rendering it an appealing subject of research across diverse domains (Mulenga *et al*., 2023). Similar to traditional machine learning, DL can be categorized into supervised, unsupervised, and semi-supervised types. A widely adopted framework for deep learning is the RBMs are used in conjunction with a Deep Belief Network (DBN) to perform unsupervised learning and provide a training model that is then used by the DBN for supervised classification or fine-tuning. Zhang and Chen (2014) presented a MapReduce-based distributed learning architecture. Distributed learning is used in the pretraining phase, where distributions of RBMs are stacked, and then distributed backpropagation using the MapReduce framework is used for fine-tuning.

D. Evaluation Metrics and Methods

In working with big data, the measure of a machine-learning algorithm's ability to classify or predict targets effectively is vital for ensuring accurate decision-making and optimizing overall system performance. Numerous evaluation metrics and methods can be used to evaluate the performance of machine learning algorithms. However, the type of machine learning algorithm and target variable play a huge role in determining the most appropriate evaluation metrics in a machine learning task. Though the conventional accuracy metric is commonly reported in several ML tasks. It is often insufficient in Mobile Big Data analytics, where datasets are highly imbalanced, dynamic, and resource restricted. Evaluation metrics like F1-score and Area Under the ROC Curve (AUC) becomes necessary to balance the trade-off between false positives and false negatives in tasks like health monitoring (Diallo *et al*., 2024). Also, latency and throughput are essential metrics in MBD, where real-time or near-real-time response is required. For instance, a delayed classification can render decisions obsolete in mobile security and traffic analytics. As such, model performance evaluation in MBD must consider both predictive and computational performance. By usage, classification and regression scenarios have suitable evaluation metrics, and as outlined subsequently.

1). Classification

In the model evaluation phase, the machine learning classifier predicts the class of the target variable using the training dataset. The classifier's ability to effectively classify the target label can be estimated by evaluating.

i. The proportion of positive instances accurately identified as positive. [True positive (TP)].
ii. The proportion of negative instances accurately identified as negative. [True negative (TN)].
iii. The proportion of negative instances inaccurately identified as positive. [False positive (FP)].
iv. The proportion of positive instances inaccurately identified as negative. [False negative (FN)].

These four conditions form the components of the confusion matrix, illustrated in Table 2.

Table 2: Confusion matrix

Actual Category	**Predicted case**	
	Positive case	**Negative case**
Positive case	TP	FN
Negative case	FP	TN

The confusion matrix forms the basis for other evaluation metrics. The most commonly used evaluation metrics for classification tasks are summarised in Table 3 along with their formula and a brief description.

Table 3: Common evaluation metric for classification cases

Evaluation metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Accuracy denotes the classification model's ability to accurately predict both positive and negative instances, representing the ratio of correct predictions.
Precision	$\frac{TP}{TP+FP}$	Precision represents the percentage of instances from the positive class that are accurately labelled as positive.
F-Measure	$\frac{2*Precision*Recall}{Precision+Recall}$	F-Measure, also known as F-score, is a measure of the harmonic mean of recall and precision.
Sensitivity/Recall	$\frac{TP}{TP+FN}$	Sensitivity or recall represents a classifier's ability to accurately identify the positive class correctly.
Specificity	$\frac{TN}{TN+FP}$	Specificity represents a classifier's ability to accurately identify the negative class correctly.
False Positive Rate (FPR)	$\frac{FP}{TN+FP}$	It is the ratio of false positives (negative samples incorrectly classified as positive) to the overall number of actual negative samples.
False Negative Rate (FNR)	$\frac{TP}{TP+FP}$	It is the ratio of false negatives (positive samples incorrectly classified as negative) to the total number of actual positive samples.
Geometric Mean (G-Mean)	$\sqrt{Precision*Recall}$	The classifier can maintain a balanced classification accuracy for both positive and negative classes.

2). Regression

For regression problems too, several metrics can be used to evaluate their model performance. This section therefore outlines the six most utilised evaluation metrics, as summarised in Table 4, along with their mathematical representations and a brief description.

Table 4: Common evaluation metric for regression cases

Evaluation metric	Formula	Description
Mean squared error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(m_i - \hat{m}_i)^2$	MSE quantifies the average magnitude of the differences between real observations (m_i) and estimated observations ($\hat{m}_i$).
Root mean squared error (RMSE)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \hat{m}_i)^2}$	The RMSE is the square root of the MSE and is employed when tolerance for small errors is desired.
Mean absolute error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert m_i - \hat{m}_i\rvert$	The MAE metric computes the average absolute difference among M predicted vectors.
Mean percentage error (MPE)	$MPE = \frac{1}{n}\sum_{i=1}^{n}\frac{m_i - \hat{m}_i}{m_i} \times 100\ \%$	The MPE evaluates the percentage deviation between the estimated and real observations.
Mean absolute percentage error (MAPE)	$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{m_i - \hat{m}_i}{m_i}\right\rvert \times 100\ \%$	MAPE calculates the average of MPE.
Coefficient of determination (R^2)	$R^2 = 1 - \frac{\sum_{i=1}^{N}(m_i - \hat{m}_i)^2}{\sum_{i=1}^{N}(m_i - \overline{m}_i)^2}$	The R^2 represents the proportion of total variation in the target variable explained by the independent variables in regression models.

E. Model Deployment and Real-world Applications

Numerous Big data sites were developed to provide businesses or users with applications for managing large-scale data, aiding in tasks such as traffic management, pollution control, safety, social management, health, and disaster management. These platforms handle the exponential growth of data originating from several sources, such as the Internet, sensors, social networks, and purchase transactions.

V. CHALLENGES AND POSSIBLE SOLUTIONS

Despite significant progress in the application of machine learning (ML) to Mobile Big Data (MBD) analytics, several key challenges continue to hinder the full realization of its potential across domains. This section prioritises and contextualizes these challenges based on their severity and practical impact in real-world applications, while briefly identifying current or emerging solution directions supported by literature:

A. Privacy and Security

Challenge: Because user data acquired by mobile devices is sensitive by nature, machine learning models in Mobile Big Data (MBD) analytics face privacy and security challenges. Large volumes of user-generated data, such as surfing history, location data, and communications logs, are constantly gathered by these devices. Because improper use or unauthorized access to this private data can result in serious privacy violations, collecting and analysing such data using machine learning models presents serious privacy problems (Rafiq *et al*., 2022). Furthermore, Data privacy and user security are at risk because adversaries can alter model behaviour or infer sensitive information about users by taking advantage of flaws in machine learning models (Cabrero-Holgueras & Pastrana, 2021).

Application: This challenge is particularly relevant in mobile health and finance applications, where privacy breaches may result in regulatory violations, reputational damage, and the erosion of user trust. In mobile recommender systems and sentiment analysis, improper handling of personal data could also lead to profiling and discrimination.

Potential Solution: Federated learning is increasingly recognised as a powerful tool for preserving data privacy in mobile environments (Nguyen *et al*., 2019; Rafiq *et al*., 2022). Other robust methods include differential privacy and secure model deployment strategies that mitigate adversarial attacks. These techniques are crucial in domains like mobile health and finance, where privacy breaches can have legal and ethical repercussions (Onah *et al*., 2021). An approach integrating cryptography approaches, access control mechanisms, and a safe model deployment procedure is necessary to tackle the privacy and security challenges in MBD ML models.

B. Legal and Regulatory Compliance Challenges

Challenge: Legal and regulatory compliance pose significant and multifaceted challenges for ML-based MBD analytics. These challenges originate from the complex nature of machine learning algorithms, the rapid pace of technological advancements, and the diverse applications of these technologies across various industries. Key concerns about legal and regulatory frameworks comprise difficulties in understanding how models generate results, risks of

biasedness or discriminatory outcomes, data privacy and confidentiality concerns, and the challenge of ensuring explainability and robustness of AI algorithms (Albahri *et al*., 2023). Also, the complexity and unexplainability of these algorithms make it challenging to mitigate the aforementioned concerns. Furthermore, existing supervisory and regulatory measures may need clarification and possible readjustment to address the perceived incompatibilities of existing arrangements with AI applications (Albahri *et al*., 2023).

Application: Regulatory and legal compliance challenges are especially critical in healthcare, finance, and smart city applications, where non-compliance with frameworks like the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA), or financial data governance policies, may hinder real-world deployment and public acceptance. Legal frameworks like GDPR and HIPAA are directly relevant to mobile applications that involve personal data collection, such as in healthcare and finance.

Potential Solution: As AI and machine learning continue to evolve, policymakers must adapt regulations to the growing landscape of these technologies, particularly in the context of mobile devices and Mobile Big Data analytics. As noted by Khan *et al*. (2023), there is a growing body of research advocating for ethics-aware AI, transparent modelling, and auditable systems to bridge the gap between AI capabilities and legal accountability.

C. Scalability and Real-Time Analytics

Challenge: The volume and velocity of data created by mobile devices pose substantial problems for machine learning models in Mobile Big Data (MBD) analytics, particularly in terms of scalability and real-time analytics. Massive volumes of data are produced in real-time by mobile settings, including location updates, sensor readings, and user interactions. Machine learning algorithms that are scalable to high throughput and velocity of data streams are needed to process and analyse this data quickly enough to derive actionable insights (Rekha *et al*., 2020). Building and running complex machine-learning models on mobile devices is challenging. This is because mobile devices often have limited memory and processing capacity. The issue of scalability arises from the necessity of creating machine learning models that are accurate and lightweight, yet able to function in mobile environments with limited resources (Anawar *et al*., 2018). Recent reviews have also emphasized the practical engineering challenges associated with full-cycle ML deployments in mobile big data ecosystems. These include pipeline orchestration, streaming feature engineering, and model versioning, especially under real-time and resource-constrained conditions (El-Sayed *et al*., 2025; Iorzua *et al*., 2025). Scalability also affects recommender systems and smart cities, where large volumes of user data must be processed in real-time for timely decision-making.

Application: The need for real-time processing is most evident in transportation and mobility analytics, where even minor latency can disrupt routing efficiency, commuter safety, or service availability. For time-sensitive applications like location-based services, tailored suggestions, and predictive maintenance, scalability and real-time analytics are essential for processing data streams as they come in, deriving insights, and making decisions almost instantly (Navaz *et al*., 2018).

Potential Solution: To tackle real-time processing limitations, researchers have proposed edge-based architectures and stream-based ML models that allow for low-latency processing directly on the device (Malek *et al*., 2021). Solutions such as incremental learning and online decision trees can support continuous learning from streaming data, while cloud-edge hybrid systems balance performance with resource availability (Anawar *et al*., 2018; Rekha *et al*., 2020). To facilitate prompt and precise decision-making on mobile devices, streaming machine learning models and methods that can handle data streams effectively and adaptively must be developed to achieve real-time analytics in MBD contexts.

D. Resource Constraints on Mobile Devices

Challenge: Machine learning-based Mobile Big Data analytics faces a significant challenge with mobile device resources. Machine learning algorithms are generally memory-intensive and computationally expensive, making them unsuitable for resource-constrained environments such as mobile devices. These constraints include memory, communication, limited runtime, and energy resources, which affect the execution of machine-learning algorithms on various mobile devices (Braun & Möller, 2024). Overall, the intersection of machine learning and resource-constrained mobile devices requires innovative approaches to optimize algorithms and models for efficient execution within limited resources.

Application: This challenge is particularly important in mobile health, IoT sensing, and recommender systems, where edge devices and wearables operate with limited battery, storage, and processing capabilities, making the deployment of resource-intensive models impractical without optimisation. Moreover, domains like mobile sensing and recommender systems employ on-device processing to balance computational efficiency with accuracy. In such scenarios, high data veracity is required (Ihnaini *et al*., 2021).

Potential Solution: To address mobile device limitations, researchers have explored model compression, quantisation, and the use of lightweight architectures like MobileNet (Ajani *et al*., 2021; Braun & Möller, 2024). Additionally, optimisation techniques are being developed to reduce the amount of processing and storage required to execute machine learning models on mobile devices (Wang *et al*., 2022). Other researchers have also focused on designing smaller models that consume fewer resources.

E. Data Imbalance and Sparsity

Challenge: Machine learning models in Mobile Big Data (MBD) analytics face major obstacles due to data imbalance and sparsity, which impact their accuracy and capacity for generalization. Imbalances, or the under- or unequal representation of some groups or classes relative to others, are frequently seen in mobile datasets. Because minority classes are less prevalent in the dataset, machine learning

algorithms may find it difficult to learn from them, which can result in biased model training and subpar performance when data distributions are out of balance (Liu *et al*., 2021). Machine learning algorithms may struggle to identify significant patterns and relationships in sparse data because they may lack the necessary knowledge to classify or predict instances with adequate accuracy (Chen *et al*., 2020). Specialized approaches and procedures designed for mobile contexts are needed to address the problems of data imbalance and sparsity in MBD analytics.

Application: Data imbalance and sparsity are highly pronounced in fraud detection, mobile healthcare, and social sentiment analysis, where minority classes often represent critical but underrepresented outcomes, making accurate detection both challenging and essential. For instance, anomalies, critical health incidents, or polarised opinions are most of the time underrepresented, leading to biased model training and poor generalisation.

Potential Solution: Solutions include oversampling techniques like the Synthetic Minority Oversampling Technique (SMOTE), undersampling, and ensemble learning for more balanced predictions. In sparse data settings, transfer learning and semi-supervised learning (Nguyen *et al*., 2019) are also promising, allowing models to learn effectively from incomplete or partially labelled data. Zhao *et al*. (2022) emphasised the importance of resampling techniques like oversampling and undersampling in reducing class imbalances by producing synthetic data or lowering the prevalence of majority classes to enable more balanced modelling. Addressing class imbalance in Mobile Big Data is a focus of several recent works (Liu *et al*., 2020; Zhao *et al*., 2022).

F. Interoperability and Integration from Heterogeneous Data Sources

Challenge: In the field of ML-based MBD analytics, interoperability and integration from different sources of data present substantial drawbacks. Mobile devices produce data, frequently in various formats and structures, from a variety of sources, including sensors, apps, and network connections. The challenges in training and deploying machine learning models arise from inconsistent data formats and semantics, which make it difficult to integrate and interoperate these heterogeneous data sources (Li *et al*., 2021). Thus, several technical issues must be resolved, such as feature extraction, data preprocessing, and model adaptation.

Application: This challenge directly impacts smart city platforms, IoT ecosystems, and mobile health systems, where sensor data, Global Positioning System (GPS) logs, application metadata, and user-generated content must be integrated across diverse formats and protocols to enable coherent modelling and decision-making. Mobile Big Data applications, such as smart cities and IoT ecosystems (Barik *et al*., 2018), often deal with heterogeneous data. Establishing smooth integration and interoperability among diverse data sources is also essential for creating efficient machine-learning models in MBD analytics.

Potential Solution: To manage this, researchers have applied data normalisation, ontology-based mapping, and cross-platform integration frameworks. Data harmonisation is essential to support accurate feature extraction and decision-making in such multi-source environments. To reconcile and resolve inconsistencies between different data formats, data preprocessing techniques like data cleaning, normalization, and transformation are crucial (Taleb *et al*., 2021). Lastly, model adaptation methods are required to guarantee that machine learning models trained on heterogeneous data sources will seamlessly combine and function within the mobile environment, taking into account variables like resource limitations and real-time requirements for processing.

G. Interpretability of Machine Learning Models

Challenge: Given the limitations and complexity of mobile contexts, the interpretability of machine learning models is a key challenge in the field of Mobile Big Data (MBD) analytics. Mobile devices frequently have resource constraints, such as limited memory and processing power (Somani *et al*., 2023). Therefore, it becomes difficult to deploy sophisticated machine learning models that prioritise interpretability because these models could need more computing power to produce explanations (Rajkomar *et al*., 2019). Moreover, it is crucial to strike a balance between model openness and data privacy since, if not properly planned and executed, highly interpretive models may unintentionally reveal sensitive information (Adadi & Berrada, 2018).

Application: The need for interpretability is most pressing in healthcare, finance, and anomaly detection, where stakeholders must understand the rationale behind predictions or classifications, particularly when the outcomes influence critical decisions such as treatment plans or financial approval. In regulated sectors like healthcare and finance, the deployment of interpretable models is becoming critical (Liu *et al*., 2020; Nguyen *et al*., 2019).

Potential Solution: Tools often called explainable AI (XAI) methods, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), are increasingly used to provide post-hoc explanations for complex models, while decision trees and rule-based systems offer intrinsic interpretability, which is especially valuable where transparency and trust are required.

H. Machine Learning-Based Challenges on Mobile Big Data

Challenge: Machine learning faces numerous challenges in Mobile Big Data. These challenges are about working with streaming data, high-dimensional data, model scalability, and distributed computing (Najafabadi *et al*., 2015). Given the so-called 4Vs of big data, which are velocity, volume, veracity, and variety, it is quite challenging to use traditional machine learning techniques. This makes it hard to manage and negatively affects the inference process at decision centres (Najafabadi *et al*., 2015). These challenges are further compounded by the unprecedented technological advancements, particularly in the context of Mobile Big Data and the AI revolution, which have led to a fresh paradigm and new learning landscape.

Application: The intersection between AI and big data in the context of Mobile Big Data analytics presents great potential in realizing highly effective teaching and learning. Furthermore, it also aids the stimulation of new research questions and designs, exploiting innovative technologies and tools in data collection and analysis.

Potential Solution: The challenges of handling velocity, variety, and volume in MBD have led to the use of distributed learning platforms. For instance, Apache Spark, TensorFlow Federated, and techniques like meta-learning and few-shot learning to address limited labelled data (Najafabadi *et al*., 2015). These innovations show promise in overcoming the traditional bottlenecks of Mobile Big Data analytics.

VI. FUTURE RESEARCH DIRECTIONS

As mobile devices become more commonplace, they generate exponentially more data, which presents opportunities and also challenges for successfully exploiting this data. This section presents the future research directions in Mobile Big Data analytics. Figure 4 provides a schematic overview of the major machine learning paradigms applied in MBD analytics, including supervised, unsupervised, and deep learning models, along with their respective strengths and typical use cases.

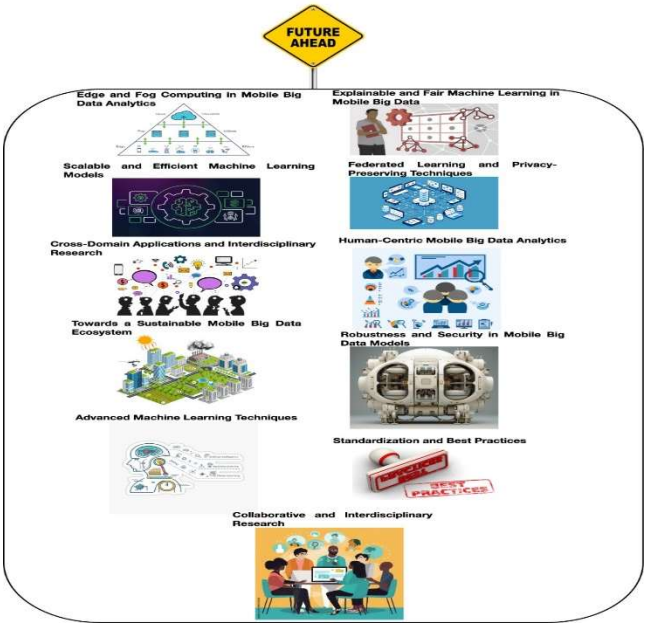

Figure 4: Future Research Directions

A. Edge and Fog Computing in Mobile Big Data Analytics

Recent observations have shown a steady increase in the overall ownership of globally connected portable devices, with estimated figures expected to be in the billions if not trillions by the year 2025 (Almolhis *et al*., 2020). Potentially, these devices are expected to produce significant amounts of data that could reach 500 zettabytes (Shi *et al*., 2016). Thus, traditional centralised architectures could potentially be limited in their ability to provide data storage, computation, and networking services. Furthermore, data processing on this type of architecture requires the transfer of data from a portable end-user device to a centralized server for processing. This, in turn, leads to a huge response time and may affect both the end-user's quality of service and experience. As a result, edge and fog computing have emerged, aiming to be the future of Mobile Big Data analytics and other IoT solutions (Laroui *et al*., 2021). Edge and fog computing offer the benefit of processing data at the network's edge, thereby reducing the load on the network and communication latency, energy consumption of mobile nodes, and increased reliability, privacy, and security. However, there remain several open areas for future consideration. For instance, several works have proposed ML and DL for Edge-IoT environments (Zafar *et al*., 2019). Future works may consider the use of distributed ML, like Deep Multi-Agent Reinforcement Learning to automatically make decisions in large-scale networks. Scalable and Efficient Machine Learning Models

ML models have shown impressive results in various fields, including renewable energy, autonomous driving, and medical diagnostics (Mulenga *et al*., 2023b; Phiri *et al*., 2023b). However, there is a substantial carbon footprint and the possibility of e-waste associated with the training and deployment of these models, which frequently need substantial computational resources and energy. Given these concerns about efficiency and sustainability, the ML community has started investigating methods for making ML models scalable and less complex (Aminizadeh *et al*., 2023). For instance, implementing algorithms that can automatically adjust compression levels in ML models to balance efficiency and performance (Srivastava, 2023). Also, considerations are being made towards the use of compression techniques in the context of federated learning. In this approach, multiple decentralized parties collaborate to train a shared model (Srivastava, 2023). Recent work, such as MobileBERT and TinyML, has focused on compressing deep learning models for mobile environments. In industry, Apple's Core ML and TensorFlow Lite are widely used to deploy efficient models on mobile devices.

B. Cross-Domain Applications and Interdisciplinary Research

The future research direction for Cross-Domain Applications and Interdisciplinary Research in Mobile Big Data Analytics aims to investigate innovative strategies for smoothly integrating diverse data sources from many areas, such as healthcare, finance, and logistics (Shahat Osman & Elragal, 2021). Researchers should concentrate on creating strong algorithms capable of handling the intricacies of heterogeneous data while maintaining interoperability and scalability (Sivarajah *et al*., 2017). Furthermore, multidisciplinary collaboration among data science experts, domain-specific professionals, and ethicists is critical to addressing privacy problems, ethical considerations, and regulatory compliance in Mobile Big Data analytics (Olayinka, 2022). The focus should be on developing novel approaches for real-time analytics, anomaly detection, and predictive modelling to extract meaningful insights from the massive and dynamic datasets generated by mobile devices, ultimately paving the way for transformative applications

and solutions with broad societal impact (Bansal *et al*., 2020).

C. Towards a Sustainable Mobile Big Data Ecosystem

Future research on a Mobile Big Data ecosystem that is sustainable is going to require a multifaceted strategy that addresses technological, economic, and environmental factors (Bibri & Krogstie, 2017). First, energy-efficient algorithms and infrastructure must be developed to reduce the carbon footprint associated with Mobile Big Data processing. Sustainability can also be enhanced by investigating renewable energy sources for data centres and maximizing resource use. It is imperative to do an economic analysis of business models that encourage ecologically conscious behaviour in the Mobile Big Data sector (Hassani *et al*., 2021). To foster user confidence in long-term Mobile Big Data ecosystems, research should also concentrate on improving data privacy and security protocols. Finally, by reducing latency, increasing efficiency, and permitting more decentralized data processing, emerging technologies like edge computing and 5G networks can be crucial in attaining sustainability (Calza *et al*., 2020).

D. Advanced Machine Learning Techniques

The need to address the growing complexity posed by the integration of blockchain, mobile data, 5G, and the upcoming 6G networks, in conjunction with IoT analytics, is poised to revolutionize the research landscape for Advanced ML Techniques in Mobile Big Data analytics (Atitallah *et al*., 2020). This development calls for investigating state-of-the-art techniques to strengthen data security, transparency, and trust, with an emphasis on using blockchain technology to support decentralized mobile environments. Furthermore, the combination of 5G and 6G capabilities promises to open up novel possibilities for real-time analytics on mobile devices, necessitating distributed structures and adaptable algorithms (Atitallah *et al*., 2020). Research studies will have to deal with issues of scalability, energy efficiency, and heterogeneous data stream management as the data ecosystem grows to include more and more IoT sources (Aminizadeh *et al*., 2023). At the same time, the rise of microservices in data analytics necessitates consideration, which drives the requirement for scalable, modular architectures to effectively implement machine learning models in distributed mobile and edge computing settings. To construct context-aware analytics solutions, transfer learning approaches will be crucial in utilizing pre-existing models and domain knowledge. Lastly, maintaining data integrity and quality within spatiotemporal dependencies will be crucial, necessitating the integration of spatiotemporal analytic methodologies along with robust preprocessing and cleaning techniques (Atitallah *et al*., 2020).

E. Explainable and Fair Machine Learning in Mobile Big Data

Explainable AI models significantly enhance the quality of decision-making and ensure that stakeholders assume accountability with confidence (Bhatt *et al*., 2020). Explainable ML models help provide improved system requirements for design, measurements, and user-performed periodic testing based on the domain and applications. These models promise to give precise, responsible, and methodical techniques of applying the fundamental values and concepts of the human decision-making process (Jagatheesaperumal *et al*., 2022). They also present excellent prospects for automated decision-making. To get the most out of using specific tools for automated and systematic decision-making, even before they are fully understood, research on creating strong Explainable ML paradigms and testing and deploying them across important use cases is highly anticipated (Jagatheesaperumal *et al*., 2022). Additionally, new strategies have emerged to develop energy-efficient IoT devices and edge devices. However, the widespread use of IoT might lead to energy waste (Bhatt *et al*., 2020). The problem remains challenging to overcome due to the scattered nature of IoT setups and flexible models. In other words, creating a low-complexity, highly flexible Explainable ML model is a crucial area of research for improving IoT system reliability while saving energy (Bhatt *et al*., 2020). Tools like SHAP, LIME, and Fairlearn are widely used in both research and industry to promote interpretable and fair ML models. Projects like IBM's AI Fairness 360 and Google's What-If Tool also support fairness and transparency evaluation.

F. Federated Learning and Privacy-Preserving Techniques

Federated learning is an ML technique in which a centralized server controls a group of clients that work together to train a model, while the training data is kept decentralized (Pham *et al*., 2021). Multiple datasets that are a component of local edge nodes are used to train ML algorithms, in particular, deep neural networks. For training purposes, the raw data from a traditional system is combined and stored in a centralized data centre. Nevertheless, in the case of federated learning, the raw data is left dispersed across the client devices and is subsequently aggregated from locally computed updates to create a shared model on the server (Pham *et al*., 2021). With federated learning in large data applications, there are several challenges. For instance, privacy is typically seen from a local or global perspective, taking into account all of the devices connected to the network (Rauniyar *et al*., 2023). Future research should focus on fine-tuning the definition of privacy because different devices have different privacy-related limitations or even different data points inside a single device. The goal is to provide privacy strategies that can manage a variety of privacy constraints. Additionally, the widespread use of FL in real-time settings is hampered by the incorporation of centralized training operations such as hyper-parameter tuning, neural architectures, debugging, and interpretability (Gadekallu *et al*., 2021). As a result, this presents a possible topic for further study. Google's implementation of federated learning in Gboard (its mobile keyboard) is a prominent real-world example. Additionally, the Flower framework and OpenMined project are supporting privacy-preserving FL at scale in research and enterprise environments.

G. Human-Centric Mobile Big Data Analytics

To build more inclusive, safe, and intuitive systems, future research in Human-Centric Mobile Big Data Analytics will potentially focus on improving the integration of cognitive computing, human-machine interaction, and cutting-edge technologies like 6G networks (Adel, 2022). These developments will put the security, privacy, and welfare of users first while utilizing blockchain technology for data management and sharing. Another important area of research will be the creation of safe and standardized procedures for obtaining personal information, especially in the fields of healthcare and education (Fatima, 2024). The objective is to develop intelligent surroundings in a more ethical, humane, and sustainable future where technology is used to empower and improve society as a whole.

H. Robustness and Security in Mobile Big Data Models

In the future, robustness and security research for Mobile Big Data models will concentrate on improving encryption methods for safe and effective data querying and storage, creating more useful systems for end users, and investigating novel approaches to privacy and security issues in mobile data analysis and sharing (Lo'ai & Saldamli, 2021). This will involve applying ML and behaviour detection methods to enhance malware detection and prevention, as well as integrating blockchain technology for enhanced security and privacy (Eke *et al*., 2024). Furthermore, the field will investigate new ML-powered signature techniques, new machine learning-based techniques for datasets with uneven network traffic, and new classification algorithms to identify malicious flows. To maintain integrity and privacy throughout the big data lifecycle, research should also take into account the difficulties associated with insecure data communication and storage, as well as the requirement for granular access control and data management (Rauniyar *et al*., 2023).

I. Standardization and Best Practices

The evolution of big data standards with additional research and the creation of new tools, services, and technologies are some possible avenues for future study in the areas of standardization and best practices in Mobile Big Data analytics (Ikegwu *et al*., 2022). To handle the difficulties of managing the growing volume and diversity of data, especially from non-database sources like cloud systems, web applications, video streaming, and smart devices, this evolution is probably going to make use of digital solutions (Chang *et al*., 2019). These solutions will also help to promote interoperability between legacy and modern systems. Furthermore, it is anticipated that ML, advanced analytics, and generative AI will all be used in new ways, leading to developments in edge computing and data processing (Aminizadeh *et al*., 2023). Additionally, given the massive amount of consumer data retrieved via mobile devices, which offers chances for more individualized and focused marketing, the influence of big data analytics on mobile marketing is anticipated to increase (Aminizadeh *et al*., 2023). To solve the difficulties and fully leverage analytics in Mobile Big Data for various fields, including business and supply chain management, the future directions recommended above will be essential.

J. Collaborative and Interdisciplinary Research

The investigation of cutting-edge ML and AI techniques customized for mobile platforms, the development of creative data visualization and interpretation techniques for Mobile Big Data, the study of privacy-preserving information sharing and analysis protocols for Mobile Big Data, and the integration of Mobile Big Data analytics with other up-and-coming tools such as edge computing and the Internet of Things (IoT) are some potential future research directions for collaborative and interdisciplinary research in Mobile Big Data analytics (Sarker *et al*., 2021b). The field would benefit from research into user-centered design for MBD applications, the ethical and regulatory elements of Mobile Big Data analytics, and the assessment of the impact of Mobile Big Data analytics on society and the environment (Chang *et al*., 2018). These research directions reflect the interdisciplinary nature of Mobile Big Data analytics, necessitating knowledge in domain-specific fields, including e-commerce, education, healthcare, and data science, in addition to computer science and mobile computing specialists.

While numerous promising research directions exist in the Mobile Big Data landscape, certain areas emerge as particularly urgent and impactful due to their relevance to real-time decision-making, privacy, and deployment feasibility. In their order of importance, these include:

i. Edge and fog computing (Section VI, Item A): Critical for low-latency, high-volume mobile applications like transportation, smart cities, and real-time monitoring.
ii. Scalable and efficient machine learning (Section VI, Item B): A key enabler for widespread adoption of ML in resource-constrained mobile and IoT devices.
iii. Explainable and fair ML (Section VI, Item E): Vital in regulated sectors requiring transparency and accountability, such as healthcare and education.
iv. Federated learning and privacy-preserving techniques (Section VI, Item F): Essential for ensuring data protection in health, finance, and human-centric applications.

These directions are not only technically relevant but also directly align with societal, legal, and ethical priorities in deploying trustworthy AI systems on mobile platforms.

In addition to their practical importance, these directions also respond directly to the key challenges identified earlier in Section V. For example, the scalability and real-time processing issues (Section V, Item A) are addressed through research into edge and fog computing (Section VI, Item A), while federated learning and privacy-preserving techniques (Section VI, Item F) directly respond to the privacy and security concerns highlighted in Section V, Item C. Similarly, efforts in interpretable and fair ML (Section VI, Item E) are targeted at overcoming challenges related to model transparency in Section V, Item E, and lightweight, efficient ML models in Section V, Item B to help mitigate resource constraints on mobile devices (Section V, Item F). These connections ensure that proposed research efforts are not just visionary but also grounded in the real-world constraints currently facing Mobile Big Data analytics.

VII. CONCLUSION

The amalgamation of machine learning with Mobile Big Data (MBD) offers hitherto unprecedented possibilities for deriving insights from the massive volumes of data produced by mobile devices, sensors, and services. This study has examined the exponential rise in data collection from several sources and the quick development of mobile devices, emphasizing the establishment of MBD as a data medium with vast analytical and insight potential. The growing significance of MBD in several sectors, including healthcare, retail, and transportation, as well as its capacity to spur innovation and decision-making, has been underlined in this study. To sum up, this study has given a thorough review of machine learning-based MBD analytics, including the most recent applications, taxonomy, difficulties, and potential future research areas. This research intends to improve understanding and offer insights for academics and practitioners in the field by analyzing the present landscape of applications and categorizing approaches using a precise taxonomy. To fully realize the potential of machine learning and MBD, future research should concentrate on resolving the issues that have been discovered, investigating novel approaches, and developing the integration of these two fields. In the end, machine learning-based MBD analytics may improve decision-making and create value in a variety of businesses by fostering growth and innovation across a wide range of areas through cooperative efforts and creative solutions.

Based on this survey of state-of-the-art research in Mobile Big Data analytics, several key insights emerge. First, while machine learning techniques have been widely applied across domains such as mobile health, finance, and transportation, real-time deployment and scalability remain persistent challenges. Second, privacy and security are critical concerns that are often acknowledged but insufficiently addressed in practical applications. Third, although emerging techniques like federated learning, edge computing, and explainable AI offer promising solutions, adoption remains limited, especially in resource-constrained environments. Lastly, there is a noticeable lack of standardisation and interoperability across platforms and data sources, which hinders effective cross-domain integration and scalability. These insights provide direction for future research and system development in the Mobile Big Data landscape.

Moving forward, researchers are encouraged to further investigate the specific technical and theoretical challenges identified in this review, including scalability, privacy, resource constraints, and interpretability. Meanwhile, practitioners and system developers are urged to explore and experiment with solutions such as edge computing, federated learning, and explainable AI, which have shown strong potential for practical application. Addressing these areas through collaborative, interdisciplinary efforts will be key to unlocking the full potential of Mobile Big Data analytics and ensuring its responsible, efficient, and equitable deployment across industries.

While this survey provides a broad and in-depth overview of machine learning applications in Mobile Big Data analytics, certain limitations should be acknowledged. The review primarily focused on peer-reviewed literature published in English, which may have excluded valuable insights from non-English or industry-exclusive sources. Additionally, due to space constraints, this work does not provide exhaustive technical analysis or architectural benchmarks for each ML approach discussed. Methodologically, the study does not employ a formal comparative framework (e.g., meta-analysis) or empirical validation of the reviewed techniques, which may limit direct generalizability. Moreover, due to the rapid pace of technological advancement, some recent developments may not have been captured, owing to publication lag. Future reviews could consider incorporating quantitative performance comparisons across domains and include grey literature or preprints for a more comprehensive landscape.

In conclusion, this review offers a holistic synthesis of ML-based Mobile Big Data analytics by mapping out its applications, identifying domain-specific challenges, and proposing actionable future directions. To bridge the gap between research and real-world adoption, it is imperative that researchers focus on developing scalable, privacy-preserving, and explainable models tailored for resource-constrained mobile environments. At the same time, practitioners and developers are encouraged to embrace interdisciplinary approaches, standardization efforts, and deployment-ready frameworks. These collective efforts are essential to unlock the full transformative potential of Mobile Big Data analytics in powering smart, ethical, and sustainable digital ecosystems.

AUTHOR CONTRIBUTIONS

C. I. Eke: Resources, formal analysis, Data Curation, **A.S. Al-Shamayleh**: Conceptualization, Methodology: **K. Maswadi**: Writing-Original draft preparation, Visualization, Validation, Investigation: **D.K. Kwaghtyo**: Data curation, Writing- Original draft preparation, Supervision, Project administration: **M. Mulenga**: Writing-Reviewing and Editing, Funding acquisition: **M. Phiri**: Resources, formal analysis, Data Curation, Conceptualization: **C.J. Iyidobi**: Methodology, Writing-Original draft preparation.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adel, A. (2022). Future of Industry 5.0 in Society: Human-Centric Solutions, Challenges and Prospective Research Areas. *Journal of Cloud Computing*, *11*(1), 40. https://doi.org/10.1186/s13677-022-00314-5

Agiwal, M., Roy, A., & Saxena, N. (2016). Next Generation 5G Wireless Networks: A Comprehensive Survey. *IEEE Communications Surveys & Tutorials*, *18*(3), 1617-1655. https://doi.org/10.1109/COMST.2016.2532458

Ahmad Jan, M., Adil, M., Brik, B., Harous, S., & Abbas, S. (2025). Making Sense of Big Data in Intelligent Transportation Systems: Current Trends, Challenges and Future Directions. *ACM Computing Surveys*, *57*(8), 1-43. https://doi.org/10.1145/3716371.

Ahmed, M., Mirza, M. A., Raza, S., Ahmad, H., Xu, F., Khan, W. U., Lin, Q., & Han, Z. (2023). Vehicular Communication Network Enabled CAV Data Offloading: A Review. *IEEE Transactions on Intelligent Transportation Systems*. https://doi.org/10.1109/TITS.2023.3263643

Ajani, T. S., Imoize, A. L., & Atayero, A. A. (2021). An Overview of Machine Learning within Embedded and Mobile Devices–Optimisations and Applications. *Sensors*, *21*(13), 4412. https://doi.org/10.3390/s21134412

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, *58*(2), 175-191. https://doi.org/10.1177/0047287517747753

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., Albahri, O. S., Alamoodi, A. H., Bai, J., & Salhi, A. (2023). A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare: Assessment of Quality, Bias Risk, and Data Fusion. *Information Fusion*, *96*, 156-191. https://doi.org/10.1016/j.inffus.2023.03.008

Alrizq, M., Solangi, S. A., Alghamdi, A., Nizamani, M. A., Memon, M. A., & Hamdi, M. (2022). An Architecture Supporting Intelligent Mobile Healthcare Using Human-Computer Interaction HCI Principles. *Comput. Syst. Sci. Eng.*, *40*(2), 557-569. https://doi.org/10.32604/csse.2022.018800

Alsheikh, M. A., Niyato, D., Lin, S., Tan, H.-P., & Han, Z. (2016). Mobile Big Data Analytics Using Deep Learning and Apache Spark. *IEEE Network*, *30*(3), 22-29. https://doi.org/10.1109/MNET.2016.7474340

Aminizadeh, S., Heidari, A., Toumaj, S., Darbandi, M., Navimipour, N. J., Rezaei, M., Talebi, S., Azad, P., & Unal, M. (2023). The Applications of Machine Learning Techniques in Medical Data Processing Based on Distributed Computing and the Internet of Things. *Computer Methods and Programmes in Biomedicine*, *241*, 107745. https://doi.org/10.1016/j.cmpb.2023.107745.

Aminzadeh, N., Sanaei, Z., & Ab Hamid, S. H. (2015). Mobile Storage Augmentation in Mobile Cloud Computing: Taxonomy, Approaches, and Open Issues. *Simulation Modelling Practice and Theory*, *50*, 96-108. https://doi.org/10.1016/j.simpat.2014.05.009.

Anagnostopoulos, I., Zeadally, S., & Exposito, E. (2016). Handling Big Data: Research Challenges and Future Directions. *The Journal of Supercomputing*, *72*(4), 1494-1516. https://doi.org/10.1007/s11227-016-1677-z

Anawar, M. R., Wang, S., Azam Zia, M., Jadoon, A. K., Akram, U., & Raza, S. (2018). Fog computing: An Overview of Big IoT Data Analytics. *Wireless Communications and Mobile Computing*, *2018*. https://doi.org/10.1155/2018/7157192.

Habeeb, R. A. A, Nasaruddin, F., Gani, A., Amanullah, M. A., Abaker Targio Hashem, I., Ahmed, E., & Imran, M. (2022). Clustering-based Real-time Anomaly Detection: A Breakthrough in Big Data Technologies. *Transactions on Emerging Telecommunications Technologies*, *33*(8), e3647. http://doi.org/10.1002/ett.3647

Atitallah, S. B., Driss, M., Boulila, W., & Ghézala, H. B. (2020). Leveraging Deep Learning and IoT Big Data Analytics to Support the Smart Cities Development: Review and Future Directions. *Computer Science Review*, *38*, 100303. http://doi.org/10.1016/j.cosrev.2020.100303

Bansal, M., Chana, I., & Clarke, S. (2020). A Survey on IOT Big Data: Current Status, Challenges, and Future Directions. *ACM Computing Surveys (CSUR)*, *53*(6), 1-59. http://doi.org/10.1145/3419634

Barik, R. K., Kandpal, M., Dubey, H., Kumar, V., & Das, H. (2018). Geocloud4GI: Cloud SDI Model for Geographical Indications Information Infrastructure Network. In *Cloud Computing for Geospatial Big Data Analytics: Intelligent Edge, Fog and Mist Computing* (pp. 215-224). Springer. https://doi.org/10.1007/978-3-030-03359-0_10

Bashabsheh, M. Q., Abualigah, L., & Alshinwan, M. (2022). Big Data Analysis using Hybrid Meta-Heuristic Optimisation Algorithm and MapReduce Framework. In *Integrating Meta-Heuristics and Machine Learning for Real-World Optimisation Problems* (pp. 181-223). Springer. https://doi.org/10.1007/978-3-030-99079-4_8

Bellini, P., Nesi, P., Paolucci, M., Soderi, M., & Zamperlin, P. (2019). Snap4city Platform: Semantic to Improve Location Based Services. *In Adjunct Proceedings of the 15th International Conference on Location-Based Services* (pp. 237-242). LBS2019. https://doi.org/10.34726/lbs2019.55

Benseny, J., Lahteenmaki, J., Toyli, J., & Hammainen, H. (2023). Urban wireless traffic evolution: The role of new devices and the effect of policy. Telecommunications Policy, 47(7), 1–44. https://doi.org/10.1016/j.telpol.2023.102595

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... & Eckersley, P. (2020, January). Explainable Machine Learning in Deployment. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 648-657). https://doi.org/10.1145/3351095.3375624

Bibri, S. E., & Krogstie, J. (2017). The Core Enabling Technologies of Big Data Analytics and Context-Aware Computing for Smart Sustainable Cities: A Review and Synthesis. *Journal of Big Data*, *4*, 1-50. https://doi.org/10.1186/s40537-017-0091-6

Braun, T., & Möller, R. (2024). Lessons from Resource-Aware Machine Learning for Healthcare: An Interview with Katharina Morik. *KI-Künstliche Intelligenz*, 1-6. http://doi.org/10.1007/s13218-024-00839-8

Brown, P. A., & Anderson, R. A. (2023). A Methodology for Preprocessing Structured Big Data in the Behavioural Sciences. *Behaviour Research Methods*, *55*(4), 1818-1838. https://doi.org/10.3758/s13428-022-01895-4

Buddha, H., Shuib, L., Idris, N., & Eke, C. I. (2024). Technology-Assisted Language Learning Systems: A Systematic Literature Review. *IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3366663

Cabrero-Holgueras, J., & Pastrana, S. (2021). Sok: Privacy-Preserving Computation Techniques for Deep

Learning. *Proceedings on Privacy Enhancing Technologies*. https://doi.org/10.2478/popets-2021-0064

Calza, F., Parmentola, A., & Tutore, I. (2020). Big Data and Natural Environment. How Does Different Data Support Different Green Strategies? *Sustainable Futures*, *2*, 100029. https://doi.org/10.1016/j.sftr.2020.100029

Chang, C., Hadachi, A., Srirama, S. N., & Min, M. (2019). Mobile Big Data: Foundations, State of the Art, and Future Directions. *Encyclopedia of Big Data Technologies*. https://doi.org/10.1007/978-3-319-63962-8_46-1

Chen, C., Zhang, M., Zhang, Y., Liu, Y., & Ma, S. (2020). Efficient Neural Matrix Factorization without Sampling for Recommendation. *ACM Transactions on Information Systems (TOIS)*, *38*(2), 1-28. https://doi.org/10.1145/3373807

Chen, F., Huang, L., Gao, Z., & Liwang, M. (2022). Latency-Sensitive Task Allocation for Fog-Based Vehicular Crowdsensing. *IEEE Systems Journal*, *17*(2), 1909-1917. https://doi.org/10.1109/JSYST.2022.3187830

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A Survey. *Mobile Networks and Applications*, *19*, 171-209. http://doi.org/10.1007/s11036-013-0489-0

Cheng, X., Fang, L., Yang, L., & Cui, S. (2017). Mobile Big Data: The Fuel for Data-Driven Wireless. *IEEE Internet of Things Journal*, *4*(5), 1489-1516. https://doi.org/10.1109/JIOT.2017.2714189

Chiu, M.-C., Huang, J.-H., Gupta, S., & Akman, G. (2021). Developing a Personalised Recommendation System in a Smart Product Service System based on Unsupervised Learning Model. *Computers in Industry*, *128*, 103421. https://doi.org/10.1016/j.compind.2021.103421

Chukwunweike, J., Anang, A. N., Adeniran, A. A., & Dike, J. (2024). Enhancing Manufacturing Efficiency and Quality through Automation and Deep Learning: Addressing Redundancy, Defects, Vibration Analysis, and Material Strength Optimisation Vol. 23. *World Journal of Advanced Research and Reviews. GSC Online Press*. https://doi.org/10.30574/wjarr.2024.23.3.2800

Cleuziou, G. (2024). Unsupervised Learning for Dynamic and Multi-View Data Analysis. Doctoral Dissertation, Université de Paris.

Diallo, R., Edalo, C., & Awe, O. O. (2024). Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. In *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA* (pp. 283-312). Springer. http://doi.org/10.1007/978-3-031-72215-8_12

Dimou, A., Iliopoulos, C., Polytidou, E., Dhurandher, S. K., Papadimitriou, G., & Nicopolitidis, P. (2022). A Comprehensive Review on Edge Computing: Focusing on Mobile Users. *Advances in Computing, Informatics, Networking and Cybersecurity: A Book Honoring Professor Mohammad S. Obaidat's Significant Scientific Contributions*, 121-152. http://doi.org/10.1007/978-3-030-87049-2_30

Ding, L., Li, C., Jin, D., & Ding, S. (2024). Survey of Spectral Clustering based on Graph Theory. *Pattern Recognition*, 110366. http://doi.org/10.1016/j.patcog.2024.110366.

Eke, C., Norman, A. A., Shuib, L., Long, Z. A. J. J. o. I. S., & Technologies, D. (2022). Random Forest-Based Classifier for Automatic Sarcasm Classification on Twitter Data using Multiple Features. *4*(2). https://doi.org/10.31436/jisdt.v4i2.345

Eke, C. I., MichaelP, E. B., & AydinP, M. Conceptual Framework for Context-Aware Service Discovery in Mobile Cloud. International Journal of Scientific Engineering and Applied Science (IJSEAS), 1(4). https://ijseas.com/volume1/v1i4/ijseas20150423.pdf

Eke, C. I., Norman, A. A., & Mulenga, M. J. A. I. R. (2023). Machine Learning Approach for Detecting and Combating Bring Your Own Device (BYOD) Security Threats and Attacks: A Systematic Mapping Review. *Artificial Intelligence Review*, *56*(8), 8815-8858. https://doi.org/10.21203/rs.3.rs-2124645/v1

Eke, C. I., Olamide, F. I., Phiri, M., Mwenge, M., Fatokun, F., & Shuib, L. (2024, November). Malware Detection Based on Stack Ensemble Using Machine Learning. In 2024 2nd International Conference on Computing and Data Analytics (ICCDA) (pp. 1-5). IEEE. https://doi.org/10.1109/ICCDA64887.2024.10867397.

El-Sayed, A., Abougabal, M., & Lazem, S. (2025). Practical Big Data Techniques for End-To-End Machine Learning Deployment: A Comprehensive Review. *Discover Data*, *3*(1), 11. https://doi.org/10.1007/s44248-025-00029-3.

Emmoh, P. U., Eke, C. I., Moses, T., & Ovre, A. (2025). Feature Selection Techniques for High-Dimensional Data Analysis: Applications, Challenges, and Future Directions. *Nigerian Journal of Technological Development*, *22*(1), 201-223. https://doi.org/10.63746/njtd.v22i1.2943

Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A Comprehensive Survey of Clustering Algorithms: State-Of-The-Art Machine Learning Applications, Taxonomy, Challenges, and Future Research Prospects. *Engineering Applications of Artificial Intelligence*, *110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743

Famoti, O., Ewim, C. P.-M., Eloho, O., Muyiwa-Ajayi, T. P., Ezechi, O. N., & Omokhoa, H. E. (2025). Revolutionizing Customer Experience Management through Data-Driven Strategies in Financial Services. *International Journal of Advanced Multidisciplinary Research and Studies. 2025a*, *5*(1), 948-957. https://doi.org/10.62225/2583049X.2025.5.1.3748

Fatima, S. (2024). Improving Healthcare Outcomes through Machine Learning: Applications and Challenges in Big Data Analytics. *International Journal of Advanced Research in Engineering Technology & Science*, *11*. https://www.researchgate.net/publication/386572106

Fettweis, G., & Alamouti, S. (2014). 5G: Personal Mobile Internet Beyond What Cellular did to Telephony. *IEEE Communications Magazine*, *52*(2), 140-145. https://doi.org/10.1109/MCOM.2014.6736754

Fullerton, E., Hellmold, S., Joshi, S., & McNamara, L. (2024). Rapid Expert Consultation on Archival Data Storage Technologies for the Intelligence Community. In Rapid Expert Consultation on Archival Data Storage Technologies for the Intelligence Community (pp. 1–27). https://doi.org/10.17226/27445

Gadekallu, T. R., Pham, Q.-V., Huynh-The, T., Bhattacharya, S., Maddikunta, P. K. R., & Liyanage, M. (2021). Federated Learning for Big Data: A Survey on Opportunities, Applications, and Future Directions. *arXiv preprint arXiv:2110.04160*. https://doi.org/10.48550/arXiv.2110.04160

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Springer. http://doi.org/10.1007/978-3-319-10247-4

Georgiadou, E., Angelopoulos, S., & Drake, H. (2020). Big Data Analytics and International Negotiations: Sentiment Analysis of Brexit Negotiating Outcomes. *International Journal of Information Management*, *51*, 102048. https://doi.org/10.1016/j.ijinfomgt.2019.102048

Ghosh, A. M., & Grolinger, K. (2020). Edge-cloud Computing for Internet of Things Data Analytics: Embedding Intelligence in the Edge with Deep Learning. *IEEE Transactions on Industrial Informatics*, *17*(3), 2191-2200. https://doi.org/10.1109/TII.2020.3008711

Gooderham, P. N., Elter, F., Pedersen, T., & Sandvik, A. M. (2022). The Digital Challenge for Multinational Mobile Network Operators. More Marginalization or Rejuvenation? *Journal of International Management*, *28*(4), 100946. https://doi.org/10.1016/j.intman.2022.100946

Grandvalet, Y., & Bengio, Y. (2004). Semi-supervised Learning by Entropy Minimization. *Advances in Neural Information Processing Systems*, *17*. https://dl.acm.org/doi/10.5555/2976040.2976107

Guo, J. (2022). Deep Learning Approach to Text Analysis for Human Emotion Detection from Big Data. *Journal of Intelligent Systems*, *31*(1), 113-126. https://doi.org/10.1515/jisys-2022-0001

Guo, Q., Zhuang, T., Li, Z., & He, S. (2021). Prediction of Reservoir Saturation Field in High Water Cut Stage by Bore-Ground Electromagnetic Method based on Machine Learning. *Journal of Petroleum Science and Engineering*, *204*, 108678. https://doi.org/10.1016/j.petrol.2021.108678

Guo, Y., Zhang, J., & Zhang, Y. (2016). An Algorithm for Analyzing the City Residents' Activity Information through Mobile Big Data Mining. In 2016 IEEE Trustcom/BigDataSE/ISPA (pp. 2133-2138). IEEE. https://doi.org/10.1109/TrustCom.2016.0328

Guruge, D. B., Kadel, R., & Halder, S. J. (2021). The State of the Art in Methodologies of Course Recommender Systems: A Review of Recent Research. *Data*, *6*(2), 18. https://doi.org/10.3390/data6020018.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues. *Information Systems*, *47*, 98-115. https://doi.org/10.1016/j.is.2014.07.006

Hassani, H., Huang, X., MacFeely, S., & Entezarian, M. R. (2021). Big Data and the United Nations Sustainable Development Goals (UN SDGs) at a Glance. *Big Data and Cognitive Computing*, *5*(3), 28. https://doi.org/10.3390/bdcc5030028

Hemmati, A., Raoufi, P., & Rahmani, A. M. (2024). Edge Artificial Intelligence for Big Data: A Systematic Review. *Neural Computing and Applications*, *36*(19), 11461-11494. http://doi.org/10.1007/s00521-024-09723-w.

Huynh, T., Nibali, A., & He, Z. (2022). Semi-supervised Learning for Medical Image Classification using Imbalanced Training Data. *Computer Methods and Programmes in Biomedicine*, *216*, 106628. https://doi.org/10.1016/j.cmpb.2022.106628.

Ihnaini, B., Khan, M. A., Khan, T. A., Abbas, S., Daoud, M. S., Ahmad, M., & Khan, M. A. (2021). A Smart Healthcare Recommendation System for Multidisciplinary Diabetes Patients with Data Fusion based on Deep Ensemble Learning. *Computational Intelligence and Neuroscience*, *2021*(1), 4243700. http://doi.org/10.1155/2021/4243700

Ikegwu, A. C., Nweke, H. F., Anikwe, C. V., Alo, U. R., & Okonkwo, O. R. (2022). Big Data Analytics for Data-Driven Industry: A Review of Data Sources, Tools, Challenges, Solutions, and Research Directions. *Cluster Computing*, *25*(5), 3343-3387. http://doi.org/10.1007/s10586-022-03568-5

Iorzua, J. T., Moses, T., Eke, C. I., Agushaka, O. J., Kwaghtyo, D. K., & Godswill, T. (2025). A Machine Learning Based Approach to Course and Career Recommendation System: A Systematic Literature Review. *Journal of Computing Theories and Applications*, *3*(1), 1-16. http://doi.org/10.62411/jcta.12603

Jagatheesaperumal, S. K., Pham, Q.-V., Ruby, R., Yang, Z., Xu, C., & Zhang, Z. (2022). Explainable AI over the Internet of Things (IoT): Overview, State-Of-The-Art And Future Directions. *IEEE Open Journal of the Communications Society*, *3*, 2106-2136. http://doi.org/10.1109/OJCOMS.2022.3215676

Jalil, Z., Abbasi, A., Javed, A. R., Badruddin Khan, M., Abul Hasanat, M. H., Malik, K. M., & Saudagar, A. K. J. (2022). COVID-19 Related Sentiment Analysis using State-Of-The-Art Machine Learning and Deep Learning Techniques. *Frontiers in Public Health*, *9*, 812735. https://doi.org/10.3389/fpubh.2021.812735

Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., & Abrahamsson, P. (2023). AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers. *IEEE Transactions on Computational Social Systems*, *10*(6), 2971-2984. https://doi.org/10.1109/TCSS.2023.3251729

Khan, N. U., Wan, W., Riaz, R., Jiang, S., & Wang, X. (2023). Prediction and Classification of User Activities using Machine Learning Models from Location-Based Social Network Data. *Applied Sciences*, *13*(6), 3517. http://doi.org/10.3390/app13063517

Kiran, M. S., Rajalakshmi, P., Bharadwaj, K., & Acharyya, A. (2014, March). Adaptive Rule Engine based IoT Enabled Remote Health Care Data Acquisition and

Smart Transmission System. In 2014 IEEE World Forum on Internet of Things (WF-IoT) (pp. 253-258). IEEE. http://doi.org/10.1109/WF-IoT.2014.6803168

Kolonia, A., & Forsberg, R. (2020). Preserving Security and Privacy: A WiFi Analyser Application based on Authentication and Tor. [Master's thesis, KTH Royal Institute of Technology]. https://www.diva-portal.org/smash/get/diva2:1469730/FULLTEXT01.pdf

Kučera, D. (2023). Challenges for Responsible Management Education During Digital Transformation. In *The Future of Responsible Management Education: University Leadership and the Digital Transformation Challenge* (pp. 35-60). Springer. http://doi.org/10.1007/978-3-031-15632-8_3

Kwaghtyo, D. K., & Eke, C. I. (2023). Smart Farming Prediction Models for Precision Agriculture: A Comprehensive Survey. *Artificial Intelligence Review*, *56*(6), 5729-5772. http://doi.org/10.1007/s10462-022-10266-6

Kwaghtyo, D. K., Eke, C. I., Abah, J., Moses, T., Agushaka, J., & Fatokun, F. B. (2024, January). Soil and Climate Parameters-based Crop Recommendation Model for Yandev, Gboko Local Government Area of Benue State. In 2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-7). IEEE. http://doi.org/10.1109/IMCOM60618.2024.10418425

Lam, J., & Abbas, R. (2020). Machine Learning-based Anomaly Detection for 5g Networks. *arXiv preprint arXiv*. https://arxiv.org/abs/2003.03474

Li, W., & Zhou, Z. (2015). Learning to Hash for Big Data: Current Status and Future Trends. *Chinese Science Bulletin*, *60*(5/6), 485-490. http://doi.org/10.1109/JPROC.2015.2487976

Li, Y., Ma, J., & Zhang, Y. (2021). Image Retrieval from Remote Sensing Big Data: A Survey. *Information Fusion*, *67*, 94-115. http://doi.org/10.1016/j.inffus.2020.10.008

Liu, C., Feng, Y., Lin, D., Wu, L., & Guo, M. (2020). IOT-based Laundry Services: An Application of Big Data Analytics, Intelligent Logistics Management, and Machine Learning Techniques. *International journal of production research*, *58*(17), 5113-5131. https://doi.org/10.1080/00207543.2019.1677961

Liu, H., Liu, Z., Jia, W., Zhang, D., & Tan, J. (2021). A Novel Imbalanced Data Classification Method based on Weakly Supervised Learning for Fault Diagnosis. *IEEE Transactions on Industrial Informatics*, *18*(3), 1583-1593. http://doi.org/10.1109/TII.2021.3084132

Liu, Y., & Wang, B. (2025). Advanced Applications in Chronic Disease Monitoring using IoT Mobile Sensing Device Data, Machine Learning Algorithms and Frame Theory: A Systematic Review. *Frontiers in Public Health*, *13*, 1510456. https://doi.org/10.3389/fpubh.2025.1510456

Lo'ai, A. T., & Saldamli, G. (2021). Reconsidering Big Data Security and Privacy in Cloud and Mobile Cloud Systems. *Journal of King Saud University-Computer and Information Sciences*, *33*(7), 810-819. https://doi.org/10.1016/j.jksuci.2019.05.007

Mahdi, M. N., Ahmad, A. R., Qassim, Q. S., Natiq, H., Subhi, M. A., & Mahmoud, M. (2021). From 5G to 6G Technology: Meets Energy, Internet-of-Things and Machine Learning: A Survey. *Applied Sciences*, *11*(17), 8117. https://doi.org/10.3390/app11178117

Malek, Y. N., Najib, M., Bakhouya, M., & Essaaidi, M. (2021). Multivariate Deep Learning Approach for Electric Vehicle Speed Forecasting. *Big Data Mining and Analytics*, *4*(1), 56-64. http://doi.org/10.26599/BDMA.2020.9020027

Mohamed, A., Najafabadi, M. K., Wah, Y. B., Zaman, E. A. K., & Maskat, R. (2020). The state-of-the-Art and Taxonomy of Big Data Analytics: View from New Big Data Framework. *Artificial Intelligence Review*, *53*, 989-1037. https://doi.org/10.1007/s10462-019-09685-9

Mphale, O., Gorejena, K. N., & Nojila, O. (2024). The Future of Things: A Comprehensive Overview of Internet of Things History, Definitions, Technologies, Architectures, Communication and beyond. *Journal of Information Systems and Informatics*, *6*(2), 1263-1286. http://doi.org/10.51519/journalisi.v6i2.738

Mulenga, M., Phiri, M., Simukonda, L., & Alaba, F. A. (2023). A Multistage Hybrid Deep Learning Model for Enhanced Solar Tracking. *IEEE Access*, *11*, 129449-129466. http://doi.org/10.1109/ACCESS.2023.3333895

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep Learning Applications and Challenges in Big Data Analytics. *Journal of Big Data*, *2*, 1-21. http://doi.org/10.1186/s40537-014-0007-7.

Navaz, A. N., Serhani, M. A., Al-Qirim, N., & Gergely, M. (2018). Towards an Efficient and Energy-Aware Mobile Big Health Data Architecture. *Computer Methods and Programmes in Biomedicine*, *166*, 137-154. https://doi.org/10.1016/j.cmpb.2018.10.008

Nguyen, D. C., Pathirana, P. N., Ding, M., & Seneviratne, A. (2019). Blockchain for Secure Ehrs Sharing of mobile cloud-based e-health systems. *IEEE Access*, *7*, 66792-66806. https://doi.org/10.1109/ACCESS.2019.2917555

Olayinka, O. H. (2022). Ethical Implications and Governance of AI Models in Business Analytics and Data Science Applications. *International Journal of Engineering Technology Research & Management*. 6(11). http://doi.org/10.5281/zenodo.15095979

Onah, J. O., Abdulhamid, S. i. M., Abdullahi, M., Hassan, I. H., & Al-Ghusham, A. (2021). Genetic Algorithm based Feature Selection and Naïve Bayes for Anomaly Detection in Fog Computing Environment. *Machine Learning with Applications*, *6*, 100156. http://doi.org/10.1016/j.mlwa.2021.100156

Oyeniyi, J. (2024). The Role of AI and Mobile Apps in Patient-Centric Healthcare Delivery. *World Journal of Advanced Research and Reviews*, *22*(1), 1897-1907. https://doi.org/10.30574/wjarr.2024.22.1.1331

Pham, Q.-V., Zeng, M., Ruby, R., Huynh-The, T., & Hwang, W.-J. (2021). UAV Communications for Sustainable Federated Learning. *IEEE Transactions on*

Vehicular Technology, *70*(4), 3944-3948. http://doi.org/10.1109/TVT.2021.3065084

Phiri, M., Mulenga, M., Zimba, A., & Eke, C. I. (2023). Deep Learning Techniques for Solar Tracking Systems: A Systematic Literature Review, Research Challenges, and Open Research Directions. *Solar Energy*, *262*, 111803. https://doi.org/10.1016/j.solener.2023.111803

Pinargote-Ortega, M., Bowen-Mendoza, L., Meza, J., & Ventura, S. (2024, January). Sentiment Analysis of Peer Feedback in Higher Education. In AIP Conference Proceedings (Vol. 2994, No. 1). AIP Publishing. http://doi.org/10.1063/5.0187956

Pouyanfar, S., Yang, Y., Chen, S.-C., Shyu, M.-L., & Iyengar, S. (2018). Multimedia Big Data Analytics: A Survey. *ACM Computing Surveys (CSUR)*, *51*(1), 1-34. https://doi.org/10.1145/3150226

Rafiq, F., Awan, M. J., Yasin, A., Nobanee, H., Zain, A. M., & Bahaj, S. A. (2022). Privacy Prevention of Big Data Applications: A Systematic Literature Review. *SAGE Open*, *12*(2), 21582440221096445. http://doi.org/10.1177/21582440221096445

Raja, G. B., & Chakraborty, C. (2023). Internet of Things based Effective Wearable Healthcare Monitoring System for Remote Areas. In *Implementation of Smart Healthcare Systems using AI, IoT, and Blockchain* (pp. 193-218). Elsevier. https://doi.org/10.1016/B978-0-323-91916-6.00004-7

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347-1358. http://doi.org/10.1056/NEJMra1814259

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J. M., & Herrera, F. (2016). Data Discretization: Taxonomy and Big Data Challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *6*(1), 5-21. http://doi.org/10.1002/widm.1173

Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2023). Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions. *IEEE Internet of things Journal*, *11*(5), 7374-7398. http://doi.org/10.48550/arXiv.2208.03392

Rayan, R. A., Tsagkaris, C., Zafar, I., Moysidis, D. V., & Papazoglou, A. S. (2022). Big Data Analytics for Health: A Comprehensive Review of Techniques and Applications. *Big data analytics for healthcare*, 83-92. http://doi.org/10.1016/B978-0-323-91907-4.00002-9.

Rekha, G., Tyagi, A. K., & Anuradha, N. (2019). Integration of Fog Computing and Internet of Things: An Useful Overview. In Proceedings of ICRIC 2019: Recent Innovations in Computing (pp. 91-102). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-030-29407-6_8

Ren, Y., Wang, T., Zhang, S., & Zhang, J. (2023). An Intelligent Big Data Collection Technology based on Micro Mobile Data Centres for Crowdsensing Vehicular Sensor Network. *Personal and Ubiquitous Computing*, *27*(3), 563-579. https://doi.org/10.1007/s00779-020-01440-0

Sarker, I. H. (2019). A Machine Learning-based Robust Prediction Model for Real-Life Mobile Phone Data. *Internet of Things*, *5*, 180-193. https://doi.org/10.1016/j.iot.2019.01.007

Shahat Osman, A. M., & Elragal, A. (2021). Smart Cities and Big Data Analytics: A Data-Driven Decision-Making Use Case. *Smart Cities*, *4*(1), 286-313. https://doi.org/10.3390/smartcities4010018

Shorfuzzaman, M., Hossain, M. S., Nazir, A., Muhammad, G., & Alamri, A. (2019). Harnessing the Power of Big Data Analytics in the Cloud to Support Learning Analytics in Mobile Learning Environment. *Computers in Human Behaviour*, *92*, 578-588. https://doi.org/10.1016/j.chb.2018.07.002

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical Analysis of Big Data Challenges and Analytical Methods. *Journal of Business Research*, *70*, 263-286. http://doi.org/10.1016/j.jbusres.2016.08.001

Somani, A., Horsch, A., & Prasad, D. K. (2023). *Interpretability in deep learning*. (vol. 2) Springer. http://doi.org/10.1007/978-3-031-20639-9

Sree Ranjani, R. (2021). Machine Learning Applications for a Real-Time Monitoring of Arrhythmia Patients Using IoT. *Internet of Things for Healthcare Technologies*, 93-107. http://doi.org/10.1007/978-981-15-4112-4_5

Stergiou, C., & Psannis, K. E. (2017). Recent Advances Delivered by Mobile Cloud Computing and Internet Of Things for Big Data Applications: A Survey. *International Journal of Network Management*, *27*(3), e1930. http://doi.org/10.1002/nem.1930

Su, Y., Agrawal, G., Woodring, J., Myers, K., Wendelberger, J., & Ahrens, J. (2014). Effective and Efficient Data Sampling using Bitmap Indices. *Cluster Computing*, *17*, 1081-1100. http://doi.org/10.1007/s10586-014-0360-5

Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (2021). Big Data Quality Framework: A Holistic Approach to Continuous Quality Management. *Journal of Big Data*, *8*(1), 76. http://doi.org/10.1186/s40537-021-00468-0

Theng, D., & Bhoyar, K. K. (2024). Feature Selection Techniques for Machine Learning: A Survey of More Than Two Decades of Research. *Knowledge and Information Systems*, *66*(3), 1575-1637. http://doi.org/10.1007/s10115-023-02010-5

Thillaigovindhan, S. K., Roslee, M., Mitani, S. M. I., Osman, A. F., & Ali, F. Z. (2024). A Comprehensive Survey on Machine Learning Methods for Handover Optimisation in 5g Networks. *Electronics*, *13*(16), 3223. http://doi.org/10.3390/electronics13163223.

Tyagi, A. K., & Sreenath, N. (2023). *Intelligent Transportation Systems: Theory and Practice*. Springer. http://doi.org/10.1007/978-981-19-7622-3

Vardakis, G., Hatzivasilis, G., Koutsaki, E., & Papadakis, N. (2024). Review of Smart-Home Security

using the Internet of Things. *Electronics*, *13*(16), 3343. http://doi.org/10.3390/electronics13163343.

Wang, Y., Kung, L., Wang, W. Y. C., & Cegielski, C. G. (2018). An Integrated Big Data Analytics-Enabled Transformation Model: Application to Health Care. *Information & Management*, *55*(1), 64-79. http://doi.org/10.1016/j.im.2017.04.001

Yazti, D. Z., & Krishnaswamy, S. (2014, July). Mobile Big Data Analytics: Research, Practice, and Opportunities. In 2014 IEEE 15th International Conference on Mobile Data Management (Vol. 1, pp. 1-2). IEEE. https://doi.org/10.1109/MDM.2014.73

Yuanxi, H., & Guang, Y. (2018, August). Study on the Impact of Internet Development on Talent Transformation. In 2018 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 1-4). IEEE. https://doi.org/10.23919/PICMET.2018.8481782.

Zhang, C., Si, Z., Ma, Z., Xi, X., & Yin, Y. (2016). Mining Sequential Update Summarization with Hierarchical Text Analysis. *Mobile Information Systems*, *2016*. https://doi.org/10.1155/2016/1340973

Zhang, C., Zhang, Y., Xu, W., Ma, Z., Leng, Y., & Guo, J. (2015). Mining Activation Force Defined Dependency Patterns for Relation Extraction. *Knowledge-Based Systems*, *86*, 278-287. https://doi.org/10.1016/j.knosys.2015.06.012

Zhang, K., & Chen, X.-W. (2014). Large-scale Deep Belief Nets with Mapreduce. *IEEE Access*, *2*, 395-403. https://doi.org/10.1109/ACCESS.2014.2319813

Zhang, Y., & Cheung, Y. M. (2014, December). Discretizing Numerical Attributes in Decision Tree for Big Data Analysis. In 2014 IEEE International Conference on Data Mining Workshop (pp. 1150-1157). IEEE. https://doi.org/10.1109/ICDMW.2014.103

Zhao, P., Luo, C., Qiao, B., Wang, L., Rajmohan, S., Lin, Q., & Zhang, D. (2022, July). T-SMOTE: Temporal-oriented Synthetic Minority Oversampling Technique for Imbalanced Time Series Classification. In IJCAI (pp. 2406-2412). http://doi.org/10.24963/ijcai.2022/331

Zhao, Q., Ye, F., & Wang, S. (2013). A New Back-Propagation Neural Network Algorithm for a Big Data Environment based on Punishing Characterized Active Learning Strategy. *International Journal of Knowledge and Systems Science (IJKSS)*, *4*(4), 32-45. http://dx.doi.org/10.1007/978-3-642-39637-3_33

Zhao, Y., Zhang, W., Zhou, L., & Cao, W. (2021). A Survey on Caching in Mobile Edge Computing. *Wireless Communications and Mobile Computing*, *2021*, 1-21. https://doi.org/10.1155/2021/5565648.