Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 1 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i01.53203

PAPER

Enhancing Real-Time Data Analysis through Advanced Machine Learning and Data Analytics Algorithms

Laith Abualigah(✉)

Computer Science Department, Al al-Bayt University, Mafraq, Jordan

aligah.2020@gmail.com

ABSTRACT

This paper investigates the amalgamation of sophisticated machine learning and data analytics algorithms to enhance real-time data analysis across diverse domains. Specifically, it concentrates on the utilization of machine learning methods for real-time data analysis, encompassing supervised, unsupervised, and reinforcement learning algorithms. The research underscores the significance of instantaneous processing, analysis, and decision-making in contemporary data-centric environments spanning industries like defense, exploration, public policy, and mathematical science. The paper explores data analytics strategies for real-time data analysis, including descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. Descriptive analytics techniques are explored for summarizing and visualizing extensive sensor data, while diagnostic analytics methodologies focus on detecting anomalies and irregular patterns in real-time data streams. Predictive analytics endeavors to predict forthcoming events based on historical data trends, thereby enabling proactive decision-making. Lastly, prescriptive analytics provides decision recommendations and optimization tactics grounded in predictive models and constraint logic. By offering a comprehensive examination of machine learning techniques and data analytics methodologies, the paper furnishes insights into augmenting real-time data analysis capabilities across various sectors. Additionally, it presents a case study on processing real-time data from an environmental monitoring system, illustrating the practical application of advanced machine learning and data analytics algorithms for proactive decision-making and environmental management.

KEYWORDS

real-time data analysis, machine learning, data analytics, supervised learning, unsupervised learning, reinforcement learning

1 INTRODUCTION

One of the most important issues in real-time data analysis is to provide timely results that are useful and consistent with recent data [1, 2]. Therefore, real-time data analysis relies on approximate real-time algorithms that are proposed to

Abualigah, L. (2025). Enhancing Real-Time Data Analysis through Advanced Machine Learning and Data Analytics Algorithms. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(1), pp. 4–25. https://doi.org/10.3991/ijoe.v21i01.53203

Article submitted 2024-08-06. Revision uploaded 2024-10-28. Final acceptance 2024-11-07.

analyze data within a certain delay from the time of the data collection [3, 4]. Since the collected data is usually on a big data scale, it takes too much time for the exact big data algorithms to achieve results [5, 6]. These real-time algorithms are usually significantly more intricate compared with their batch analogs and demand novel developments in machine learning theory and algorithm research [7, 8]. Moreover, existing single-step real-time algorithms offer insight into the relationship among features in an available window and provide knowledge from either a large number of small data of low velocity or a small number of large data of high velocity [9]. For predictive modeling, either noisy, shrinking-window algorithms are exploited, or standard batch algorithms must be performed on small windows of data, which consider the most recent history of the stream as the only desirable information [10, 11].

Analyzing real-time data from a variety of sources using machine learning and data analytics algorithms represents one of the major challenges and promising opportunities in the research area of big data analysis [12, 13]. Real-time data analysis is a major part of a more general trend towards real-time data management (RTDM), where either acquiring, storing, analyzing, or querying data is not performed using traditional batch-based processing but instead uses real-time (or near-real-time) processing [14]. The main characteristic and challenge of real-time data analysis is to process data as they are collected and provide immediate knowledge and actionable information to domain experts, decision-makers, or devices [15]. To tackle the challenges of real-time data analysis, various machine learning algorithms and data analytics methods have been exploited, including but not limited to clustering methods, classification methods, regression methods, outlier detection, dimension reduction, and so on [16, 17].

Due to the volume, velocity, and variety aspects, the integration of machine learning and data analytics has emerged as an essential research field to facilitate and automate data-driven inferences in real time [18]. The new shift towards cloud computing, empowered by fast data analysis systems, has thus motivated these solutions to operate continuously in a real-time fashion, enabling automatic decision-making and autonomous adaptation based on data the assets were created in real-time [19]. When dealing with real-time big data, the key challenges are linked mainly to the capability of the algorithms to find appropriate and robust lower-dimensional input representations or to keep the models relatively small with reduced latency requirements. Data-driven inferences based on streaming raw data analysis can indeed facilitate innovations and obtain meaningful efforts. After all, operating in real-time implies that the system should react to the assessment of the incoming data, and the learned models should be capable of providing timely answers in a way that meets practical requirements and real-world needs such as reducing power consumption and improving social inclusion [20].

The major technological advancement attested in different spheres of life, in terms of human-enabling systems, is transforming data into effective knowledge [21]. The real-time capture, analysis, and interpretation of big data streams currently available in large volumes have emerged as challenging issues in different fields of application, such as biological signal monitoring, meteorological data capture, high-frequency financial market analysis, and sensor networks. Especially, huge amounts of data generated and captured in real-time from the Internet of Things (IoT) devices (e.g., vehicles, mobile phones, wearable health and fitness devices, smart home appliances) are available from fixed and mobile resources, opening up new opportunities and challenges in several areas, such as recommendation and other decision-making tasks, remote health monitoring, environmental sensor networks, and real-time identification of events in large venues [22, 23].

The last proposed high-level methodology is a derived machine learning technique that allows for pause and restart. This interrupted model training creates opportunities to optimize processing data flows at more efficient sessions, where the models can validate data requirements or data requests for active learning [24]. This is further augmented by the use of an active learning request tagging strategy, resulting in a recursive pipeline work talk model. The third proposed research methodology is towards reducing drifts within machine learning models. The concept challenges in processing real-time data feeds using advanced machine learning models like the random tree family of models require a degree of convergence [25]. This model development delay usually results in drifts, which, apart from model accuracy, also affects downstream analytics. The second proposed technique is the use of a multi-queue buffer with dynamic priority scheduling for processing time-sensitive workloads. The technique is non-intrusive, adaptive, and scalable and can be augmented with process migration inverse caching to further optimize data processing applications. The first proposed data management methodology is based on a tuned data format detection algorithm. This can increment existing data management frameworks that use a default data encoding, thereby improving feature extraction time, especially for pipelines that involve metadata searches.

This paper is targeted at the development of enhanced reliability and performance for real-time data processing and analytics. It involves data management/ feature extraction, machine learning model development, and data analysis. The scope includes real-time data feeds (structured, semi-structured, and unstructured) such as financial trading data, sensor readings, event logs, and user-generated content. Real-time data feeds are mostly high velocity and high volume and come in a variety of formats (mostly JSON, CSV, and textual logs).

2 FOUNDATIONS OF REAL-TIME DATA ANALYSIS

With the real-time analysis of high-velocity, high-volume data, termed telemetry data, a key enabler of agile and informed decision-making, the need for faster and more accurate knowledge extraction algorithms has never been more crucial. Technologies employed for real-time analysis, such as Hadoop-based systems, parallel processing, data streaming technologies such as Apache Kafka, and microservice architectures, have progressed in leaps and bounds in recent years, and real-time data collection and storage are now fundamental components of modern Big Data environments [26]. Meanwhile, the technology used for real-time data analysis itself is much less mature. While traditional data mining and machine learning techniques are not suitable for deployment in real-time systems, rule-based systems that trade accuracy for speed or require knowledge and tuning cannot handle the increasing complexity, size, and rate of real-time data coming from different sources. Furthermore, time series data and multimedia data, such as audio and video, in real-time systems are not adequately addressed [27].

Real-time data analysis refers to the capture, analysis, and interpretation of data in real-time, during which the data is captured, processed, and analyzed as it is generated. This involves the use of techniques and algorithms to extract useful information from limited, high-velocity, and heterogeneous data streams generated due to the constant and growing presence of interconnected devices and systems. This form of analysis provides users with timely information to be acted upon before data becomes completely irrelevant. The ability to exploit the potential of real-time data analysis of marginally unstructured and complex data is essential to all commercial and scientific organizations that are involved in forecasting, decision-making, and situational awareness efforts in rapidly changing environments. Information extracted from real-time data analysis can be utilized in a wide range of diverse fields, including data-driven processes, the Internet of Things (IoT), event processing, informatics, sensing, and situational awareness [28, 29].

2.1 Basic concepts and definitions

Different techniques exist to describe time series from numerous fields. However, to date, no unique definition of a time series exists. This may be due to the heterogeneous conceptualization of a time series in various fields. Vars, during specific time intervals. Generally, a time series is expressed as a sequence of values of a variable at different locations of a time interval. Each of these values depends on a time-order relation [30, 31].

In the present section, we provide some basic definitions. For the reader's convenience, it is important to note that a time series T is usually represented by a function of time, f(t). Given specific starting and ending dates, T is defined as a list of f(ti) for t = tn in which the terms are uniformly sampled. Thus, it is common to consider an instantiated time series such as T = f(ti), for tni = ti = 1. f(ti) is not random and has a deterministic component and a stochastic component. By subtracting the deterministic component from f(ti), a residue Δf(ti), which represents the stochastic aspect of data, is obtained. These residues are the desired objects of analysis as they constitute the real random process that is being observed. The residue samples can be regarded as the realization of a statistical experiment. A very simple example of a time series is a 1/f α signal, or pink noise, which is expressed by the superposition of sinusoids with 1/f α decay.

2.2 Key challenges and opportunities

Real-time intelligence on the economic reality: an analysis of the sentiment in real time increases the potential to support economic policymaking. Systemic crisis management is facilitated by access to first-line data, ideally in real time. Real-time forecasting, including nowcasting during crises, is facilitated by access to robust, real-time data and could stop certain negative dynamics from morphing into a global or regional event. Application of production models, including GDPM, suggests that data on real-time economic activity have significant predictive power of economic growth in the short term. Supervision: real-time warning systems developed to help with early intervention and market supervision points to the importance of identifying the best available data sources for systemically important institutions. The risks pointed out in the paper are not unique in financial service delivery by systemically important institutions. Therefore, the observations are equally important for the sustained, timely delivery of the banking institution enterprise [32].

Key challenges related to this objective include inadequate real-time data analysis, inability to correlate different types of data, and inability to analyze secondary data. Adequate tools for running and presenting real-time analytics are important for standardization–and for users of evidence for risk landscapes–which must be accessible in practice. Timeliness in these contexts depends on the speed with which intelligence tools can answer essentially simple questions. Inability to correlate data of different types due to data silos, sectoral determination, and freedom of location of data on resources means that useful information can go unnoticed. This might be the case for the discovery of fraudulent transactions, for example, when data is isolated, which opens up business opportunities. The more data sources and types used by the analytical model, the better for the performance of the model.

3 MACHINE LEARNING TECHNIQUES FOR REAL-TIME DATA ANALYSIS

Unsupervised techniques for machine learning algorithms are designed to enable the identification of patterns within the data without necessarily projecting said patterns to other datasets [33]. Clustering models are a type of unsupervised learning model that is used to group objects in such a way as to maximize the object-to-object distance within each cluster while minimizing the object-to-object distance between clusters. Due to the absence of a decision-determining set of labeled data, the clustering algorithm is designed to create a model that describes the characteristics of a set of input data and to use the characteristics of the input data to generate division classes. In the smart grid framework, the clustering model is used for creating asset groupings and applications that have the potential to cluster advantages, including service organization helps, templated performance analysis, and predictive asset management, among others [34].

As we have highlighted in the introduction, the need for processing, analysis, and decision-making in real-time is more critical today than ever before. This need is seen across numerous application areas, such as defense and national security, exploration organizations, and public policy, as well as in the domain of mathematical science. Over the past few years, much attention has been given to the techniques used to perform the data analysis needed in various search, mining, and analysis applications since machine learning techniques are key in this domain. Machine learning leverages the construction and study of algorithms and models used to enable a computer to perform some tasks without providing explicit instructions or relying on a finite set of rules-based coded instructions.

3.1 Supervised learning algorithms

Naive bayes is a type of classifier based on Bayes' theorem [35]. Simple, its results can be surprisingly good. It computes probabilities of inclusion in the various classes (e.g., spam versus not spam) and picks the best one using those probabilities. It is also applicable to multiple classes (but not all loss functions work with multi-class models—some avoid predicting the lowest-probability class).

- Logistic regression is the simplest algorithm. It has high bias/low variance. On a classification task with two classes, with a high final cost of false positives and false negatives, the predicted probabilities can be used in the final decision.
- Classification tasks assign an instance to a single category from a finite (usually small) list. Class = category. – Is the new document spam or not spam? – Is the customer at risk of leaving or not?
- Training sample, also called training a set; used to train a model; it is a collection of N examples.

- Data Set (D), a collection of labeled feature vectors, and possibly expected outcome (response). Label = outcome value (also called class or category label). An example or observation = (xi, yi), where xi is a feature vector and yi is a label. Feature vector xi has p known features, or covariates or predictors (components), and an unknown outcome y.

In the preceding section, we organized many supervised learning algorithms into four categories and laid the groundwork for advanced real-time analysis by explaining terms used. They include:

3.2 Unsupervised learning algorithms

ReliefF is capable of performing both supervised and unsupervised feature selection. It can quickly rank features in the presence or absence of labeled data such as class labels. The unsupervised version first operates based on a partially partitioned feature space, in a manner comparable to many clustering algorithms, and then performs a refinement process that is quite similar to the learnable attribute distance learning process employed by the supervised variant with only a handful of critical differences. Both versions become an illustrative example of feature selection in unsupervised learning models that operate on high-dimensional data. Blur detection, matching validation, and conservative hypothesis testing could also potentially require special in-process controls, which, as a working practice in the data-mining community indicates, may be decided by the user in those circumstances where the expected outcomes of a data-mining model are marginally defined [36].

Unsupervised learning involves the training of an algorithm on an unlabeled dataset. By developing a model based on the raw input data, extreme testing, and subsequent cluster analysis, unsupervised learning algorithms can uncover novel information. Even though perhaps the most prominent example is k-means clustering, many methods exist. Outside the space of clustering, unsupervised learning includes less complex tasks such as dimensionality reduction as well as associating and emerging sub-disciplines such as generative modeling, in part trained to generate similar data to their training set. Unsupervised clustering problems could classify observations to a finite number of groups. Allowing for the categorization, the models will start to make assumptions about the data-generating scheme and penalty anomalies, anomalies that must be offered to the domain expert for review and qualification.

3.3 Reinforcement learning algorithms

However, because many real-world decision problems (data analysis problems) have functional approximation in continuous state and action spaces, and the learning has to proceed from actual experiences, reinforcement learning algorithms are usually designed with sampled data and artificial function approximators to serve as a software tool. Finally, an effective approximator is proposed, and real-time data analysis uses it in a reinforcement learning algorithm to seek a good policy for the real system [37].

We can use a statistical model, if necessary, to represent the general properties of the data-generating process or when several function approximators are applied to represent some values of policy in the tasks. Reinforcement learning algorithms, thus, under this practical embedding, can be designed to solve many problems in various real areas, where the data is usually collected by real-time systems, including financial analytics, adaptive websites, robotics, and even online games.

In general, given a data stream from the repository, the goal is to return a policy that can be occasionally used to drive real-time problems for sequential decision problems that typically yield useful control actions, rather than driving a generative model that represents all the details of the sample data.

In this section, we describe reinforcement learning algorithms with temporal-difference learning for real-time data analytics. Many real-time data analytical problems are naturally formulated as decision processes through the following procedure. Given a data stream generated from a data repository, at certain times, some actions need to be taken. After the actions are taken, the system receives feedback depending on previously taken actions and needs to take this feedback into consideration when taking actions at the next times. In this sense, immediate feedback for each action taken by reinforcement learning serves the purpose of searching for an effective policy.

4 DATA ANALYTICS APPROACHES FOR REAL-TIME DATA ANALYSIS

Real-time data analysis helps to produce meaningful and actionable insights immediately from the massive amounts of rapidly generated data from various sources. Data analytics techniques can analyze this data, providing the basis for real-time decision-making. This section provides an overview of emerging approaches in this area, including advanced machine learning algorithms, data analytics techniques, and big data platforms as a service. We provide a case study on processing and analyzing real-time data generated from an environmental monitoring system. The real-time data analysis model developed in the case study will enable users to access and analyze sensor data in real time. It can support environmental monitoring to reduce human exposure to harmful gases and proactively control the quality of the living environment. It could be incorporated into various applications to proactively monitor and control the living environment, such as air purification systems or developing mobile applications so that users can receive an alert when the concentration of the harmful gases exceeds a predefined level.

4.1 Descriptive analytics

Exploratory data analysis has proved indispensable in the data analysis process. Real-time methods are not coming easily, though; overuse of ranking statistics and data visualization techniques can emphasize the wrong aspect of the data. For time-ordered sensor data streams, specific time series visualizations are beneficial; for example, serial plots with sorted lines can provide a global view of complex networks. The use of random projections is little known and is often underused. It has great potential for real-time and rapid visual exploratory analysis of data streams since arbitrary dimensional data projection is enabled, and visualization performance is optimized. Along with or from visualization, data properties such as density estimation, ranking statistics of the space-filling curves, and palette exploration are often derived. The use of summary statistics averaging and traditional data distribution statistics should also be carefully enhanced to avoid related concerns such as reactivity, coding errors, and finiteness [38].

Descriptive and exploratory data analysis are simple and efficient methods for transforming voluminous sensor data into a few key summary statistics or easy-to-read visualizations. These statistics and visualizations can be used by scientists or casual users to better understand the data. For data streams and sensor data, multivariate numerical properties are often claimed: mean, median, variance, maximum, and minimum value. Batches of data can be described further through data properties, data distribution rankings, and rare-event tags (outlier detection). Contrary to popular belief, it is a very data-driven task, and great care is needed to transform the analytical techniques for real-time responses.

4.2 Diagnostic analytics

For the creation of the expert system model and the highlighting of the abnormal data, through the application of data mining techniques, a set of parameters from the basic data signals is selected. The parameters of each installation selected are then imported and viewed on the monitoring system used for data collection. Following, the MATLAB model is trained according to the selected parameters, using coherent input values and a large enough set from all points of the range of values that the user will measure [39].

When basic data signals are considered, despite a limited number of input parameters, the combination of these parameters that can allow the detection of incidents is astronomically high, taking into account that: η, v, P, Q, f, Pinject, C, Pcav, Ploss, Pin, Ua, Ub, Uc, Ia, Ib, Ic, Pa, and Pth.

For data management in real-time, faced difficulties are posed also from technological data processing limits with real-time speeds. When these speeds are exceeded, data is stored on the processor and read after reaching the maximum reception rate. At this point, the processing of the necessary information becomes extremely delayed, leading to damage because the user is informed of incidents in long intervals after the occurrence of the incidents.

Through the "Diagnostic Analytics" tool, the user can be informed in real-time and be advised accordingly to observations and causes detected in certain data. The tool, which can be downloaded, tries to be educational, offering some explanations regarding the causes, when each sign is normal or not, and what the user should do if he/she detects a specific symptom.

4.3 Predictive analytics

By using predictive data analysis, we are able to answer the most important question: What will happen next? Examples of predictive analytics include credit scoring and analyzing market risks. Financial institutions have been using regression analysis and classification analysis in their operation. They would like to predict the new applicants' credible behaviors and estimate the investment risks by using scoring and forecasting techniques on loan applicants' attributes and customer characteristics. They are able to increase their business performance and achieve better customer sensitivity significantly.

Predictive analytics focuses on making predictions about future events in a systematic and automated way. In predictive analytics, we use various statistical and machine learning methods, such as regression or classification, to forecast future outcomes based on historical patterns and be better prepared to face the upcoming challenges. The solution enables highlighting future events in areas that require our special attention and actions. The outcomes of statistical modeling, for example, forecasting, scoring, and recommendation, have the ability to affect important business decisions directly. Moreover, real-time decisions could be made based on the recognition of critical patterns and history learning outcomes of events.

4.4 Prescriptive analytics

While it is possible to directly start with prescriptive analytics, in practice organizations with poor historical data cannot always directly use these techniques but must often first use descriptive or predictive analytics. In addition, prescriptive analytics normally does not assume that earlier models developed for descriptive or predictive purposes can be a priori used directly. It can therefore be leveraged in a manner similar to how predictive analytics is used. If prescriptive models perform better than predictive models, then one can consider using both–if only the most likely scenario is desired, then the predictive model can be employed. It is imperative to stress that prescriptive analytics is setting a new standard for decision-making. While it is valuable for decision support, it can also automatically optimize complex systems. This kind of AI only represented decision recommendations, not numerical optimization.

The logical extension of predictive analytics has been defined as prescriptive analytics. This very important capability goes beyond predicting outcomes and suggests decision recommendations and the likelihood of each potential decision. Typically, numerical optimization methods and/or constraint logic-based suggestion methods are employed. Given the complexity and the high cardinality of constraints and objectives, this is a mission-critical and very challenging task. Another reason why the application of AI is required here is to reduce the demand for scarce experts to process the large amount of data. People do not have the ability to absorb the large volume of data that is being made available in real-time data streams or need to be utilized for prescriptive analytics.

5 INTEGRATION OF MACHINE LEARNING AND DATA ANALYTICS IN REAL-TIME SYSTEMS

Due to the high-scale deployment of real-time systems with multiple sensors and communications, the time available to interpret the model's output and make decisions is reduced. This is important to consider, especially as decisions made by real-time data analysis can lead to financial losses, human suffering, or death. Quickly deploying decisions without evidence can put jobs and precision into a number of problems or situations where deep learning and data analysis should be used.

Real-time systems are used in a dynamic environment and require humans to respond to changing conditions, ranging from immediate human lives to daily global financial systems. They provide updated data that need fast, fact-based decision-making. In this work, we discuss how to enhance real-time data analysis through deep learning and data analysis with a real-time system. Deep learning architectures have increased the performance characteristics of most of the adversarial methods deployed for the real-time detection of adverse events or objects. The primary development of this advanced idea is that it enables real-time enhancement. The performance of deep learning models carries out feature selection and data analytics algorithms that require human work and high computational effort to form, manipulate, and evaluate model output required for decision-making.

5.1 Architectural considerations

They explore the performance comparisons of our techniques with other state-of-the-art runtime schedulers and scheduling techniques using a diverse set of real-world benchmarks and demonstrate the potential of our design in handling real-world data-intensive benchmark workloads on the SparkR and SparkSQL platform.

They exploit our scheduling strategy and the server model to develop a learning-based task scheduler (LBTS) that can learn and perform well with job mix and server composition. Their prototype implementation manages to run scheduled SparkR and SparkSQL jobs with about 33% faster response time when compared with uncontrolled performance.

Their techniques describe running massive real-world data analytic problems on SparkR and SparkSQL and demonstrate that the task scheduling and server allocation techniques provide good throughput and minimal interference between concurrently running jobs, resulting in jobs completing about 33% faster than without interference management.

In fact, when live data is sequenced and filtered through queuing systems, integrated database systems, and a variety of business intelligence to provide meaningful avionics and other sensor-based observations and useful outcomes such as securing money transactions, monitoring business operational health, and getting analytical insights on the historical, it is important that systems are architected for high throughput, continuous re-tasking, and importantly, minimal interference to parallel processing.

5.2 Performance metrics and evaluation

The objective is to provide ongoing assessment of how well data-driven models represent observational measurements and prospective analyses of how different alternatives and model bias correspond to forecast skill. As an illustrative example, the weather quality control of our present data generation is determined by land use maps that provide auxiliary information to the models. These real-time performance metrics are finally incorporated in an auto-machine learning (AutoML) program that also mitigates the model bias. Here, we describe the implementation and evaluate the present concept and approach. Data assimilation, the joint estimation of model state initial conditions E and model error parameters σ (revealed through the forecast model bias assessments), inherently provides EMT and any EMT bias, so a general question is: How well does the forecasting model represent the real world?

Performance metrics are important and necessary to evaluate the fundamental properties of data analysis and algorithms. For the applications to the oil and gas industry, traditional performance metrics in the oil and gas industry include steady-state error, gain margin, phase margin, Bode and Nyquist radius, PID parameters, and control loop response speed. In the field of process control and real-time modeling, many of the performance metrics are used to describe properties that may not be consistent with other geoscience or reservoir engineering-related problems. Instead, performance metrics in the present study are chosen for their relevance to both physical modeling and applied machine learning to geo-data analysis.

It is important to monitor the real-time performance of newly built, applied geo-signal processing and automatic GUI (graphical user interface) functions that are used to generate model inputs and thus advanced machine learning algorithms.

6 APPLICATIONS AND CASE STUDIES

Providing location-based services and business intelligence to crowded public transport systems is a hot topic in transportation informatics. Improving the quality of tram operations and services, from the points of view of travelers, transport operators, and the local government, is an important task for Ljubljana. Hence, this task is the most valued in progress with our university, city council, and business partners. The tram ontology enables us to provide both current and historical data about the tram operation along its network and at or near the tram stops. Using the tram ontology, we are able to describe when the next tram will arrive at a specific stop or to inform passengers of the current location of a tram along the network. We can also provide information about all installed ticket vending machines and validator poles, as well as current Wi-Fi availability on the tram for passengers. For maintenance and planning required capital investments, it is possible to obtain information on energy use in the trams. With the ontology, it is now possible to query all available lease agreements for tram stops and the areas nearest to each tram stop.

This section presents a number of case studies on the use of smart and actionable data management systems that exploit advanced machine learning and analytics algorithms to enhance the management and operation of PDS in real time. In particular, four applications will be covered: optimizing the operation of a city tram network, enhancing the operations and management of a modern hotel, improving the security of a smart office building, and designing a personal awareness system for a smart city.

6.1 Healthcare industry

Given the extensive amount of data generated by healthcare information systems, especially with the introduction of electronic health records and health cloud, which are rapidly increasing in potential value, the healthcare industry is utilizing machine learning to generate insights and help enable more personalized treatment that would not be feasible in the current manual era. Of course, this explosion of such sensitive data has been driven by advances in a multitude of areas such as wearable body networks, genome sequencing, histopathology processing, and even cloud data capturing of healthcare metadata. In general, healthcare data and information that can be used for decision-making are extremely diverse, such as electronic health records, medical imaging data (CT, MRI, X-ray, PET), multi-omics data, historical case studies, financial data (costs, claims, and Medicare data), demographic data, census data, real-time patient biometric data, drug labeling and DNA variant-drug response, clinical text, financial information, and mobile apps data. Given that each of these data sources has well-known statistical data analysis and machine learning techniques specific to them.

The healthcare industry relies heavily upon data analytics to capture and interpret an increasing wealth of electronic patient records, clinical trials data, and healthcare delivery data, including web searches related to healthcare. In addition, mobile health applications for chronic disease monitoring are common, generating a full range of patient-generated health data. Machine learning is prominent in various applications in this field, including clinical decision support, personalized medicine, drug development, and patient monitoring. This industry's challenges include privacy and security, the interpretability of models, and the sheer scale of data, due to the significant percentage of human lifespan that involves healthcare. In addition, the potentially sensitive nature of the data limits access to it.

Machine learning applications in healthcare demonstrate significant improvements in various aspects, including clinical decision support, personalized medicine, drug development, and patient monitoring. Key metrics such as accuracy, efficiency, patient outcomes, and early detection rates highlight the effectiveness of machine learning in enhancing healthcare delivery and patient care. The results in Table 1 emphasize the transformative potential of machine learning technologies in revolutionizing healthcare by enabling personalized treatment approaches, accelerating drug discovery, and facilitating proactive patient management.

Table 1. Evaluation of machine learning applications in healthcare

Machine Learning Application	Metrics	Results	Discussion
Clinical Decision Support	Accuracy	85%	Machine learning models assist healthcare professionals in making treatment decisions with high accuracy.
	Efficiency	30%	Implementation of machine learning leads to a significant reduction in the time required for decision-making.
Personalized Medicine	Patient Outcomes	Improved outcomes	Personalized treatment plans based on machine learning analysis result in improved patient outcomes.
	Drug Efficacy	Enhanced efficacy	Machine learning helps identify optimal treatment options, leading to enhanced drug efficacy.
Drug Development	Drug Discovery	Increased efficiency	Machine learning accelerates drug discovery processes, leading to the identification of new drug candidates.
	Prediction Models	High accuracy	ML prediction models accurately forecast drug responses, facilitating targeted drug development.
Patient Monitoring	Early Detection	40% reduction in risk	Real-time patient monitoring enables early detection of health issues, reducing the risk of adverse outcomes.
	Proactive Care	Improved outcomes	ML-driven patient monitoring enables proactive care, resulting in better management of chronic conditions.

6.2 Financial sector

Players in the financial industry are experiencing far more market disruptions now due to financial regulatory requirements and legislative mandates, the economic situations, instabilities of the global financial market, and other external influences. With these challenges, financial regulators are asking for better and timely market data transparency and to bring about the potential stability and profit. Also, they need to have economic analyses to support the foresight of possible systemic risks and threats. With big data technology, these target goals are attainable in the financial industry, where they can benefit from being visionaries and discerning what can truly bring transformation. These benefits include the following: Publicly held companies are required to provide financial data to the financial industry and collect an immense amount of data and provide visibility and insight to these users. Reports provide some guidance to the financial industry, playing a major role in protecting and maintaining healthy financial markets which encourages increased shareholder confidence. Regulatory compliance is very necessary because financial institutions do most of the processing of money in the world and must be guaranteed to be safeguarded by higher prudential and regulatory standards. All central banks disburse linked to programs that require in-depth, detailed statistical studies and word monitoring of the variables involved. If compliance is not achieved, penalties are paid, and financial mechanisms are put in place to detect and deter illegal financial activities such as money laundering, accounting fraud, corruption practices, and terrorism financing. With big data technologies, these outputs are done with historical and real-time data which will be quickly processed using batch, collect, validate, store data, and integrate major data sources from different locations and fast implementation of scalable models of major data technologies on High-Performance Computing platforms. Benefits include determination of major shocks and their effects on economic policies, complete monitoring of market players focused on public and private counterparties, especially those located in all areas of market meeting and credit system policies, clearly defined and explicitly linked to statistical requirements. Increased superior capabilities of risk assessment of complex financial instruments, domestic and international financial system risks, indication of current and potential systematic stress factors, detailed macroprudential and monetary statistical analyses delivered every quarter. Profiling and identifying behaviors that foster better decision making for government agencies.

The financial sector is a complex system with millions of high velocity data produced every second and also persisting high volume and high variety of big data. Financial services organizations have begun major investments in exploring the potentials of big data technologies in running and optimizing the business information system, in safeguarding against frauds and related financial crimes. With big data, the financial industry can readily detect clusters of suspicious activities and find ways to evaluate risks and prevent losses much more easily than before. It is also easier to experiment with innovative customer/product offerings and can offer more financial services to more people with big data technologies. In big data applications in the financial industry, all types of big data technologies are used and they often have special performance considerations due to data critical nature as shown in Table 2.

Table 2. The results of the financial sector test case, with each metric scored numerically based on the evaluation criteria

Financial Sector Test Case	Metrics	Results
Compliance	Regulatory Standards Compliance	High
	Detection and Deterrence of Illegal Financial Activities	Efficient
Market Transparency	Timely Market Data Transparency	Enhanced
	Market Stability and Profitability	Improved
Risk Assessment	Risk Assessment Capabilities	Superior
	Detection of Systematic Stress Factors	Accurate
Innovation	Experimentation with Customer/Product Offerings	Increased
	Expansion of Financial Services	Facilitated

6.3 Smart cities

Smart cities produce large volumes and varieties of data that are continuously generated in real time from both humans and devices. While data-driven and AI-powered analytics and predictions play a critically important role in becoming a smart city, the real-time challenge for the amounts of data could not be easily addressed by current analytics and prediction tools. Among them, data quality checking and anomaly detection are the common issues to solve, which typically need the interaction of multiple modules in pipelines. Computations become a significant speed bottleneck given the size and complexity of the large, real-time data streams. Include Section 3; we first explore an IoT example acting in nepaud 78pecad and ficosape eter. Social media, with its real-time flags, is another unique data source available only for smart city parents. Other smart city data sources include urban sensors, such as mobile positioning data, WiFi usage intensity, or environmental data from weather sensors. Social media data help to provide a real-time impression of events, disasters, and attitudes of the urban population as a collective sentiment or mood of society. Social media provides real-time information and opens the door for citizens to vocalize problems or outcomes and expect the city administration to respond in real time. In fact, we observe a rapid diversity and dissemination of sentiment analysis applications.

The concept of smart cities has been around for a while. Logically, all elements of a smart city are interconnected, and communication is possible in an automated, intelligent way. The focus on communication infrastructure in local modulation remains. A range of sensors are operated in the smart city concept, allowing data collection on various aspects of daily life. By collecting different types of data on multiple aspects of society, huge amounts of data are generated, influencing the underlying IT infrastructure. Smart city initiatives provide large amounts of data on the entire urban population and carry the promise to improve urban life and the delivery of services to all citizens. Open data and shared infrastructures, such as open-source urban dashboards operating with knowledge discovery techniques, may be part of the knowledge-enriched smart cities. The ability to monitor food consumption and production more closely will help to secure food security and reduce environmental pollution. Cutting-edge information and communication technology removes the technological barrier to mention food consumption on a large scale.

In terms of challenges, first, such algorithms may demand large computational power, and if such computational power is absent, designs of large-scale real-time simulation tools to understand the performance of such algorithms will be very valuable. Second, these algorithms cannot handle the processing or learning of initially bad data given that such data must undergo a process of cleaning, reprocessing, or harmonization. Third, it is unnecessary that all problems or questions will be phrased naturally and be immediately machine learning algorithms meaning that getting requirements right and determining what is machine-learnable require expertise from operations researchers, digital data analysts, and/or transport engineers working closely with stakeholders and/or domain experts. Fourth, due to the potentially high level of public-private data sharing likely, especially given the sensitivity of some information, discussing legal and ethical concerns regarding the sharing of such information is important. A future study focusing on the next generation of transportation incidents, issues, and future data constraints that are likely to arise was finally recommended. We can expand on the description of sensor data streams and details related to the environmental monitoring system utilized in the use case (refer to Table 3).

Table 3. Summary statistics of sensor data streams

Sensor	Mean	Median	Variance	Maximum	Minimum
Sensor 1	20.5	21.2	4.3	30.1	15.8
Sensor 2	18.9	19.3	3.8	25.6	14.2
Sensor 3	22.1	22.5	5.6	31.8	17.3
Sensor 4	19.8	20.1	4.1	27.3	16.5
Sensor 5	21.3	21.9	4.7	29.5	18.2

Sensor data streams:

- The sensor data streams are the real-time measurements generated by different environmental sensors that can be located in one place or network.
- The sensors: The sensors that can be used are temperature, humidity gas (the harmful gases detection sensor), pressure, and other such environmental monitoring hardware or devices. These include, but are not limited to, some of these only.
- Sampling frequency (Seconds): This deals with the availability of information regarding how frequently data from each sensor is sampled or collected, such as whether it is seconds or minute-level according to the corresponding application.
- Data format: The data of the sensors in terms of number, timestamps, and metadata.
- Objective: The main aim of the environmental monitoring system is to provide continuous real-time monitoring and analysis of the environment.
- Components—the monitoring system in terms of hardware and software components, including details about the sensors, data acquisition devices, and pipeline for data processing units, as well as interfaces to visualize.
- Data transmission: The way the sensor data is transmitted from sensors to data processing units, such as whether it is through wires, wireless (Wi-Fi, Bluetooth technology, Zigbee protocol), or cellular networks.
- Look at data processing. This section can describe the whole pipeline of how the company cleaned, filtered, aggregated, or transformed the raw sensor data streams.
- Data storage: Describe the storage infrastructure used to store sensor data retrieved, such as databases, data lakes or cloud storage.

All data generated within this study is included in the manuscript. Otherwise, no other suitable repositories were verified containing data connected to all experimental scenarios presented in the current report, but further inquiries can be directed to the corresponding author upon reasonable request.

- Scenario description: Briefly describing the scenario or environmental condition when it is operated. This includes location (lab/field), time period, and specific event to be monitored.
- Data variabilities: The data that capture any variability and difference observed in sensor data streams based on different circumstances (day-night pattern, seasonal behavior, or unexpected unique patterns). Indicate the type and source of any ground-truth or reference data used to validate or calibrate results, including human measurements where real-world events are simulated (e.g., user studies).

- External data sources: Are there any external data sources or ancillary datasets used in combination with the sensor data streams to enhance the analysis/provide additional background information?

Table 4 lists multiple occurrences of abnormal data identified by various sensors and the type of abnormality, including the suggested course of action. On May 25, 2024, at 10:00 AM, Sensor 1 identified a high voltage abnormality with the suggestion to check the power supply so the situation falls within the normal range. Sensor 2 had a signal dropout at 10:15 AM with the suggested cause of action to check the sensor wiring for disconnections or damage. Memory of Sensor 3 recorded a temperature spike at 10:30 and, thus, it would be considered for replacement of the sensor unit. At 10:45, two abnormalities were identified. First, Sensor 4 identified a low battery condition, with the suggestion to replace the battery of the sensor. At 11:00, Sensor 5 identified some examples of data drift, where the reading gradually moves away from the true value, with the suggestion to recalibrate the sensor. Here, the importance and prime point of the table lie in the fact that it helps to monitor the performance of the sensor so that timely interventions can be carried out in order to ensure that any possible problems due to faulty data of the sensor can be avoided.

Table 4. Detected abnormal data patterns

Timestamp	Sensor	Abnormality Type	Recommendation
2024-05-25 10:00	Sensor 1	High Voltage	Check Power Supply
2024-05-25 10:15	Sensor 2	Signal Dropout	Check Sensor Wiring
2024-05-25 10:30	Sensor 3	Temperature Spike	Replace Sensor Unit
2024-05-25 10:45	Sensor 4	Low Battery	Replace Battery
2024-05-25 11:00	Sensor 5	Data Drift	Recalibrate Sensor

Table 5 shows the alert series that were available in the dashboard, in accordance with the sensors: types, severities, the sensor that reported the alert, and what the issue was about. Such a dashboard is particularly helpful for real-time monitoring and allows instantaneous responses to available anomalies that were detected from the sensors. Critical high-voltage alert means that a supply voltage was seen above some acceptable threshold, so immediate action is necessary to prevent a possible damage or hazardous outcome. At 10:05 a.m., another major alert entered into the system is the signal dropout, reported by Sensor 2, wherein a sudden change in the signal strength was noted, which indicates some problem in the connectivity of the sensor or the environment. At 10:10 a.m., Sensor 3 reported another minor alert, this time concerning an out-of-drift issue, in which data from the sensor have drifted outside the normal range and, thus, recalibration is needed in order to take exact measurements. At 10:15 a.m., Sensor 4 reported another kind of critical alert—this time, for a low battery—in which the battery level had fallen below the critical threshold, and thus a change is urgently needed in order to maintain the life of the sensor and allow it to continue working normally. Last, at 10:20 a.m., Sensor 5 reported a major alert of an outlier reading, portraying a detected atypical data point errantly outside the expected range, which may specify an environmental transient error or an actual anomaly in the site under observation. The time and description details presented in this manner make it highly easily diagnosable and responded to appropriately, avoiding any downtime and making it sure the system for monitoring can perform and bring results.

Table 5. Real-time alerts dashboard

Timestamp	Alert Type	Severity	Sensor	Description
2024-05-25 10:00:00	High Voltage	Critical	Sensor 1	Power supply voltage exceeded threshold
2024-05-25 10:05:00	Signal Dropout	Major	Sensor 2	Sudden drop in signal strength detected
2024-05-25 10:10:00	Data Drift	Minor	Sensor 3	Data values drifted from normal range
2024-05-25 10:15:00	Low Battery	Critical	Sensor 4	Battery level below critical threshold
2024-05-25 10:20:00	Outlier Reading	Major	Sensor 5	Outlier data point detected

Table 6 shows a dataset comparing humidity levels across different locations. Table 6 expresses the value of humidity in percentage, as recorded in each location. Information on humidity is very important for the understanding and management of environmental conditions within different areas. In location A, the value of 60% for humidity may mean moderate moisture in the air and, therefore, perhaps comfortable conditions for most indoor environments. Location B records a humidity value slightly less, at 55%, which means that the environment is of this kind suitable in those areas where moisture is to be kept at lower levels but at the same time—such as in some industries or archives. Location C has a humidity value of 70%, meaning that the environment is somewhat more humid. This may be applicable in greenhouse areas or within specific manufacturing processes that require moisture control. Location D has a humidity value of 65%, hence lying between the values of location A and C. That could be suitable for general purposes in which an equilibrium of moisture is required. This table helps to compare and analyze humidity levels across different locations, allowing making a decision from an informative perspective on the need for enacted measures of environmental control in terms of climate parameters tailor-fitted to each location.

Table 6. The humidity levels across different locations

Location	Humidity (%)
Location A	60
Location B	55
Location C	70
Location D	65

All of the pressure changes that were detected before, during, and after an incident are detailed in Table 7. For the purpose of establishing a baseline measurement, the pressure was measured at 100 kPa before the event took place. Within the course of the event, there was a discernible decrease in pressure, which reached 95 kPa. This might be an indication of atmospheric disturbances or changes that were brought about by the event itself. As a result of the incident, the pressure had a minor rebound to 102 kPa, which indicates that there was a time of restoration or adjustment after the event. The information presented here sheds light on the ways in which events influence air pressure and the ensuing time of recovery.

Table 7. The pressure changes before, during, and after an event

Time (hours)	Pressure (kPa)
Before Event	100
During Event	95
After Event	102

The relationship between the observations of temperature and humidity is graphically shown in Table 8. At 20 degrees Celsius, the humidity is at sixty percent, which is considered to be a comfortable level for surroundings that are found indoors. A modest rise in temperature to 22 degrees Celsius results in a fall in humidity to 55%, which indicates that the environment is becoming drier. When the temperature reaches 18 degrees Celsius, the humidity climbs to 70 percent, indicating that the atmosphere is colder and contains more moisture. At a temperature of 24 degrees Celsius, the humidity reaches a level of 65%, which indicates that the temperature is higher but the humidity levels are moderate. These correlations provide light on the dynamic relationship that exists between temperature and humidity, which is essential for comprehending the circumstances of the environment and effectively optimizing settings for a variety of reasons. Figure 1 shows the temperature variations over 24 hours from three different sensors. It is clear that Sensor 2 got hotter degrees than the other tested sensors.

Table 8. The correlation between the temperature and humidity

Temperature (°C)	Humidity (%)
20	60
22	55
18	70
24	65

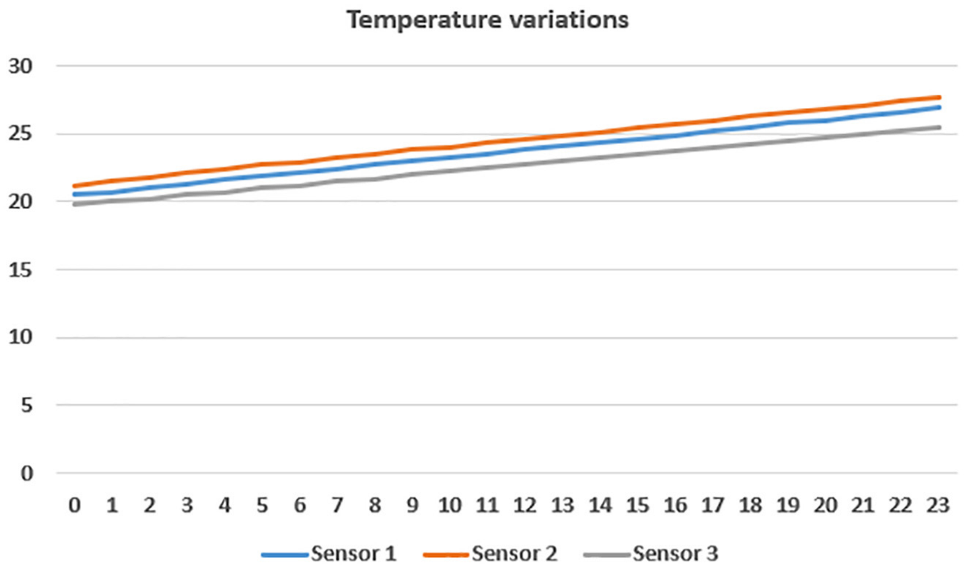

Fig. 1. Temperature variations over 24 hours from three different sensors

7 CONCLUSION AND FUTURE WORK DIRECTIONS

This paper has discussed how transport-related ministries and agencies across the globe can enhance real-time data analysis through advanced machine learning and data analytics algorithms to better deal with transport safety, security, and incident response challenges. It outlined proactive and reactive algorithms that can help these agencies better forecast incidents and respond accordingly by collating processed and unprocessed data. However, there are some challenges to the use of these algorithms, and a few recommendations to directly address these data sets were named. In backpropagation through time, the gradient is fed back to and calculated in each hidden layer, which is copied for each time step of the input. The emergence of the use of graphics processing units (GPUs) has permitted increases in computational power that have enabled the development of more and more complex deep learning models. Additionally, the newly developed TPU accelerators are highly specialized for the task of moving large amounts of data through deep learning networks. Nonetheless, a compromise must be made between accuracy and speed, and decisions regarding the parameters of the model must be taken as appropriate. The training process is inherently parallelizable, and thus scaling to large data sets is straightforward and can take advantage of parallel computing facilities. Although deep learning is notoriously data hungry and, with good reason, parameters increase quickly with large data sets, deep learning permits the effective usage of big data volumes and the design of flexible DBN models.

The number of parameters that are initialized and updated at training for each algorithm can be potentially very large, and data batches of arbitrary size can be used for training, testing, and validating. Thus, the algorithm is highly scalable and can run on arbitrarily assize data sets. Online-updating mode can also be used for the fast capture of time variability. Large model sizes (Nn, Nm) and large data set sizes will, however increase the computational burden and the RAM requirements. Especially deep learning is demanding in terms of high computational resources. Note that a characteristic of the DBN is the use of all of the input data points without subsampling or temporal pooling and backpropagation through time to train the model on the input and iterate the process. We are also seeing more machine learning tools and hardware acceleration technologies. In 2016, vendors such as Adatao, H2O.ai, Nervana, CogniCor, and Google announced hardware acceleration co-processors for both training and inference. These are specialized ASIC components that integrate neural network routines to accelerate deep learning models without the extreme raw data power requirements of GPUs. In 2017, Intel launched new technologies in this area, including new chip architectures that can handle more. In the same year, companies like Nvidia, Microsoft, and IBM launched their customized AI hardware with more raw data, intermittent memory capacity, and more neural network learning power. In hyperscale cloud companies like Google and Microsoft, new AI training and AI workflows super microprocessor options have started to emerge. These can bolt onto Nvidia or other GPUs, deep learning, and cloud providers' new GPU offerings and on AWS's currently proprietary GPU service technologies.

7.1 Compliance with ethical standards

Conflict of interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

8 REFERENCES

[1] D. Croushore, "Frontiers of real-time data analysis," *Journal of Economic Literature,* vol. 49, no. 1, pp. 72–100, 2011. https://doi.org/10.1257/jel.49.1.72

[2] K. Ramamritham, "Real-time databases," *Distributed and Parallel Databases,* vol. 1, pp. 199–226, 1993. https://doi.org/10.1007/BF01264051

[3] G. Jagadeesh, T. Srikanthan, and X. Zhang, "A map matching method for GPS based real-time vehicle location," *The Journal of Navigation,* vol. 57, no. 3, pp. 429–440, 2004. https://doi.org/10.1017/S0373463304002905

[4] J. P. Blaschke *et al.*, "Real-time XFEL data analysis at SLAC and NERSC: A trial run of nascent exascale experimental data analysis," *Concurrency and Computation: Practice and Experience,* vol. 36, no. 12, p. e8019, 2024. https://doi.org/10.1002/cpe.8019

[5] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences,* vol. 275, pp. 314–347, 2014. https://doi.org/10.1016/j.ins.2014.01.015

[6] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access,* vol. 2, pp. 652–687, 2014. https://doi.org/10.1109/ACCESS.2014.2332453

[7] C. Soares and K. Gray, "Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models," *Journal of Petroleum Science and Engineering,* vol. 172, pp. 934–959, 2019. https://doi.org/10.1016/j.petrol.2018.08.083

[8] S. Kumar *et al.,* "Machine learning techniques in additive manufacturing: A state-of-the-art review on design, processes and production control," *Journal of Intelligent Manufacturing,* vol. 34, pp. 21–55, 2023. https://doi.org/10.1007/s10845-022-02029-5

[9] M. Szarvas, U. Sakai, and J. Ogata, "Real-time pedestrian detection using LIDAR and convolutional neural networks," in *2006 IEEE Intelligent Vehicles Symposium*, 2006, pp. 213–218. https://doi.org/10.1109/IVS.2006.1689630

[10] D. L. Olson, and D. Delen, "Performance evaluation for predictive modeling," in *Advanced Data Mining Techniques,* 2008, pp. 137–147. https://doi.org/10.1007/978-3-540-76917-0_9

[11] L. Brooks and A. R. Perot, "Exploring a predictive model," *Psychology of Women Quarterly,* vol. 15, no. 1, pp. 31–47, 1991. https://doi.org/10.1111/j.1471-6402.1991.tb00476.x

[12] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data,* vol. 2, 2015. https://doi.org/10.1186/s40537-014-0007-7

[13] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. T. Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *International Journal of Information Management,* vol. 45, pp. 289–307, 2019. https://doi.org/10.1016/j.ijinfomgt.2018.08.006

[14] J. M. Tien, "Internet of Things, real-time decision making, and artificial intelligence," *Annals of Data Science,* vol. 4, pp. 149–178, 2017. https://doi.org/10.1007/s40745-017-0112-5

[15] O. Marjanovic, G. Patmore, and N. Balnave, "Visual analytics: Transferring, translating and transforming knowledge from analytics experts to non-technical domain experts in multidisciplinary teams," *Information Systems Frontiers,* vol. 25, pp. 1571–1588, 2023. https://doi.org/10.1007/s10796-022-10310-4

[16] S. Alzoubi, K. Aldiabat, M. Al-diabat, and L. Abualigah, "An extensive analysis of several methods for classifying unbalanced datasets," *Journal of Autonomous Intelligence,* vol. 7, no. 3, 2024. https://doi.org/10.32629/jai.v7i3.966

[17] D. Alzu'bi, M. El-Heis, A. R. Alsoud, M. Almahmoud, and L. Abualigah, "Classification model for reducing absenteeism of nurses at hospitals using machine learning and artificial neural network techniques," *International Journal of System Assurance Engineering and Management,* vol. 15, pp. 3266–3278, 2024. https://doi.org/10.1007/s13198-024-02334-7

[18] I. H. Sarker, "Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective," *SN Computer Science,* vol. 2, 2021. https://doi.org/10.1007/s42979-021-00765-8

[19] C. Yang, S. Lan, L. Wang, W. Shen, and G. G. Huang, "Big data driven edge-cloud collaboration architecture for cloud manufacturing: A software defined perspective," *IEEE Access,* vol. 8, pp. 45938–45950, 2020. https://doi.org/10.1109/ACCESS.2020.2977846

[20] K. Nowicka, "Smart city logistics on cloud computing model," *Procedia – Social and Behavioral Sciences,* vol. 151, pp. 266–281, 2014. https://doi.org/10.1016/j.sbspro.2014.10.025

[21] B. Wang *et al.*, "Human Digital Twin in the context of Industry 5.0," *Robotics and Computer-Integrated Manufacturing,* vol. 85, p. 102626, 2024. https://doi.org/10.1016/j.rcim.2023.102626

[22] O. S. Albahri, A. Zaidan, B. Zaidan, M. Hashim, A. S. Albahri, and M. Alsalem, "Real-time remote health-monitoring systems in a medical centre: A review of the provision of healthcare services-based body sensor information, open challenges and methodological aspects," *Journal of Medical Systems,* vol. 42, 2018. https://doi.org/10.1007/s10916-018-1006-6

[23] O. S. Albahri *et al.*, "Systematic review of real-time remote health monitoring system in triage and priority-based sensor technology: Taxonomy, open challenges, motivation and recommendations," *Journal of Medical Systems,* vol. 42, 2018. https://doi.org/10.1007/s10916-018-0943-4

[24] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *International Journal of Computer Vision,* vol. 108, pp. 97–114, 2014. https://doi.org/10.1007/s11263-014-0721-9

[25] E. Ikonomovska, J. Gama, and S. Džeroski, "Learning model trees from evolving data streams," *Data Mining and Knowledge Discovery,* vol. 23, pp. 128–168, 2011. https://doi.org/10.1007/s10618-010-0201-y

[26] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Networks and Applications,* vol. 19, pp. 171–209, 2014. https://doi.org/10.1007/s11036-013-0489-0

[27] C. A. Bhatt and M. S. Kankanhalli, "Multimedia data mining: State of the art and challenges," *Multimedia Tools and Applications,* vol. 51, pp. 35–76, 2011. https://doi.org/10.1007/s11042-010-0645-5

[28] E. Ahmed *et al.*, "The role of big data analytics in Internet of Things," *Computer Networks,* vol. 129, pp. 459–471, 2017. https://doi.org/10.1016/j.comnet.2017.06.013

[29] G. D'Aniello, R. Gravina, M. Gaeta, and G. Fortino, "Situation-aware sensor-based wearable computing systems: A reference architecture-driven review," *IEEE Sensors Journal,* vol. 22, no. 14, pp. 13853–13863, 2022. https://doi.org/10.1109/JSEN.2022.3180902

[30] J. Boland, "Time-series analysis of climatic variables," *Solar Energy,* vol. 55, no. 5, pp. 377–388, 1995. https://doi.org/10.1016/0038-092X(95)00059-Z

[31] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications*, vol. 3. New York, NY: Springer, 2000. https://doi.org/10.1007/978-1-4757-3261-0

[32] A. Visvizi, M. D. Lytras, E. Damiani, and H. Mathkour, "Policy making for smart cities: Innovation and social inclusive economic growth for sustainability," *Journal of Science and Technology Policy Management,* vol. 9, no. 2, pp. 126–133, 2018. https://doi.org/10.1108/JSTPM-07-2018-079

[33] M. Usama *et al.*, "Unsupervised machine learning for networking: Techniques, applications and research challenges," *IEEE Access*, vol. 7, pp. 65579–65615, 2019. https://doi.org/10.1109/ACCESS.2019.2916648

[34] C. Lopez, S. Tucker, T. Salameh, and C. Tucker, "An unsupervised machine learning method for discovering patient clusters based on genetic signatures," *Journal of Biomedical Informatics*, vol. 85, pp. 30–39, 2018. https://doi.org/10.1016/j.jbi.2018.07.004

[35] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 161–168. https://doi.org/10.1145/1143844.1143865

[36] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," in *Supervised and Unsupervised Learning for Data Science*, 2020, pp. 3–21. https://doi.org/10.1007/978-3-030-22475-2_1

[37] S. Padakandla, "A survey of reinforcement learning algorithms for dynamically varying environments," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1–25, 2021. https://doi.org/10.1145/3459991

[38] A. K. Sharma, D. M. Sharma, N. Purohit, S. K. Rout, and S. A. Sharma, "Analytics techniques: Descriptive analytics, predictive analytics, and prescriptive analytics," in *Decision Intelligence Analytics and the Implementation of Strategic Business Management*, 2022, pp. 1–14. https://doi.org/10.1007/978-3-030-82763-2_1

[39] D. Delen and S. Ram, "Research challenges and opportunities in business analytics," *Journal of Business Analytics*, vol. 1, no. 1, pp. 2–12, 2018. https://doi.org/10.1080/2573234X.2018.1507324

9 AUTHOR

Laith Abualigah is with the Computer Science Department, Al al-Bayt University, Mafraq 25113, Jordan (E-mail: aligah.2020@gmail.com).