



Paper Type: Original Article

## Big Data Mining Techniques In IoT, Challenges and Architectures

Mohammad Saber Iraji<sup>1,\*</sup>, Ali Yavari<sup>2</sup>

<sup>1</sup> Department of Computer Engineering and Information Technology, Payame Noor University, Tehran, Iran; iraji.ms@pnu.ac.ir.

<sup>2</sup> Department of Computer Engineering, National University of Skill, Tehran, Iran; ayavari@nus.ac.ir.

### Citation:

Received: 05 August 2024

Revised: 11 September 2024

Accepted: 16 February 2025

Iraji, M. S. & Yavari, A. (2025). Big data mining techniques in IoT, challenges, and architectures. *Transactions on soft computing*, 1(1), 46–60.

### Abstract


Nowadays, data is globally viewed as the most valuable resource, and the Internet of Things (IoT) has been playing an essential role ever since it emerged. Modern data sets are so complicated that they cannot be handled by traditional software and hardware. Three significant characteristics of the present time generated data are volume, velocity, and variety, which have resulted in the development of a concept called big data. Such characteristics have turned the routines of receiving, storing, processing, analyzing, and visualizing big data into a challenging issue. In the current competitive world, analyzing big data is critically important. The significance of big data doesn't refer to the amount of data that a company or organization accesses, but rather it depends on how the data is used. Processing and analyzing the collected data helps enterprises to gain the desired insights and benefit from them, compatible with strategic decisions. Over the past few years, some novel frameworks and tools have been presented for storing, processing, and analyzing big data, so that their relevant know-how and thus, working with such large-scale data can provide the specialists in this field with various research areas and job opportunities. This paper has addressed big data in the IoT, in which the issues about data mining architectures have been discussed. One of the prominent architectures raised in this field is the IoT-based multi-layered data mining model, which is divided into four layers: data collection, data management, event management, and data processing service. Another architecture considered in this paper is the distributed data mining model, whose primary goal is to pre-process the distributed data before being submitted to the central receiver (core infrastructure) in order to reduce energy consumption in the central nodes. Grid-based data mining infrastructure in the IoT, pursuing the objective to focus on the strategies to increase portability and situational awareness, is another model that has been dealt with in the references. Another architecture is the data mining model that predicts the integration of several technologies in the IoT. In this architecture, the integration of several technologies, including cloud computing, grid computing, pervasive networks, secure networks, etc., is dealt with. Another architecture under discussion is the IoT-based data mining on the cloud computing platform, whose objective is to integrate the layers of extraction, management, exploration, and interpretation. Finally, a three-dimensional five-layered architecture was proposed for mining big data in the IoT, the main idea of which is a three-dimensional device-layer architecture consisting of the device layer, raw data layer, data collection layer, data processing layer, and service layer. The other two dimensions cover issues like complying with the standards, security and privacy, data management, data perception, and data interpretation.

**Keywords:** Data mining, Big data, Internet of things, Data mining architecture.

## 1 | Introduction

Currently, considering the constantly growing use of mobile phones and electronic devices, accompanied by the potential of network connections, has revolutionized data generation. Every day, millions of mobile

 Corresponding Author: iraji.ms@pnu.ac.ir

 <https://doi.org/10.48314/tsc.vi.38>



Licensee System Analytics. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

devices are carried by humans, and the embedded internet-enabled sensors, such as speedometers, are producing and sending vast amounts of information. Mobile devices record humans' habits and living styles, and their health, personal, and social life-related details. Moreover, this companionship has tremendously contributed to the research activities of scientists. The rapid growth of mobile devices, accompanied by the augmentation of computing power, has provided the befitting opportunity for doing real-time and smart analytics applicable to vital applications such as health, smart transportation systems, smart homes, and crisis management [1]. The goal behind the Internet of Things (IoT) is to bring internet connectivity to the world of physical objects. The physical objects "Things" should possess a distinguished identity, be automatically identifiable, be able to communicate with humans, be equipped with the potential to make their own decisions, and follow human instructions.

Consequently, there is nothing wrong with claiming that in the IoT, the internet can be viewed as a global platform rendering the machines and smart things the power to communicate, compute, make decisions, and coordinate with humans across the world. The IoT is a technological revolution, and its purpose is to turn various things into smart objects. Despite the IoT being widely used in miscellaneous fields such as urban development, commercial, medical, etc., it suffers from some drawbacks. One of the challenges this technology is tackling is the problem with big data, that is, analyzing and extracting the required information from vast amounts of data has become a challenge. The IoT will vastly increase the data available for analysis by all kinds of enterprises [2]. Yet, there are salient obstacles that need to be removed before fully harvesting the potential benefits.

The IoT is a perpetually evolving constellation of internet-enabled sensors connected to various types of "Things". The sensors can potentially measure several processes, whether the internet connections are wired or wireless. In contrast, "Things" can literally be an object (Animate or inanimate) on which a sensor can be pasted or embedded. For example, if you carry a smartphone, you become a multi-purpose IoT, and most of your daily activities can be tracked, analyzed, and made effective [2].

Meanwhile, big data is described by four characteristics: Volume, Variety, Velocity, and Veracity. That is, big data appears in large quantities (Volume), is a combination of structured and unstructured information (Variety), arrives (Often in real time) (Velocity), and can be of unknown source (Authenticity).

The IoT and big data are vividly associated: Billions of "Things" connected to the Internet, which generate vast amounts of data, as the definition implies. However, this won't by itself lead up to another industrial revolution, transform the routine digital life, or provide an early planet-saving warning system anymore. As mentioned by EMC and IDC in their latest Digital World report, organizations must be rich in valuable data, that is, "The target information" should be: 1) easy to access, 2) have real-time availability feature, 3) leave a sizable footprint (Influencing a large portion of the organization or its customer base), and 4) able to create dramatic changes through proper analysis and follow-up measures.

More and more devices and objects are now connected to the internet, which are transmitting the data they collect for analysis. The goal here is to use such data to acquire more knowledge on the procedures and patterns that can be employed to exert positive impacts on lifestyle, energy conservation, transportation, and health. Nevertheless, the data itself does not produce such goals; instead, it creates their solutions through analysis and finding the responses we need. When the required data is received from an IoT-based device, an infrastructure has to be there to analyze the data [4]. This infrastructure can be built in or provided through an external system, but in the end, it provides this capability for companies and researchers to acquire some crucial data and convert the raw data into practical information. In the present paper, the IoT is addressed, through which we'd get the chance to investigate big data and its architecture [5].

## 1.1| Big Data and Internet of Things

The IoT technology is applied in most of the industries and businesses, which has led to boosting the quality of various data. Most IoT-based devices are equipped with the ability to collect various data from the environment, which includes smart speakers for listening to voice commands or uncrewed aircraft (Drones) in charge of collecting data under certain conditions. The collected data is then sent to the company servers, and businesses can use this data in numerous applications [6]. For instance, some companies might intend to use data for instant weather mapping and immediately delete the data. In contrast, some companies store such collected data in order to get it analyzed at the right time. Anyway, taking the big data into account becomes possible when you run into some practical cases for use. There is no single way to access IoT data, just as there is no unique approach to building an IoT-based device. Big data can help to extract some useful and valuable information from the data collected from millions of IoT devices instantaneously [7]. Big data analytics platforms receive the IoT devices-collected unstructured data (Ranging from traffic analytics to smart home information) and organize qualitative information to help companies optimize operations.

## 2| Big Data Extraction Architecture in Internet of Things

In this section, the architecture for extracting the IoT-based big data is presented. There are a number of architectures which have already been proposed from different perspectives of the IoT, for example, proposed a Blockchain architecture of things, which introduced a composite layer of the Blockchain between the network layer and the application layer, owning the advantages offered by a 5G-enabled network. Marjani et al. presented a big five-layer architecture of cloud-based IoT, the bottom-up layers including the IoT devices, network devices, gateway, cloud, and data analytics, investigating the power of IoT-based big data analytics in the IoT applications. Considering the above cases, in the big data mining architecture, we connect various applications such as AAL smart home, smart healthcare, smart traffic and parking systems, industrial IoT, and smart agriculture. The lowest layer of the system architecture consists of various sensing and actuating applications [8].

These devices include sensors, actuators, cameras, and small embedded systems for automated home, healthcare, traffic, parking, automobiles, industries, and agriculture, serving diverse applications. Different devices' generated raw data, like time series data and detected sequence of events, visual and audio data, etc., are collected by the Application Layer Gateway (ALG) and preprocessed for noise removal. Sequencing/recurrent events use different types of gateway processing units, i.e., Bluetooth, WiFi, or ZigBee routers with other electronic devices, including smartphones and small-scale embedded systems. Even the local server can be a gateway. Besides noise removal, the heterogeneous data of the IoT environment requires feature extraction, data fusion, and projection, which is operated by a gateway [9].

The abundance of IoT-generated data results in a new challenge known as the big data of the IoT. The inherent characteristics of this raw data are large volume, heterogeneity, production velocity, and rapidly changing data. The big data of the IoT possesses time as an integral element, i.e., unless it is processed in real-time or particularly within a short period, the processing result will be in vain after a particular deadline. Following the gateway, the preprocessed data is sent to the decentralized data centers via the internet. After that, the centralized processing and control stations extract knowledge via various data mining and Machine Learning (ML) algorithms at the respective ends.

Having extracted knowledge and analysis, the decentralized units provide services and perform the due smart operations in their constrained environment. Eventually, the extracted high-level proper IoT environment data are transformed into interpretable and understandable information for humans by the centralized processing and control station through coordinating the decentralized data processing and processing servers. If necessary, such data can be adapted to make reasonable decisions in order to optimize the performance and service quality of IoT applications and their infrastructure. In short, several global research panels are intensively surveying to come up with smarter knowledge techniques for extracting practical higher-level information out of IoT big data [10].

### 3|IoT-Based Data Stream Mining

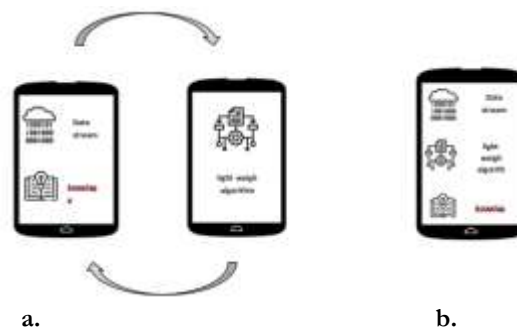
Regarding the nature of the IoT and the effective role of mobile devices in daily data generation in IoT, one of the data types that is generated, transmitted, and mined in large volumes is the data stream often generated on mobile platforms. Data stream is a subset of big data playing a key role in the development of many vast businesses, for example, analyzing mobile data streams by vehicles helps to enhance traffic management services. One of the primary principles for data stream mining in mobile devices is to build an integrated and extensible toolkit equipped with the potential to support mobile data mining applications [11] rapidly.

#### 3.1| Mobile Data Stream Mining Platforms

Mobile Data Stream Mining (MDSM) Platforms are implemented in several different topologies with special features. The fundamental communication models in MDSM platforms are made up of a variety of computing devices and middleware (middle edge) systems with various factors. These devices and systems are mobile devices, the internet, application servers, cloud data centers, and so on. This section presents the topological details of the above-mentioned platforms and how to implement the respective MDSM platforms [12].

##### 3.1.1| Topology of far-edge devices

Far-edge mobile equipment includes all systems and devices capable of wireless communication, data production, and processing. Smartphones, wearable sensors, wireless body networks, and smart vehicles in IoT are all examples of such devices, which have limited computing resources and low battery power. Therefore, such limitations have to be considered in their applications when it's time to run the processing cases. The MDSM application is implemented in the mobile environment, considering such limitations. The components of mobile data mining in this platform are depicted (*Fig. 1a*).



**Fig. 1. MDSM applications; a. in Far-edge mobile devices, b. in F2F interactions.**

##### 3.1.2| Far-edge to far-edge topology

F2F communications is based on sets of Far-edge devices capable of directly communicating with each other, for which they require no communication point control mechanisms. You can see an instance of this model in *Fig. 1.b*. This model is suitable for a person owning several devices, enabling point-to-point and group communication among different devices through Bluetooth, WiFi, infrared, and so forth, and via collaborative meetings; thus, implementing MDSM applications can be realized [13].

##### 3.1.3| Topology of Mobile Edge Computing (MEC) Servers

Any mobile device or any mobile system being placed at a single-hop distance from the Far-edge mobile devices is called a mobile edge server. A mobile edge server is a network architecture that enables data stream mining for thin and thick devices. Thin Far-edge mobile devices act as a source of data acquisition and transmission. Thick Far-edge devices are in charge of supplying further facilities for data stream mining algorithms. Of course, Far-edge systems are some instances of mobile edge servers frequently placed in the

same environment, such as personal mobile phones at home or colleagues' mobile phones at the workplace, in addition to some office equipment in an office building. The same location and simultaneous mobility of Far-edge devices and mobile edge servers minimize reliance on large-scale central systems.

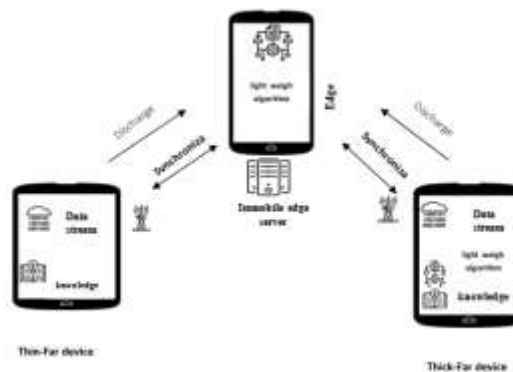


Fig. 2. MDSM applications in edge servers.

### 3.1.4 | Immobile edge servers

Physical static systems and the systems with full computing resources located at one-hop distance in wireless networks from Far-edge mobile devices are known as immobile edge servers, including smart cloud, microdata centers, Radio Access Network (RAN) servers, application servers, and smart routers in local networks. Fig. 2 displays a communication model, in which the immobile edge servers' built-in limitation forces Far-edge devices to cooperate with such servers [14].

### 3.1.5 | Mobile cloud computing

Mobile Cloud Computing (MCC) refers to mobile systems producing heterogeneous computing, networking, and storage services for Far-edge mobile devices through big data centers. The application models of data stream mining for cloud computing based on thin/thick Far-edge mobile devices are displayed in Fig. 3. For example, wearable mobile devices directly upload information to the cloud, and data stream mining operations are performed in the cloud.

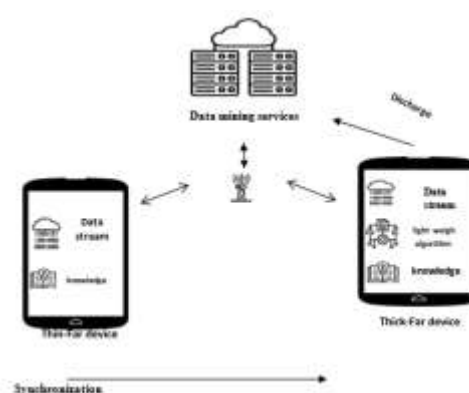


Fig. 3. Applications of MDSM in cloud systems.

### 3.1.6 | Mobile edge cloud computing

Multi-Access Edge Computing (MECC) is the extended network architecture of the traditional MCC at the edge of a cellular network at one-hop distance in wireless networks from mobile devices via mobile and immobile edge servers. MECC enables distributed MDSM applications. This capability takes place via

replicating the traditional infrastructure of cloud services in edge servers, such as multiple-level partitioning of applications.

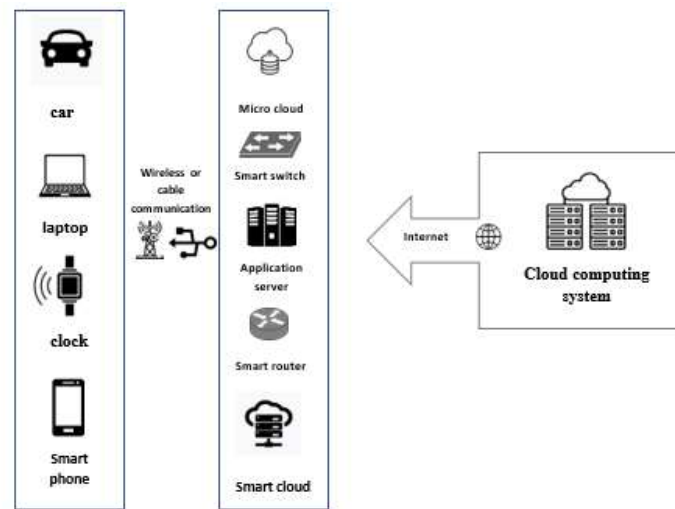


Fig. 4. MDSM applications in mobile edge cloud-based mobile systems.

## 4 | Mobile Data Stream Mining Application Platforms

This section briefly deals with some MDSM Application Platforms in IoT.

**Mine fleet:** A distributed data stream mining platform suitable for analyzing traffic and transportation workflows. This platform is responsible for online receiving and mining the vehicle-generated data stream.

**Context Aware Real-time Data Analytics Platform (CARDAP):** A distributed data stream mining platform for mobile sensing applications under dense and congested conditions. CARDAP is made up of three different strategies for transferring data from mobile devices to the cloud.

**Mobile Sensor Data Engine (MOSDEN):** Mobile sensor data processing engine as a plug-in-based IoT middleware for mobile devices, which allows random sensing for applications in highly congested environments. The component-based nature of this platform provides the potential for being generalized and programmed for any proposed architecture.

**Middleware Analysis and Retrieval System (MARS):** MARS prototype, a platform capable of processing the automatic information retrieved from the mobile phone embedded speedometer sensor and exploiting the statistical information and tagging data streams from specific physical activities such as walking, running, and standing.

**Star:** Stream learning through detecting the mobile operation framework. This platform addresses change detection through data streaming to categorize a specific operation.

**Point Distributed Model (PDM):** PDM is an agent-oriented distributed data stream mining system. PDM architecture is based on three general agents: mining agents, source discovery agents, and decision-making agents.

**CARA:** A cloud-based device for implementing cloud-oriented analyses. CARA works based on downloading a general learning model to detect the cloud environment operations.

**Service-Oriented Architecture (SOA):** A client-server computing model, sending the data mining service request from the mobile device to the cloud, and the cloud server provides a specific web-driven service.

**MobiSens:** A general sensor architecture to detect large-scale operations. This architecture is based on the client-server model, where the mobile device, client, and back-end server act as servers.



Mobile Waikato Environment for Knowledge Analysis (WEKA): A multipurpose tool for the mobile implementation of the WEKA data mining library.

MSM: A multipurpose tool for exploring data mining dependency rules from the sequences of repeated operations.

MobileMiner: A tool with the potential for mining the co-occurrence pattern via GPS data, call logs, and reporting applications.

Three-tier data mining architecture: Proposed by MDSM researchers, this architecture functions in three layers. Highly small smart devices with microcontrollers perform row-level learning based on samples and implement filtering methods.

## 5| Internet of Things-based Data Mining Challenges

- I. The first challenge is accessing and extracting data from different storage sources in different locations. The data retrieved from such sources are diverse, heterogeneous, and noisy.
- II. The second challenge is how to explore uncertain and faulty data in big data applications. The key issue in data mining systems is a safe and effective solution for data sharing.
- III. -Resource limitation and battery limitation in Far-edge devices, F2F communication models, and mobile edge servers are a fundamental bottleneck.

## 6| Key Data Mining Methods

The environment engulfing us is replete with heterogeneous data. It seems that when you lack the knowledge of how to use the data mining technologies properly, the extremely appealing environment becomes useless. Data mining can be a supervised, unsupervised, or reinforcement learning application. Computer-aided learning grows more precisely when it is done hierarchically in several layers. Hierarchically supervised or unsupervised learning-driven automated feature extraction is known as ML. Data mining is an integral part of knowledge discovery. The data collected from various IoT devices is first sent to a preprocessing unit, where several operations such as feature selection and extraction, noise separation, dimensionality reduction, and other tasks are carried out so that the raw data is placed in a suitable format for analysis. Then, the formatted data is sent to the data mining unit, where different data mining techniques perform their task to extract useful information at a higher level. The combination of preprocessed units and sufficient data is located in a single framework. The output of this evaluation unit is stated as machine-interpretable and human-understandable knowledge, which is further used by the IoT infrastructure.

## 7| Classification

Classification refers to the object assignment process to some predefined categories, whose goal is to predict the destination class for each data object accurately. Since the target labels are assumed to be known before processing, this is a supervised learning process. Classification requires training before being used to classify unlabeled or unknown objects/data. Therefore, the labeled or known data can be used to train the prediction function. For example, in a particular medical care center, the data of the patients suffering from a disease are available in three stages: the initial, intermediate, and critical treatment stages, respectively, or three specific methods for their treatment as Treatment-P, Treatment-M, and Treatment-S, respectively. First off, the classification/prediction function is built from a set of rules defined by a medical researcher or the recorded data during the due treatment. The existing data are divided into two classes:

A labeled training data set and an unlabeled training data set. The training set first builds and analyzes the classifier, and according to the detected stage of the disease, it gets to do class assignment, i.e., Treatment-P, Treatment-M, or Treatment-S.

Most algorithms are classified into two stages: the first one for estimating the probability of an item belonging to a particular class, and the second one for comparing and classifying based on the cut value. Evaluating the classification model's performance based on the number of items assigned to the correct category, that is, accuracy, and assigned to the incorrect category, i.e., error rate.

There are several classification models for data classification in different classes depending on the characteristics and conditions of the data, including the inferential decision-tree classification, Bayesian classification, rule-based classification, backpropagation classification, Support Vector Machine (SVM), K-nearest neighbor, deep neural network, and ensemble methods. Moreover, a set of classifiers can be applied to the complex problems of large-scale IoT applications by integrating different classification techniques.

## 7.1 | Clustering

A cluster stands for a group of similar objects. A clustering algorithm classifies the collected objects into a certain number of clusters; the objects of a particular cluster have similar characteristics. In contrast to classification, clustering is an unsupervised learning method. That means no previous segmentation is required for directing the process. For example, in a specialized medical care center where a number of patients suffer from an unknown disease, it has been recognized that the medical researchers only possess the information about the observed symptoms of the disease and the progression of the patient through following up on a large number of treatments. Under such circumstances, clustering is available by classifying the patients into a number of groups for the proper treatment based on the recognized symptoms and past treatment data.

## 7.2 | Association Analysis or Recurrent Pattern Extraction

A data type object, or a set of data objects, or a sequence of events repeatedly appearing in a system, is called a recurrent pattern. Extracting such recurrent patterns gives a good analytical understanding of user activity in a suitable environment. Finding affinity rules has comprehensive applications in the market basket, such as the situations where it is possible to predict a customer's buying pattern after analysis, which can enhance the business or user experience. In recurrent pattern extraction, the recognized events are essential in terms of order. The patterns observed in a specific order are sequential, and their extraction is called sequential pattern extraction.

The technological revolution enables extracting recurrent patterns in such environments as medical care centers or smart homes to assist us in diagnosing diseases in their embryonic stages. For instance, no disease emerges abruptly in the human body; instead, it goes through a step-by-step progressive process. Daily routine monitoring sensors can be employed to detect the symptoms of diseases. Recurrent pattern extraction can extract functional patterns/sequential information from daily operations or routine checkup data.

## 7.3 | Continuity Analysis

Continuity in data mining focuses on market basket analysis or data transaction analysis. It aims to discover the rules representing valuable relationships intermittently emerging, and it helps to generate general and qualitative knowledge, which in turn assists the decision-making process.

The initial list of continuity analysis algorithms includes those in which the order of data processing occurs. Algorithm-based prioritization has been used to discover the continuity of internal connections and then to locate the connections existing in many extended algorithms. Typical of the format of data records, there are two types of clusters: the horizontal database and vertical database format algorithms.

The prevalent algorithms include MSPS and LAPIN-SPAM. The growth algorithm is very complex, but it is able to calculate a vast quantity of data. The FP-Growth algorithm is the common one.

In some areas, the data are event data streams, and as a result, it gets tough to detect the pattern of events occurring recurrently and concurrently. This is divided into two parts: event-based algorithms and event-oriented algorithms, where the common one is the PROWL algorithm.



Such algorithms as Par-CSP have been extended in order to reap the benefits granted by the distributed memory parallel computer systems.

## 7.4 | Time Series Analysis

A time series is a collection of time-driven data objects having characteristics like large data sizes and constant updating. Typically, time series rely on three components, including representation, the degree of similarity, and indexing. One of the main reasons for representing time series is the reduction of dimensions, and it is divided into three categories: display-based model, inconsistent data display, and consistent data display, where the first one intends to discover the main parameters for presentation. The critical research cases include a bitmap study of ARMA modeling of time series. In the inconsistent data representations, the transition parameters remain the same for each time series, regardless of its nature, encompassing DFT, PAA, etc. In consistent data representation, the transition parameters change according to the data at hand, and their related tasks include the representation versions of DFT, PAA, and PLA indexing.

The similarity in time series analysis is usually measured approximately. The research direction includes sub-sequence matching and complete-sequence matching. The indexing of time series analysis is closely associated with representation and similarity measurement, and the relevant studies are SAMs and TS-Tree-based access for spatial databases.

## 7.5 | Other Analyses

Outlier detection refers to the detection of complicated data patterns that are significantly different from the rest of the data, constructed based on some appropriate criteria. Such a model often contains valuable information concerning the system's weird behavior description given by the data. Distance-based algorithms rank the distance between objects in irrelevant data as low-density models of data. Density-based algorithms estimate the density distribution of the input space, and after that, identify the irrelevant data as low-density models of data. The heterogeneous sets-based algorithms introduce heterogeneous sets or fuzzy heterogeneous sets to identify irrelevant out-of-range data.

# 8 | Internet of Things–Based Data Mining Architectures

The inherent characteristics of the IoT data, literally a subset of the macro data, such as their volume, variety, and speed, require data frameworks suitable for this data structure. This section presents some data mining architectures in the IoT.

## 8.1 | Multilayer Architecture of Internet of Things Data Mining

The multilayered data mining architecture is based on the IoT framework and data mining in RFID, the model which consists of four layers: data collection layer, data management layer, event processing layer, and data mining service layer.

**Data collection layer:** It collects information from the RFID readers and sinks, etc. This process tackles and solves several problems, such as harmful reading energy consumption, reading duplicate information, error tolerance, data filtering, and communication.

**Data management layer:** This layer uses a consolidated or distributed database or data warehouse for data management. After being detected, summarized, and compressed, various data in the corresponding databases are stored in the data warehouse. Later, after resolving the conflict, the conflict-free data can be stored in an appropriate format.

Then, a data warehouse called RFID-CUBOID can be hired to store and manage data. Based on RFID-CUBOID, users can perform online analysis, and XML language can also be used to describe the data in the IoT.

Event processing layer: Events, as a collection of data, time, and other factors, provide high-level mechanisms for data processing. This layer is designed to process data-related events. The primary observed events are filtered, and following that, the critical events of interest for the user are separated.

Data mining service layer: This layer is based on two lower layers, i.e., the event processing layer and the data management layer. In this layer, there are various event-based or goal-based services in data mining, namely classification, prediction, clustering, outlier data detection, dependency analysis, exploring applications for diverse Supply Chain Management (SCM), and so on.

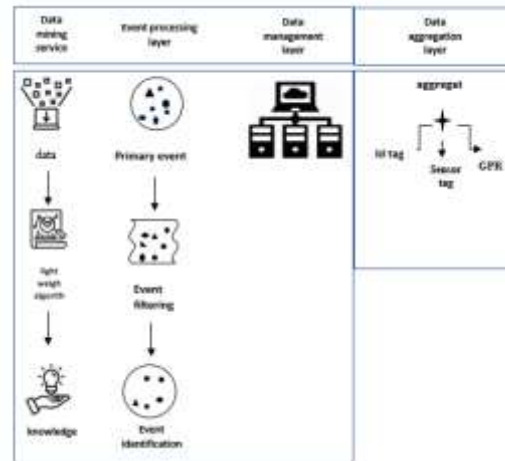


Fig. 5. Multilayered data mining model in internet of things.

## 8.2 | Distributed Architecture for Internet of Things -Based Data Mining

The IoT data possess special characteristics, for example, being different in terms of volume, distribution, and dependency on time and place. Moreover, the IOT data generation sources are heterogeneous. Such properties create lots of nuisances during data mining. Such vast amounts of data are stored in different sites. As a result, it requires mining during distribution exploration. Data preprocessing is a must for real-time processing.

Regarding some security considerations, legal restrictions, and other factors, it's impossible to comply with the strategy to stack the data next to each other. Besides that, in order to boost the central node's energy consumption, it's not optimal to transmit all data to a central node. Consequently, the distributed nodes' data must first be preprocessed, and then the required information should be transmitted.

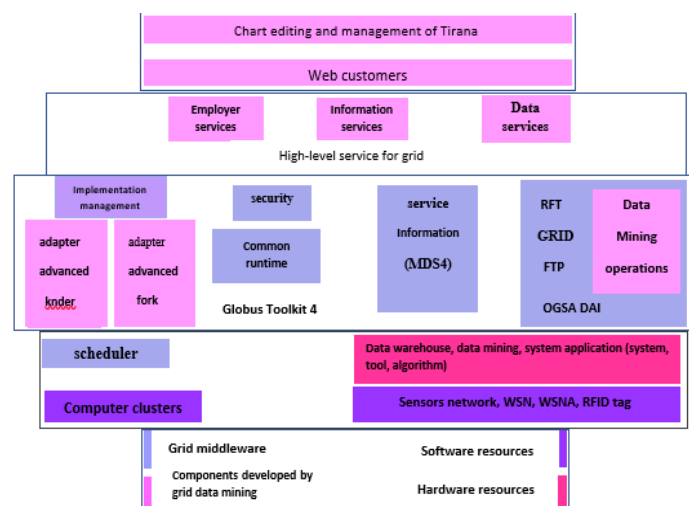


Fig. 6. Distributed model of internet of things data mining.

### 8.3 | Architecture for Grid Data Mining in the Internet of Things

Grid computing is a convenient option for a large, heterogeneous, and efficient infrastructure. Various computing resources and data sources are easily accessible and applicable. The main idea behind the IoT approach is the interconnected small objects over the Internet. Thus, smart objects are portable and reality-aware. Therefore, smart objects in the IoT are like grid computing resources. Thus, data mining services on grid computing can also be used for the IoT.



Fig. 7. Grid data mining in internet of things.

### 8.3 | Architecture for Internet of Things -Based Data Mining Using Multi-Technology Preaggregation

Today, plenty of technologies are being developed, which include cloud computing technology, grid computing, pervasive networks, secure networks, etc. In this architecture, the aggregation of several technologies has been taken into account. The data is gathered from location-aware sources, smart objects, and the environment. This architecture is compatible with 128-bit IPV6, in which various approaches realize the internet connection.

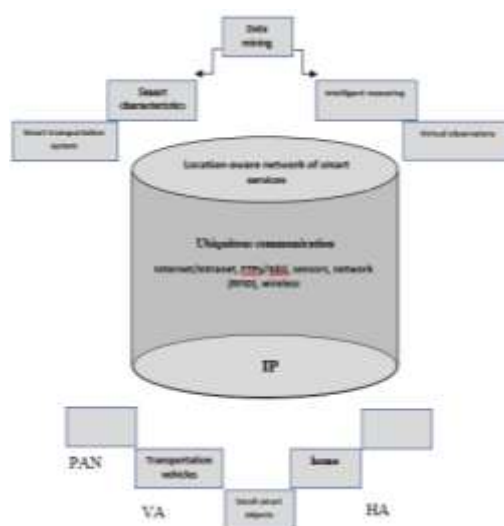


Fig. 8. Data mining model predicting multi-technology aggregation in the internet of things.

## 8.4 | Architecture for Internet of Things Data Mining Based on Cloud Computing

An architecture has been depicted in the Figure for supporting social networks and cloud computing in the IoT. In this architecture, big data and KDD have been aggregated into the extraction, management, and exploration layers, as well as the interpretation layer. The extraction layer is mapped into the receiving layer. Unlike the traditional KDDs, the extraction layer in this framework considers the agents' behavior with their things.

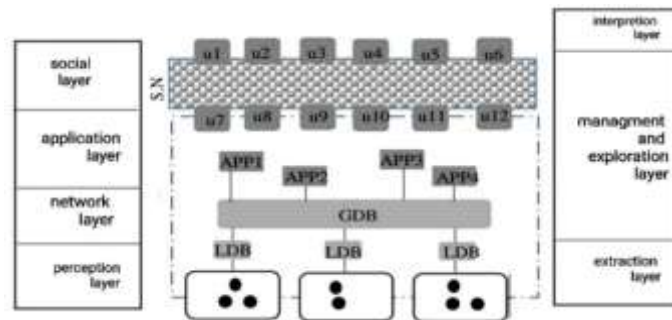


Fig. 9. Proposed architecture for big data mining in internet of things.

Regarding the presented models in previous sections, a three-dimensional five-layered architecture is presented for the IoT-based big data mining. In this architecture, besides considering the conventional data mining layers and the IoT infrastructure, new layers have been taken into account in two other dimensions for big data management. In the first dimension, there are five layers as follows:

**Device layer:** A large number of IoT devices, such as sensors, RFID, cameras, and other devices that can be gathered in this system for understanding the real world and continuously generate data, are located in this layer.

**Raw data layer:** In the big data mining system, various types of data, such as structured, unstructured, and semi-structured, are aggregated.

**Data collection layer:** Real-time data and batch data are supported, and all types of data can be analyzed and integrated in this layer.

**Data processing layer:** Lots of open source solutions like Hadoop, Storm, HDFS, and Oozie are collected in this layer.

**Service layer:** Data mining functions are provided as a service in this layer.

The other two dimensions point out the issues such as complying with the standards, security and privacy, data management, data perception, and data interpretation.

Table 1. Comparison of architectures.

Title	Function	Data Collection Tool	Advantages	Challenges
Multilayered architecture of IoT-based data mining	Consisting of four layers, including: Data collection layer, Data management layer, Event processing layer, Data mining service layer	Reader RFID, GPS...	Clear data mining	Energy consumption, bad reading, reading duplicate data, fault tolerance, data filtering, and communication
Distributed Architecture of Data Mining the IoT	Preprocessing of distributed data before sending to the network core to reduce energy consumption in central nodes	RFID, WSNs...	Reducing energy consumption in processing nodes	Bandwidth costs and Challenges related to distribution networks

Table 1. Continued.

Title	Function	Data Collection Tool	Advantages	Challenges
Grid data mining architecture in IoT	Focused on strategies to increase portability and situational awareness	Source grid, RFID, WSNs ...	Increasing the availability of resources	Overhead costs to increase availability and create different data distributions
Data mining model predicting the integration of several technologies in the IoT	Aggregation of several technologies: cloud computing technology, grid computing, ubiquitous networks, secure networks, and...	Internet, intranet, sensors, RFID, wireless, and...	Architecture IPV6	Challenges in collecting data from location-aware sources
Data mining architecture of cloud computing in the IoT	Integrating the extraction, management, exploration, and Interpretation layers	Internet, intranet, sensors, RFID, wireless, and...	Supporting social networks and cloud computing in the IoT	Security challenges in social networks
A 3-D five-layered architecture for big data mining in the IoT	A 3-D five-layered architecture consisting of a device layer, a raw data layer, a data collection layer, a data processing layer, and a service layer. In the other two dimensions, issues such as compliance with standards, security & privacy, data management, data perception & data interpretation	A large number of IoT devices, such as sensors, RFID, cameras, and other devices	Complying with standards, Security & privacy, data management, data perception & data interpretation	Overhead resulting from two-layered processing added to more traditional models

## 9 | Conclusion

The current paper has outlined the study of data mining in the IoT, where the issues about data mining architectures have been discussed. Also, since a considerable portion of the generated data comes from various data-flow platforms and communication topologies in the IoT, different data stream mining scenarios and diverse MDSM platforms have been investigated in this respect, and a novel model has been proposed for big data mining architecture in the IoT. Finally, the challenges that MDSM is exposed to in the IoT have been surveyed. The study about big data in the IoT has been presented, in which the issues related to data mining architectures have been set forth. One of the prominent architectures in this field is the multilayered architecture of data mining in the IoT, which is based on four layers, i.e., data collection layer, data management layer, event processing layer, and data mining service layer. In this architecture, data is collected through RFID readers, GPS, etc., and clean data is delivered to the final layer for processing.

One of the most significant challenges in this architecture is energy consumption, bad reading, reading duplicate information, fault tolerance, data filtering, and communication. Another architecture discussed in this paper is the distributed architecture of data mining in the IoT, whose primary purpose is to preprocess distributed data before being transmitted to the network core in order to reduce energy consumption in the central nodes. In this architecture, tools such as RFID, WSN, etc., are hired to collect data. Reducing energy consumption in processing nodes is one of the main advantages of this architecture. Another architecture

addressed in the references is the grid data mining architecture in the IoT, focusing on the strategies to increase portability and situational awareness. In this architecture, grid resources, RFID, WSN, and other components are used for data collection, and increasing resource availability is one of its noteworthy achievements. Overhead costs to increase availability and to create different distributions of data are one of the significant challenges in this architecture.

Another architecture is the data mining model equipped with the power to anticipate the multi-technology aggregation in the IoT, in which the aggregation of several technologies, including cloud computing technology, grid computing, pervasive networks, secure networks, and so on, is used. In this architecture, the internet, intranet, sensors, RFID, wireless, etc., are employed to collect data. A key property of this architecture is the potential to use IPv6, which is one of the main challenges in data collection from location-aware applications. Another architecture under discussion is the one for IoT data mining based on cloud computing, whose primary goal is to aggregate the extraction, management, exploration, and interpretation layers. Here, like other existing architectures, the internet, intranet, sensors, RFID, wireless, and so on are used, and its most significant advantage is supporting social networks and cloud computing in the IoT. Therefore, it is challenging with the security in social networks.

At last, a three-dimensional five-layered architecture has been proposed for big data mining in the IoT, the central concept behind which is a three-dimensional five-layered architecture made up of device layer, raw data layer, data collection layer, data processing layer, and service layer. The other two dimensions include issues such as complying with the related standards, security and privacy, data management, data perception, and data interpretation. This also involves using a large number of IoT devices, including sensors, RFID, cameras, and other devices for data collection. The main merit of this architecture is complying with the standards, security, and privacy, data management, data perception, and data interpretation, and the overhead resulting from two-layered processing added to more traditional models is one of its fundamental challenges.

## Author Contributions

Mohammad Saber Iraji created the research framework, performed the literature review, established the methodology, and drafted the initial version of the manuscript. Ali Yavari played a role in conceptualizing the study, offered critical insights on data mining architectures, and helped revise and finalize the manuscript. Both authors reviewed and approved the final version.

## Funding

This study did not receive any external financial support from public, commercial, or non-profit entities.

## Data Availability

The research is based on previously published literature, theoretical frameworks, and an architectural analysis of IoT-based data mining. All data that support the findings of this study are included in the article. Further materials can be requested from the corresponding author upon a reasonable inquiry.

## Conflicts of Interest

The authors declare that there are no conflicts of interest concerning the publication of this article.

## References

- [1] Venu, N., Kumar, A., & Vaigandla, K. K. (2022). Review of internet of things (IoT) for future generation wireless communications. *International journal for modern trends in science and technology*, 8(3), 1–8. <https://ssrn.com/abstract=4232170>



- [2] Zhong, Y., Chen, L., Dan, C., & Rezaeipannah, A. (2022). A systematic survey of data mining and big data analysis in internet of things. *The journal of supercomputing*, 78(15), 18405–18453. <http://dx.doi.org/10.1007/s11227-022-04594-1>
- [3] Sunhare, P., Chowdhary, R. R., & Chattopadhyay, M. K. (2022). Internet of things and data mining: An application oriented survey. *Journal of king saud university-computer and information sciences*, 34(6), 3569–3590. <https://doi.org/10.1016/j.jksuci.2020.07.002>
- [4] Bi, Z., Jin, Y., Maropoulos, P., Zhang, W. J., & Wang, L. (2023). Internet of things (IoT) and big data analytics (BDA) for digital manufacturing (DM). *International journal of production research*, 61(12), 4004–4021. <http://dx.doi.org/10.1080/00207543.2021.1953181>
- [5] Qi, Q., Xu, Z., & Rani, P. (2023). Big data analytics challenges to implementing the intelligent Industrial Internet of Things (IIoT) systems in sustainable manufacturing operations. *Technological forecasting and social change*, 190, 122401. <http://dx.doi.org/10.1016/j.techfore.2023.122401>
- [6] Brohi, S., Marjani, M., Hashem, I., Ramiah Pillai, T., Kaur, S., & Amalathas, S. (2019). A data science methodology for internet-of-things. In *Emerging technologies in computing* (pp. 178–186). [http://dx.doi.org/10.1007/978-3-030-23943-5\\_13](http://dx.doi.org/10.1007/978-3-030-23943-5_13)
- [7] Shirvanian, N., Shams, M., & Rahmani, A. M. (2022). Internet of things data management: A systematic literature review, vision, and future trends. *International journal of communication systems*, 35(14), e5267. <https://doi.org/10.1002/dac.5267>
- [8] Li, X., Liu, H., Wang, W., Zheng, Y., Lv, H., & Lyu, Z. (2022). Big data analysis of the Internet of Things in the digital twins of smart city based on deep learning. *Future generation computer systems*, 128(10), 167–177. <http://dx.doi.org/10.1016/j.future.2021.10.006>
- [9] Ali, A., Hussain, T., Tantashutikun, N., Hussain, N., & Cocetta, G. (2023). Application of smart techniques, internet of things and data mining for resource use efficient and sustainable crop production. *Agriculture*, 13(2), 1–22. <https://doi.org/10.3390/agriculture13020397>
- [10] Zhang, H., & Yuan, S. (2023). How and when does big data analytics capability boost innovation performance? *Sustainability*, 15(5), 1–19. <https://doi.org/10.3390/su15054036>
- [11] Abughazala, M., & Muccini, H. (2023). *Modeling data analytics architecture for data-driven IoT applications using DAT*. IEEE. <http://dx.doi.org/10.1109/ICSA-C57050.2023.00066>
- [12] Yang, H., Zhou, L., Cai, J., Shi, C., Yang, Y., Zhao, X., ... & Yin, X. (2022). *Data mining techniques on astronomical spectra data. II: Classification analysis*, 518(4), 5904–5928. <https://doi.org/10.1093/mnras/stac3292>
- [13] Alrehaili, G., Galam, N., Alawad, R., & Albraheem, L. (2023). *Cloud-Based big data analytics on IoT applications*. IEEE. <http://dx.doi.org/10.1109/ITIKD56332.2023.10100150>
- [14] Finogeev, A. G., Parygin, D. S., & Finogeev, A. A. (2017). The convergence computing model for big sensor data mining and knowledge discovery. *Human-centric computing and information sciences*, 7(1), 1–11. <https://doi.org/10.1186/s13673-017-0092-7>