Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

Important Features Selection Method for Temporal and Information Aware Clustering of Internet of Things Stream Data

[1]Jyoti Yadav, [2],*Ihsan Hamza Jumaa, [3] Masoumeh Norouzifard

[1],[2] Department of Computer Science, Savitribai Phule Pune University

[2],[3] Department of Information Technology, Rwandz Private Technical Institute

[2] Department of Information Technology, Aynda Private Technical Institute

[1] yadav.jyo@gmail.com, [2],*Ihsan.hamza89@gmail.com, [3] Mas.norouzifard@gmail.com

Abstract

The Internet of Things (IoT) is essential because it creates new services that allow multiple disparate devices to be seamlessly connected over the Internet. The Internet of Things is growing every second in several industries, including smart cities, smart homes, smart factories and oil and gas. The high dimensionality of data has made machine learning and data mining more challenging. This study presents a comprehensive overview of advanced dimensionality reduction techniques that aid in the selection of essential features for machine learning and IoT-based data analytics. The proposed method for clustering of IoT data uses Finite Differential method (FDM), Principal component analysis (PCA) and K-mode Clustering. These techniques are based on criterion measures, and datasets, and are inspired by soft computation technology. In Internet of Things applications, the resulting clustered features can help identify and reduce uncertainty, improve data clarity, and minimize data loss. The article discusses potential avenues for future research and provides readers with information regarding the suitability of various data reduction techniques by providing an overview of Dimensionality Reduction applications across various fields.

Keywords: Internet of Things; Dimensionality reduction; Principal component analysis; Finite Differential method; K-Mode Clustering; Machine Learning.

1. Introduction

The Internet of Things (IoT) is intended to give solutions to practical issues to enhance quality of life. Connecting physical things to the Internet through wired or wireless technology is the cornerstone of the Internet of Things. IoT offers a wide range of solutions, including e-health, smart grid, smart homes, intelligent transportation systems, and environment monitoring [1]. A vast number of IoT devices are being created as a result of the design of complex IoT solutions and their adaptation. There will be 29 billion connected devices by 2025. analysts at Ericson predict [2]. These numerous devices produce a significant amount of data that is transmitted over the Internet in a variety of different formats, including plain messages, pictures, audio, and video. Future habitation in smart cities and corporate operations may be

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

significantly impacted by the knowledge gained from the data gathered by billions of IoT devices. The heterogeneous data produced by various sources can offer in-depth knowledge about the environment, users, and physical things.

Low-power gadgets, a distributed structure, meager computational capability, and heterogeneous data are a few of the IoT's major characteristics. It is difficult to read, write, and analyze a high volume of data due to low-power IoT devices. So, to handle a significant amount of data while keeping in mind the limitations of IoT devices, we need efficient and reliable ways. Additionally, there is a requirement to combine heterogeneous data produced by various IoT device types. Finally, it is necessary to extract important data from a large amount of data. Dimension reduction, classification, regression, clustering, outlier identification, and other activities can be used to analyze the data produced by large-scale IoT devices [3]. To lower the communication and computing costs for data transmission and task-specific design change, we need innovative algorithms and methodologies for data analysis.

Future life will be drastically altered by the Internet of Things [4]. However, processing the huge amounts of data with high dimensions produced by IoT applications is quite difficult [5].

High dimensional data contains redundant and unnecessary features, which not only increases computational complexity but also significantly lowers classification systems' efficacy and accuracy. High dimensional data could be the source of the dimensionality curse in the interim. Dimensionality reduction techniques are suggested to address these issues by analyzing the relevance and redundancy of IoT big data, preserving relevant features, and eliminating as many irrelevant and redundant aspects as possible. These techniques can lower the cost of computation, improve the learning model, and prevent over-fitting issues when processing massive data for the Internet of Things [6].

The two basic areas of dimensionality reduction are feature extraction and feature selection [7]. The original features are projected into a new low-dimensional feature space during feature extraction, which lowers the dimensionality. Ng et al.'s [8] introduction of a useful dissimilarity measure for the K-Modes clustering technique expands on the conventional simple matching method by taking into account the frequency of mode components in the current cluster.

When creating new, creative company ideas, it could be used as a preparation task while keeping security, assurance, and interoperability in mind, which uses a variety of machine learning methods to cut down on the number of variables. Both feature extraction and feature selection are involved in this [9]. Finding a subset from a multi-dimensional dataset is the goal of feature selection dimension reduction strategies. In this area, filter, wrapper, and embedding are the most important techniques. However, only a small subset of the multi-dimensional data is recovered during feature extraction. Principal component analysis (PCA) and linear discriminant analysis fall under this category of approaches. We primarily give an overview

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

and a list of categories for dimension reduction. We list several difficulties with dimension reduction in IoT systems. We also go over several examples of dimension reductions utilizing various methods and strategies. Finally, we emphasize unresolved challenges in dimension reduction research for Internet of Things systems.

Well-known numerical techniques for solving differential equations are finite difference methods (FDM), which approximate the derivatives using various differentiation schemes [10].

In this case study, we examine reviews of the literature that contain comparable works.

For intrusion detection in IoT backbone networks, authors in [11] developed a two-layer dimension reduction method. Component analysis and dimension reduction using linear discriminate analysis make up these layers. Additionally, the authors looked into the two-tier classification module, which uses Naive Bayes and the certainty factor version of K-nearest Neighbor to spot suspicious activity. For the study of huge data, authors in [12] developed a novel strategy based on PCA. The recommended approach can If the linear regression method is then applied to the data analysis, give an exact solution. In [13], authors recommended a framework for crowdsourcing improved comprehension of social data supplied by users based on IoT. Dimension reduction lowers both the amount of data that needs to be kept and the cost of communication. At the customer end, a framework for massive data reduction in IoT is put forward [14]. A commercial model for end-to-end data reduction in enterprise applications was also offered by the authors. In [15], scientists recommended that camera sensor networks should cover the entire field of view. It is demonstrated that choosing a certain set of points can minimize the minimal number of full-view area coverage to the minimum number of full-view points. Based on the investigation of the geometric relationship between full-view coverage and conventional coverage, the authors presented a greedy algorithm and a set-covered-based approach. A method for real-time data reduction at the network edge was presented by authors in [16]. The recommended solution automates changing between various data handling algorithms. Based on the concept of perceptually important points, three variations of the suggested algorithm are shown. In [17], the idea of a "-kernel dataset" is put out, which uses a tiny subset of data to represent a significant amount of information from wireless sensor networks. The recommended algorithm's information loss rate is less than, which is an arbitrarily low figure. To conserve energy and computing resources, authors proposed distributed approaches (accurate algorithm and sampling-based approximate algorithm) to decrease kernel dataset.

The Internet of Things (IoT) applications produce a data flood that is full of many kinds of important information. The crucial questions, however, are how to process these data and how to extract the valuable information from data [18]. A potential solution for the IoT's large data issue is to collect the data that is helpful. Feature selection that lessens the complexity of input data is one of the representative research fads. Finding a subset of the least redundant and

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

irrelevant features is the aim of feature selection to achieve satisfactory performance on predetermined evaluation criteria. Subset creation, subset evaluation, stop condition, and result validation are the four main components of filter methods [19]. Given that most real-world datasets lack previous knowledge for evaluation, the prediction accuracies of particular classifiers are utilized to assess the chosen feature subset following feature selection [20].

Typically, feature ranking and feature subset selection are two widely used subset creation techniques for filter systems [21]. Each feature is ranked according to its importance ratings, which are determined by predetermined criteria based on information theory, distance, and other factors [22]. The Top-K features are then retained by a threshold value of important score set by the users after removing irrelevant or weakly relevant features. In other words, feature redundancy cannot be handled by feature ranking techniques, although they are nonetheless effective in terms of the linear computing complexity of the feature dimension. Typical feature selection experiments were more lucrative the rating of features. Kira's method of representative feature ranking is called relief likewise Rendell's [23]. First, a single instance is chosen at random, after which two further instances, one belonging to the same class and the other to a different class, are chosen. The Manhattan distance of the chosen instance and its two closest instances is then used to update each feature's essential score. A robust approach dubbed ReliefF, proposed by Kononentko [24], which can even handle noisy and missing data, expanded the original relief. There is also widespread use of information metrics such as the information gain (IG) [25], information gain ratio [26], and SU [27] Provides a comparison of various feature ranking techniques. The findings demonstrate that, due to the existence of, the Top-K most significant features chosen by those approaches may not yield the highest classification performance.

The limitation of proposed works data standardization is necessary, which could have an impact on the variables' initial scale and meaning. If the number of principal components is not selected appropriately, information loss could result. The number of features in the dataset equals the maximum dissimilarity score between two data points. In K-Modes clustering, the final clusters rely significantly on the initial centroids.

2. Proposed Methodology

2.1 IoT data collection

The setting of the smart home was specifically created for this study to facilitate data collecting. A Google Nest Hub, a Kasa Cam, a SmartThings Multipurpose Sensor, a SmartThings Motion Sensor, and a SmartThings Outlet for smart homes were all present in the environment.

To gather data, the steps below were taken.

(a) Device activation;

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

(b) Companion app download & device enrollment;

(c) Identifying functions of devices;

(d) Experiment with the devices.

The Smart Thing Hub acts as a base station for data transmission between low-end smart home gadgets linked to it. Each smart home equipment used in the experiment is compatible with supported smart speakers, but it is also possible to use it independently. Given its popularity, the Google Nest Hub was used as a smart speaker in this study. A smart speaker or a smartphone app can use voice commands to control the installed smart home appliances.

The functions of each experimentation tool. The Google Nest Hub has a variety of home controls, including the ability to operate a camera or an outlet. With the help of the hub's display, connected devices may be viewed and managed in a single window without switching between apps. The smart camera Kasa cam's videos can also be seen on the display. The video chatting Google Duo app enables calls between smartphones and the Google Nest Hub. Using the Kasa Smart app, you may view the Kasa cam live. The camera records footage and notifies a smartphone when it detects motion and sounds. Alternatively, a smartphone can record the feed. The camera contains an integrated speaker and microphone that allow two-way communication between the camera and its companion app. Power on/off can be controlled via the SmartThings Outlet. The SmartThings Multipurpose Sensor can recognize vibrations, open and closed states, and other events. Both the motion sensor and the multifunctional sensor can measure temperature. Figure1 shows our steps of work from our data collection till clustering.

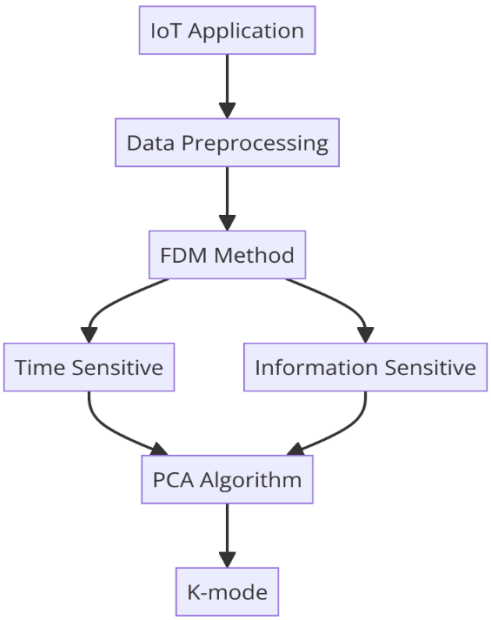

Figure 1. proposed flow diagram

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

2.1.1 Features of dataset

For the reason of enabling farm management and increasing productivity, the AgTech sector is utilizing satellite imaging data, Landsat data, IoT data, insights, and driving through spatial analytic methodologies. In the end, it will increase agricultural productivity and efficiency while lowering food production costs. Cloud computing in agriculture enables extensive data collection and retrieval from various sources. Soil conditions, crop mapping, agricultural environment monitoring, satellite photos, yield information management, and many more sorts of data are just a few examples. They provide information quickly and precisely. Since the data is still kept in the cloud, it is always available. Farmers can solve agricultural production issues by using historical data.

Different Government line departments, the business sector, community organizations, and academic institutions all contribute to the smooth operation of cities by offering infrastructure, services, research, co-creation, and helpful feedback. Data that is available to these institutions is locked up in silos and needs to be shared with them. Therefore, it is essential to unlock urban data and make it the universal language of collaboration in the urban environment in order to unleash its transformative potential. Therefore, the DataSmart Cities initiative was conceived to successfully institutionalize a "Culture of data", drive data governance and policy formulation, promote data sharing and exchange, and promote multidisciplinary research to achieve co-creation, open innovation, and citizen empowerment.

Using the latest IoT and industrial internet technologies, including smart sensors and sensing, computation and predictive analytics, and robust control technologies, the smart factory outlines a new approach to multi-scale manufacturing. To collect, transfer, interpret, and analyze the data, as well as to properly control the manufacturing process, these technologies must be linked together.

For the O&G sector, handling massive amounts of data is not a novel problem, in contrast to other businesses. Petroleum corporations have long been investing in seismic data processing and visualization tools to assist them in planning how to collect and comprehend what is beneath the surface of the earth. This opens the door for the acceptance of IoT technologies in the oil and gas industry more quickly and confidently. Nevertheless, the shift to digital and IoT technology marketing lags in the O&G sector behind other industries. The networks and services that are linked to the use of IoT also pose problems about data privacy and cyber security. An O&G operation's legacy asset base is not designed for cyber security, and the system is susceptible to cyber-attacks due to a deficiency of monitoring tools in the current networks. Examining these problems promotes technology adoption and improves knowledge of the advantages of the Internet of Things.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

The industrial data contained should be a smart city, smart farming, smart factory and oil and gas. The consumer data contain should be healthcare, wearable, self-driving, and traffic management. Another one commercial data should be retail, smart home, entertainment venues, and hotels these are all our feature data sets.

2.1.3 Number of records

The data has been collected from three IoT smart applications such as smart home, Healthcare and smart city. These applications can be broadly classified as consumer application, industrial application and commercial application.

Table1. Dataset Information

Dataset Name	Number of Records	Number of Features
Healthcare	2127	21
Smart Home	503911	30
Smart City	10000	23

- ***Data preprocessing***

The data collected from the IoT sources are preprocessed using different stages such as handling missing values, label encoding, and Z-score normalization techniques. These techniques are discussed below:

1. Handling outliers and missing value

The rapid expansion of IoT technology has led to the collection of vast amounts of data, primarily from sensor devices. It is common to encounter significant amounts of missing values in this data when managing large volumes [28]. Sensor data often has missing values due to errors in data collection and transmission. Common-mode failures cause data with missing values to persist as a longstanding challenge in IoT architecture, potentially leading to bias and loss of precision [29]. Various factors contribute to incomplete data, such as weak network connectivity, malfunctioning sensor systems, external influences, and synchronization issues. Finally, an IoT model is developed for imputing missing values, using the model to estimate the missing data. The total number of missing values in different columns of Healthcare dataset that were handled are tabulated in Table 2. In Table 2, there are 918 missing values for the "operator" feature, which are addressed using the '*mean'* method. A total of 8 features had missing values and were handled for further processing.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

Table 2. Handling Missing Values from Healthcare Dataset

Feature Name	#Missing Values
Speed	3
Sound	4
Longitude	4
Date	6
gsm_signal	11
Operator	918
battery_temperature	8
app_version	5

2. Label encoder (convert categorical to numerical)

The label encoding technique is applied to convert the categorical variables into numerical values. This helps the machine learning models to process the data samples and generate the binary output. In this method each category is given a unique integer label. Categories are assigned integer values starting from 0. For instance, in the "network_type", "operator", and "device model" columns, each unique category has been replaced with a distinct integer label. In order to apply this technique, the categorical features in the dataset are identified and are grouped into separate categories. These are the variables that are subjected to label encoding. Categorical features are typically represented as text or discrete values which can be either nominal or ordinal. The results after converting categorical variables into numerical variables are tabulated in Table 3.

Table 3. Conversion of Categorical Features to Numerical Features in Healthcare Dataset

Rec. No	Categorical Feature Name			Numerical Feature Name		
	network_type	Operator	device_model	network_type	operator	device_model
0	UMTS	Movistar	GT-I9195	1.44	-1.27	-0.81

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

1	UMTS	Movistar	GT-I9196	1.44	-1.27	-0.81
2	UMTS	Movistar	GT-I9197	1.44	-1.27	-0.81
3	UMTS	Movistar	GT-I9198	1.44	-1.27	-0.81
4	UMTS	Movistar	GT-I9199	1.44	-1.27	-0.81
...	...	...	...	...	...	...
9995	Cido	Telenor HU	Aquaris E5 HD	-2.22	-0.49	-1.77
9996	GPRS	Pannon	Aquaris E5 HD	-1.18	0.57	-1.77
9997	GPRS	Telenor HU	Aquaris E5 HD	-1.18	-0.49	-1.77
9998	GPRS	Telenor HU	Aquaris E5 HD	-1.18	-0.49	-1.77
9999	GPRS	Telenor HU	Aquaris E5 HD	-1.18	-0.49	-1.77

3. Normalization used z-score

The standard score, also known as a z-score, is a very helpful statistic because it enables comparison between two scores that come from different normal distributions and allows programmers to determine the likelihood that a given score will occur within our normal distribution (a). To put it another way, the standard score achieves this by standardizing scores from a normal distribution by converting them to z-scores in a standard normal distribution. (Source:) Every input value has been normalized by the following Eq. 1:

$$Z(i,j) = \frac{a(i,j)-\mu}{\alpha}, \quad (1)$$

where

$Z(i,j)$new value

$a(i,j)$old value

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

μ-column of the input value

α-standard deviation of the column of the input value.

$i = 1 - n$ (n number of rows or inputs)

$j = 1 - k$ (k number of bits that represents each input 32 bits)

Data normalization is a technique in data mining that transforms dataset values to a common scale. This is crucial because many machine learning algorithms are sensitive to the scale of input features and tend to generate more accurate outcomes when the data is normalized. This technique helps in comparing two scores obtained from different normal distributions based on which the probability that a given score will occur within a specific normal distribution can be obtained.

The values after applying the Z-score normalization technique are tabulated in Table 4. It can be observed from Table 4 that after Z-Score normalization the feature values lie between -3 and 3.

Table 4. Z-Score Normalization on Healthcare Dataset

Rec. No	Z-Score Normalization (Before)			Z-Score Normalization (After)		
	altitude	battery_level	magnetism	altitude	battery_level	magnetism
0	148	98.00	6,160,551	0.095	1.12	0.97
1	148	98.00	61,565,234	0.095	1.12	1.17
2	148	84.00	5,835,965	0.095	0.73	0.51
3	148	84.00	5,841,368	0.095	0.73	0.51
4	178	74.00	6,170,769	0.26	0.45	0.98
...	...	...	...	...	...	...
9995	236.8946	128.90	-34,642,971	0.59	1.99	-1.6
9996	247.0286	134.46	-41,681,927	0.63	2.15	-1.59
9997	257.1626	140.03	-48,720,883	0.69	2.30	-1.59

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

9998	267.2966	145.60	-55,759,838	0.74	2.46	-1.58
9999	267.2966	146.60	-55,759,838	0.74	2.49	-1.58

- ***Finite Differential method (FDM) for data clustering***

Finite-difference methods (FDM) are a set of numerical techniques used in numerical analysis that approximate derivatives using finite differences in order to solve differential equations. Discretization, or the division of the spatial and time domains into a finite number of intervals, is used to approximate the values of the solution at the ends of the intervals. This is done by solving algebraic equations that involve values from close points and finite differences.

The basic idea as well as the FDM are covered in this part of the section. Finite variance differential equations computational solutions give us the values at separate grid points. Let's look at a domain in the xy plane. considering Δx and Δy, the grid point spacing in the x- and y-directions is considered to be identical. It is not always required for Δx and Δy to be equivalent or uniform. Index i represents the grid points in the x-direction, and index j does the same in the y-direction. The idea behind FDM is to use algebraic variance methods to maintain the position of the derivatives products in the governing differential equation. This produces an algebraic equation system that can be solved using any common analytical or numerical technique to determine the variables that are dependent at the discrete grid points.

Forward difference method

$$\Delta y_i = \frac{dy(x)}{dx}|x = x_i \approx \frac{y_{i+1}-y_i}{x_{i+1}-x_i} = \frac{y_{i+1}-y_i}{\Delta x} = \frac{y_{i+1}-y_i}{h}. \quad (2)$$

Where, h is the increment step size

Backward difference method

$$\nabla y_i = \frac{dy(x)}{dx}|x = x_i \approx \frac{y_i-y_{i-1}}{h}. \quad (3)$$

Central difference method

$$\Delta y_i \approx \frac{y_{i+1}-y_i}{x_{i+1}-x_i} = \frac{y_{i+1}-y_{i-2}}{2h}. \quad (4)$$

The application of FDM for all the three IoT based smart applications is illustrated in six figures, viz. *Figure 1 to Figure 6* for some features: fetal_movement, uterine_contractions, summary, use [kw], pressure and magnetism. The following six graphs display the comparison between the original data and the finite differences for different features from three IoT applications for first 500 records. The blue line depicts preprocessed data and the orange line

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

represents the finite differences calculated from the original data that helps in categorizing the features as time or information sensitive feature.

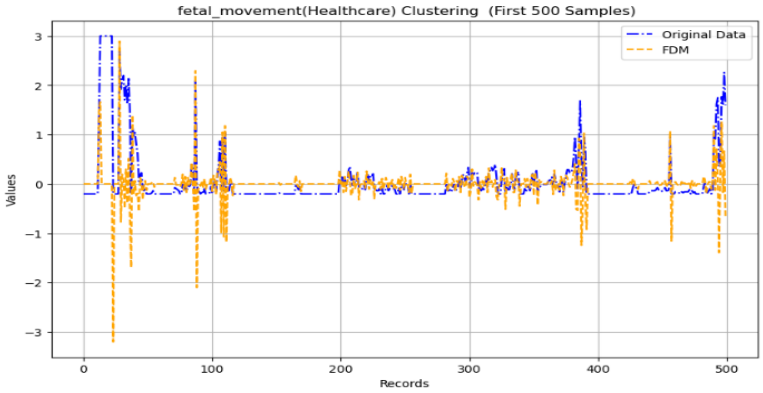

Figure 1. Information Sensitive Feature in Healthcare

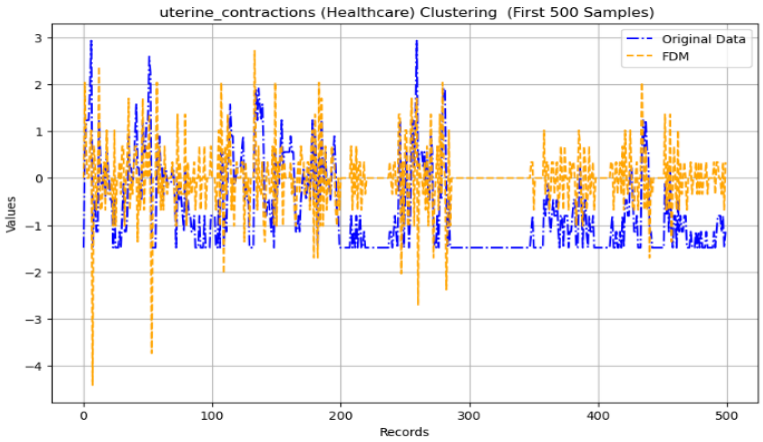

Figure 2. Time sensitive Feature in Healthcare

The FDM method classifies features as either information-sensitive or time-sensitive based on their statistical properties and patterns observed in the data. The explanation of why the feature fetal_movement is classified as information-sensitive and uterine_contractions as time-sensitive is as follows:

Fetal Movement (Information-Sensitive)

- Low Variability Over Time: The fetal_movement data shows less variability and more stability over time. This suggests that the feature provides more information about the underlying state without being highly influenced by time-dependent changes.
- Predictable Patterns: The stable and predictable nature of the data makes it easier to extract meaningful information that can be directly linked to specific conditions or states, which is a characteristic of information-sensitive features.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

Uterine Contractions (Time-Sensitive)

- High Variability Over Time: The uterine_contractions data shows significant variability over time, indicating that this feature is highly influenced by time-dependent changes.
- Temporal Dynamics: The patterns in uterine contractions are more dynamic and change significantly over time, reflecting the temporal nature of this feature. This makes it time-sensitive as it captures the changes and variations that occur over specific time periods.

Visualization Insights:

- Fetal Movement: The graph shows less fluctuation and more consistent values, which aligns with the idea of being information-sensitive. The FDM clustering lines (dashed orange) closely follow the original data (dashed blue), indicating consistent information extraction.
- Uterine Contractions: The graph shows more fluctuations and variations, indicating that it captures dynamic and time-varying patterns. The FDM clustering lines (dashed orange) show more deviation from the original data (dashed blue), capturing the time-sensitive nature of the feature.

The classification by the FDM method likely relies on analyzing the statistical properties of these features. Fetal_movement is more stable and consistent, making it information-sensitive, whereas uterine_contractions show high temporal variability, making it time-sensitive. This differentiation helps in understanding and processing the data more effectively in Healthcare applications. Similar conclusions are observed in Smart Home (features summary and use [kW]) and Smart City (features pressure and magnetism) applications.

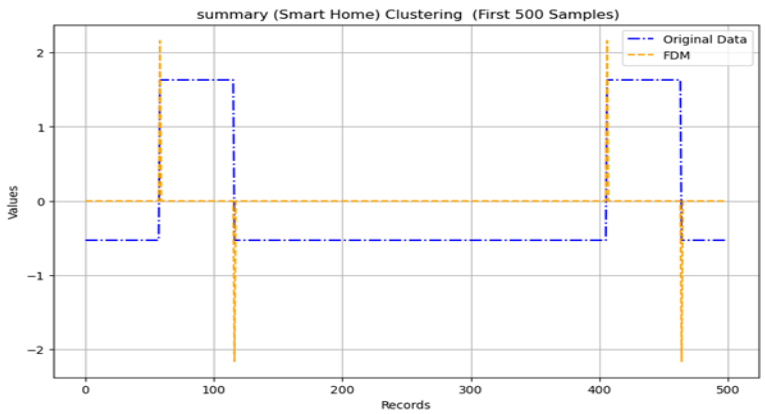

Figure 3. Information Sensitive Feature in Smart Home

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

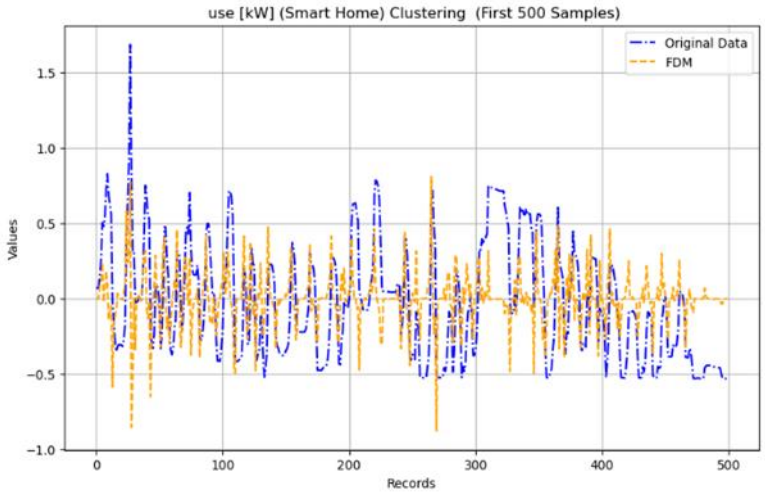

Figure 4. Time Sensitive Feature in Smart Home

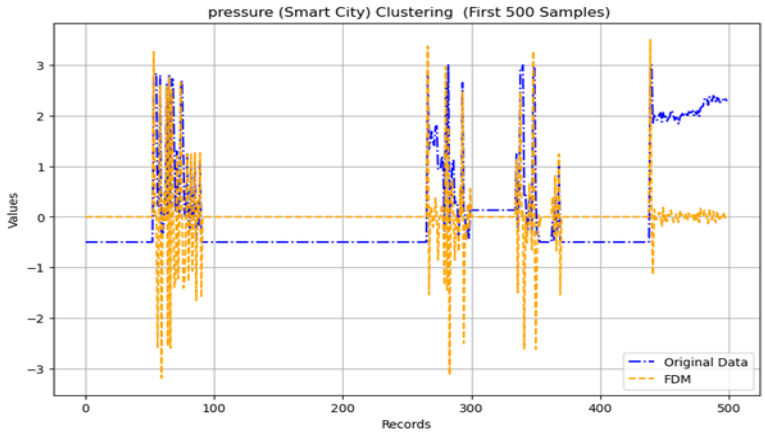

Figure 5. Information Sensitive Feature in Smart City

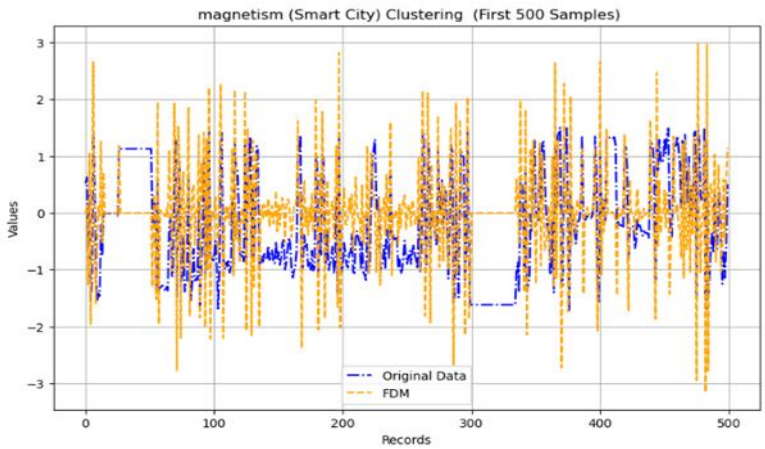

Figure 6. Time Sensitive Feature in Smart City

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

Based on the data clustered using the FDM technique, the features are classified as time or information sensitive for all three different smart applications and the same have been represented in *Tables 5, 6 and 7* respectively.

Table 5. Classification of Features in Healthcare Dataset

Time Sensitive Features	Information Sensitive Features
uterine_contractions,	baseline value
abnormal_short_term_variability	Accelerations
mean_value_of_short_term_variability	fetal_movement
mean_value_of_long_term_variability	light_decelerations
histogram_width	severe_decelerations
histogram_number_of_peaks	prolongued_decelerations
histogram_mean	percentage_of_time_with_abnormal_long_term_variability
histogram_median	histogram_min
histogram_variance	histogram_max
----------	histogram_number_of_zeroes
----------	histogram_mode
----------	histogram_tendency

Table 6. Classification of Features in Smart Home Dataset

Time Sensitive Features	Information Sensitive Features
Time	dishwasher [kw]
use [kw]	kitchen 38 [kw]
gen [kw]	well [kw]
furnace 1 [kw]	Temperature

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

furnace 2 [kw]	Humidity
home office [kw]	Visibility
fridge [kw]	Apparenttemperature
wine cellar [kw]	Pressure
garage door [kw]	Windspeed
kitchen 12 [kw]	Windbearing
kitchen 14 [kw]	Precipintensity
barn [kw]	Dewpoint
microwave [kw]	Precipprobability
living room [kw]	Icon
solar [kw]	Summary

Table 7. Classification of Features in Smart City Dataset

time sensitive features	information sensitive features
user_id	Temperature
sm_id	Light
Sound	Humidity
Latitude	Pressure
Longitude	Speed
Altitude	Rotation
gsm_signal	Gravity
battery_level	network_type
battery_temperature	Operator
Magnetism	device_model
Date	app_version
Acceleration	----------

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

○ ***Time Sensitive Data***

When a task is time-sensitive, it must be finished by a certain date and time. The task's required time is what we use to determine the key's age. Nevertheless, current data aggregation schemes are unable to satisfy the demands of personalized services if they do not account for the time factor.

Example

If compromised, a variety of sensitive data kinds could seriously hurt an individual, company, or government organization. These are a few typical instances of sensitive data.

○ ***Information Sensitive Data***

Sensitive data is information that is kept private, processed, or managed by a person or organization. Only authorized users who have the necessary rights, privileges, or clearance to view it can access it. Because of the potential consequences, if this information falls into the wrong hands, it is regarded as sensitive.

● ***PCA for feature importance selection***

The proposed method for performing feature selection using PCA is presented in this section. As is well known, dimension reduction is typically the aim of both feature extraction and feature selection. We begin with feature extraction based on PCA. Assuming x to be an eigenvector of the PCA covariance matrix, we understand the feature extraction outcome of any random sample vector.

Therefore, choosing features that enable a precise description of the defect condition and, in turn, trustworthy defect classification, diagnosis, and prognosis is the aim of feature selection. To decrease the dimensionality of the input features for supervised and unsupervised classification, the PCA technique was created. This is predicated on the idea that while increasing the computational load, a large number of inputs may not always improve the efficacy of defect classification.

In generally, the PCA method transforms n vectors $(x_1, x_{2,\dots}x_{i,\dots,}x_n)$ from d- dimensional space to n vectors $(x_1', x_2', x_i', \dots, x_n')$ in new d' dimensional space.

$$X_i' = \sum_{k=1}^{d'} a_{k,i} e_k, d' \leq d, \tag{5}$$

where e_k- eigenvectors, d'-largest eigenvalues, $a_{k,i}$-projections of the original vectors on the eigenvectors.

Using principal component analysis (PCA), one can determine which features are crucial for accurately characterizing variation in a data collection. When working with huge data sets that were originally high dimensional, it is most frequently employed to reduce the dimensionality

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

of the data to make machine learning more feasible. Use the components_ property to determine the significance of each feature on each component. A PCA loading array with "rows" denoting components and "columns" denoting the original features is the end product. *Table 8 to Table 13* for Healthcare, smart home and smart city datasets respectively. The details of the tables are explained as follows:

1. **PC:** This identifies the principal component to which the feature is associated. The principal components are new variables created in PCA that aim to capture the most variance in the data. These are denoted as PC1, PC2, PC3, etc.

2. **Feature:** This lists the original features from the dataset that were included in PCA. Each feature represents a specific aspect or measurement within the dataset.

3. **Loading:** The loading value indicates the contribution of a feature to the corresponding principal component. A higher absolute value suggests a stronger contribution. Positive loadings imply a direct relationship, while negative loadings imply an inverse relationship with the principal component.

4. **Type:** This categorizes the features based on their significance or relevance to the principal component. "Best" indicates features with strong loadings that significantly contribute to the principal component, while "Weak" indicates features with lower loadings that contribute less to the principal component.

Table 8. Importance of Features using PCA for Healthcare (Time sensitive)

	PC	Feature	Loading	Type
0	PC1	mean_value_of_short_term_variability	0.45	Best
1	PC2	histogram_median	-0.53	Best
2	PC3	abnormal_short_term_variability	-0.55	Best
3	PC4	uterine_contractions	0.81	Best
4	PC5	abnormal_short_term_variability	-0.60	Best
5	PC6	histogram_variance	0.71	Best
6	PC7	mean_value_of_short_term_variability	0.81	Best
7	PC3	mean_value_of_long_term_variability	0.55	Weak

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

8	PC1	histogram_width	0.43	Weak
9	PC6	histogram_number_of_peaks	-0.55	Weak
10	PC2	histogram_mean	-0.51	Weak

Table 9. Importance of Features using PCA for Healthcare (Information sensitive)

	PC	Feature	Loading	Type
0	PC 1	histogram_min	-0.45	Best
1	PC 2	histogram_max	0.51	Best
2	PC 3	Accelerations	-0.51	Best
3	PC 4	prolongued_decelerations	0.55	Best
4	PC 5	severe_decelerations	0.51	Best
5	PC 6	severe_decelerations	0.82	Best
6	PC 7	histogram_number_of_zeroes	0.78	Best
7	PC 8	prolongued_decelerations	0.67	Best
8	PC 9	percentage_of_time_with_abnormal_long_term_var	0.78	Best
9	PC 10	Accelerations	-0.55	Best

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

10	PC 4	baseline value	0.39	Weak
11	PC 5	fetal_movement	-0.5	Weak
12	PC 10	light_decelerations	-0.54	Weak
13	PC 1	histogram_mode	-0.45	Weak
14	PC 3	histogram_tendency	0.50	Weak

Table 10. Importance of Features using PCA for Smart Home (Time sensitive)

	PC	Feature	Loading	Type
0	PC1	gen [kw]	-0.59	Best
1	PC2	use [kw]	0.49	Best
2	PC3	wine cellar [kw]	0.5	Best
3	PC4	living room [kw]	0.55	Best
4	PC5	barn [kw]	0.65	Best
5	PC6	kitchen 12 [kw]	0.83	Best
6	PC7	garage door [kw]	0.71	Best
7	PC8	microwave [kw]	0.54	Best
8	PC9	home office [kw]	0.53	Best
9	PC10	fridge [kw]	0.76	Best
10	PC11	living room [kw]	-0.68	Best
11	PC12	Time	-0.70	Best
12	PC13	furnace 1 [kw]	0.77	Best

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

13	PC13	furnace 2 [kw]	-0.48	Weak
14	PC8	kitchen 14 [kw]	-0.52	Weak
15	PC1	solar [kw]	-0.59	Weak

Table 11. Importance of Features using PCA for Smart Home (Information sensitive)

	PC	Feature	Loading	Type
0	PC1	Precipprobability	0.42	Best
1	PC2	Temperature	-0.54	Best
2	PC3	Windspeed	0.60	Best
3	PC4	well [kw]	0.68	Best
4	PC5	kitchen 38 [kw]	0.87	Best
5	PC6	well [kw]	0.71	Best
6	PC7	pressure	0.55	Best
7	PC8	summary	-0.61	Best
8	PC9	windbearing	0.73	Best
9	PC10	visibility	-0.8	Best
10	PC11	humidity	-0.58	Best
11	PC4	dishwasher [kw]	0.68	Weak
12	PC2	apparenttemperature	-0.54	Weak
13	PC8	precipintensity	0.52	Weak
14	PC2	dewpoint	-0.46	Weak

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

15	PC1	icon	0.39	Weak

Table 12. Importance of Features using PCA for Smart City (Time sensitive)

	PC	Feature	Loading	Type
0	PC1	latitude	0.61	Best
1	PC2	battery_level	-0.48	Best
2	PC3	battery_temperature	-0.54	Best
3	PC4	date	0.49	Best
4	PC5	battery_temperature	-0.52	Best
5	PC6	acceleration	0.52	Best
6	PC7	magnetism	-0.74	Best
7	PC8	gsm_signal	0.66	Best
8	PC9	acceleration	0.52	Best
9	PC10	battery_level	-0.56	Best
10	PC1	user_id	0.4	Weak
11	PC4	sm_id	-0.48	Weak
12	PC9	sound	0.50	Weak
13	PC1	longitude	0.44	Weak
14	PC8	altitude	0.61	Weak

Table 13. Importance of Features using PCA for Smart City (Information sensitive)

	PC	Feature	Loading	Type
0	PC1	gravity	0.51	Best
1	PC2	network_type	0.49	Best
2	PC3	temperature	0.63	Best
3	PC4	pressure	0.49	Best

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

4	PC5	speed	0.81	Best
5	PC6	light	0.68	Best
6	PC7	app_version	0.62	Best
7	PC8	rotation	0.60	Best
8	PC9	temperature	-0.70	Best
9	PC9	humidity	0.69	Weak
10	PC1	operator	-0.48	Weak
11	PC7	device_model	-0.49	Weak

- ***K modes for data source clustering***

The purpose of clustering, a method for learning unsupervised, is to separate the population or data points into several groups so that the data points inside a group are more similar to each other than they are to the data points outside of that group. In essence, it is an assemblage of items chosen for their similarities and differences.

In data science, the KModes clustering technique is used to organize related data points into clusters according to their category characteristics. KModes operates by determining the modes, or most frequent values, within each cluster to establish its centroid, in contrast to conventional clustering methods that employ distance measures. KModes is the best tool for clustering categorical data, including survey results, market segmentation, or consumer profiles. It is an effective tool that helps scientists and data analysts understand their data and come to intelligent decisions.

The K-means cluster technique uses different measures, it is unable to cluster categorical data. The K-Mode cluster algorithms remove the numerical data limitation while maintaining the K-mean pattern as their foundation.

This K-Mode method removes the restriction imposed by the Kmeans following modification to enable the clustering of categorical data using the K-mean pattern:

- Hammering distance or simple match dissimilar evaluation are utilized for categorical data objects.
- Modify the cluster's mean using the modes.

$$d(x,y) = \sum_{i=1}^{f} \delta(X_j Y_j), \quad (6)$$

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

$d(x, y)$ assigns equal weight to each type of attribute. Let Z be a collection of categorical data objects, with A1, A2, ..., Am serving as their categorical attributes. Although the above is utilized since categorical data objects' dissimilarity determines the cost function,

$$C(Q) = \sum_{i=1}^{n} d(z_i Q_i), \quad (7)$$

where Q_i is near the cluster center and z_i is the element. Equation 4 defines the cost Function, which is minimized by the K-Modes technique.

The K-Modes, which comprise the following steps, operate under the assumption that the number of a probable group of data (K) is accessible.

1. Create K clusters by selecting data objects at random and designating K initial cluster centers, one for each cluster.
2. Using equation 4, assign the data object to the cluster whose cluster center is closest to it.
3. Calculate the K most recent modes for each cluster and update the K cluster based on the distribution of data objects.
4. Repeat steps 2 through 3 until the cluster relationship between the data objects has not changed.

- ***Consumer Application***

One of the most significant uses of consumer IoT technology is home security. To prevent unwanted entry, it enables homeowners to keep an eye on their front door. One well-liked product is the video doorbell, which notifies your computer, tablet, or phone whenever someone approaches the gate. Users can use emergency services and access control to keep an eye on their house in real time while they are away.

Numerous wearables for personal healthcare are available in the consumer electronics market thanks to IoT. For example, these devices are easily able to monitor vital signs like blood pressure, pulse rate, respiration rate, and body temperature. The gadgets are linked to the user's smartphone, and the relevant app collects the data. When medical assistance is needed, this can give medical professionals crucial information.

- ***Industrial Application***

Manufacturing facilities benefit greatly from the efficiency and convenience that IoT-based package management offers. Smart sensors can track every step of the packing process and provide real-time status updates. Vibrations, ambient factors like temperature and humidity, and feedback in the event of a problem during storage or transit can all be detected by embedded sensors.

- ***Commercial Application***

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

A smart home is the most obvious use of the Internet of Things. Sensors are used in a smart home's resource management, lighting, and security systems. A smart home is a more compact and self-contained form of a smart city.

A smart greenhouse doesn't rely on shifting weather patterns to grow crops; instead, it uses microclimate. All parameters are monitored and controlled by sensors, which also have automated water and light systems. Categorical data clustering is a significant research challenge in IoT. Clustering is a widely used method that partitioning objects into groups so that objects within the same group are more similar to each other than to those in other clusters [100]. K-Mode algorithm clusters the features based on the feature importance which are obtained from the PCA technique. The features are categorized as best and Weak Features Based on Time sensitivity and information sensitivity. The obtained features are tabulated in *Figures 7 to 9*.

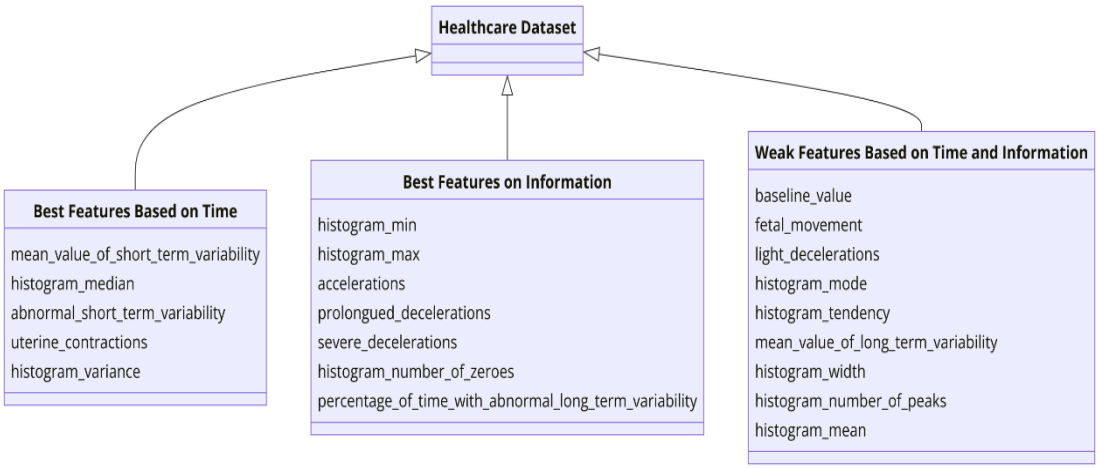

Figure 7. K-Mode Clustering of Features from Healthcare Dataset

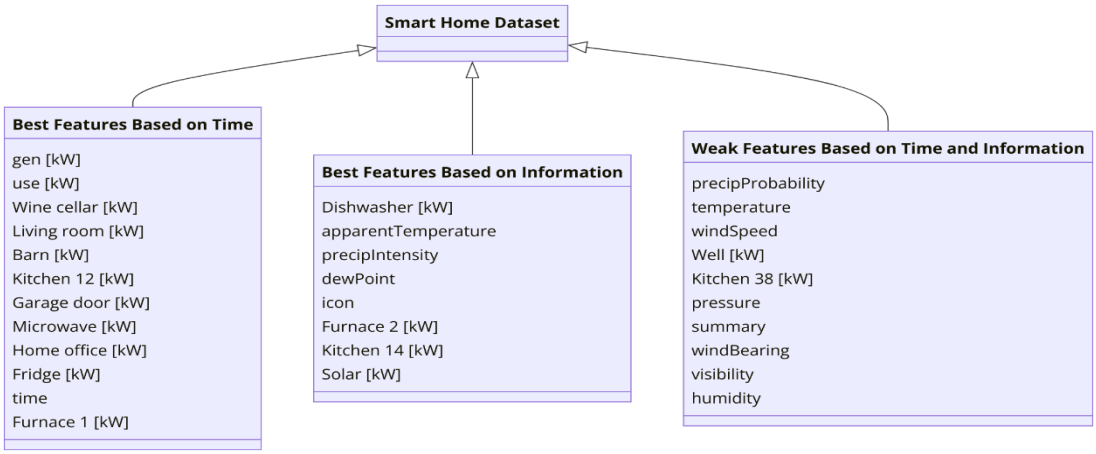

Figure 8. K-Mode Clustering of Features from Smart Home Dataset

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

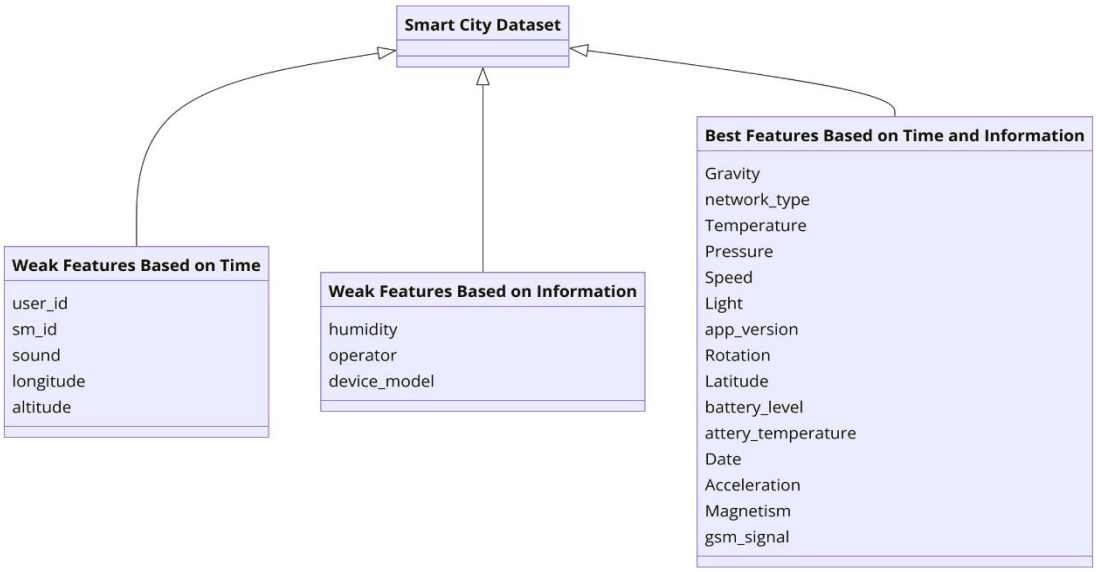

Figure 9. K-Mode Clustering of Features from Smart City Dataset

3. Results and Discussion:

This section discusses the results of the experimental analysis carried out in this research. Experimentation is carried out for all three smart applications.

3.1. Results for Healthcare Application

The classified features are categorized as follows:

1. ***Best features based on time***
2. ***Weak Features Based on Time and Information*** **and**
3. ***Best features based information***

The three categories obtained are considered as three new datasets for further processing. An overall 21 features are classified from the Healthcare dataset consisting of 2127 records. FDM is applied to group features into information and time sensitive features. By reducing variance, FDM facilitates PCA to extract significance of individual features within each data frame. These features are further divided into clusters that represent best and weak features by applying the K-Mode algorithm.

The results for Healthcare application using the proposed methodology are depicted in Figure 10.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

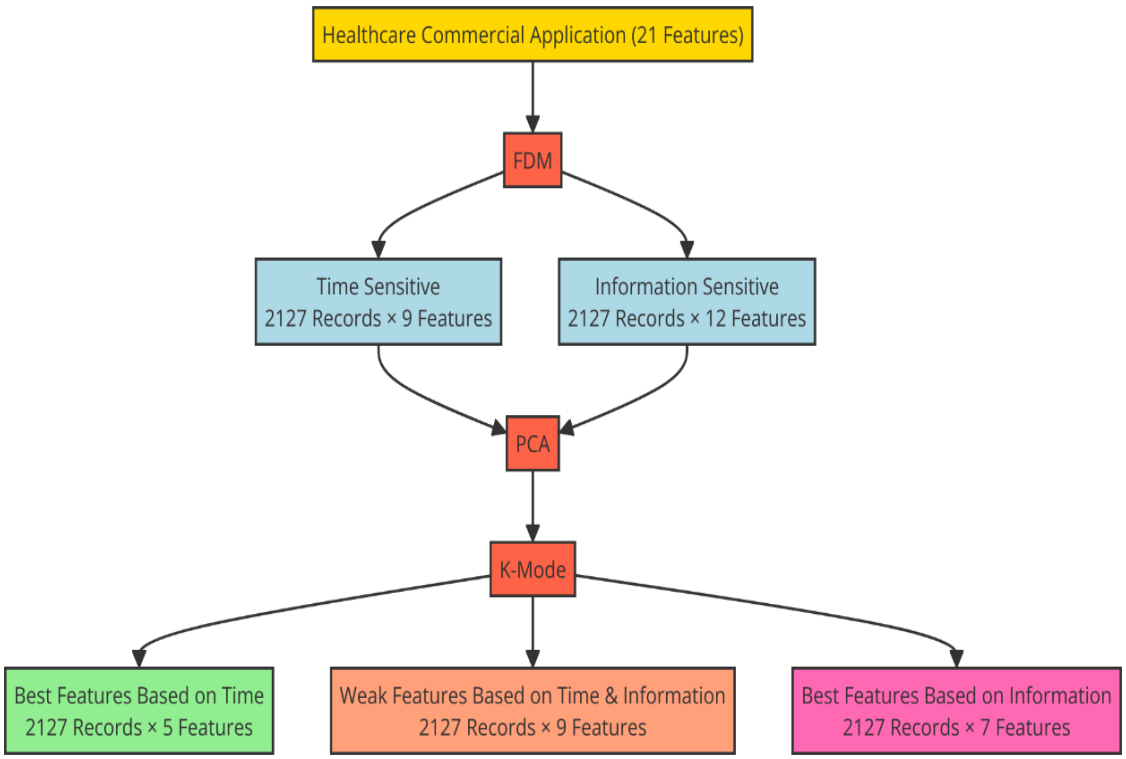

Figure 10. Healthcare Commercial Application

The K-Mode algorithm categorizes the 21 features into Best features based on time that contains 5 features, Weak Features Based on Time and Information containing 9 features, and Best features based information containing 7 features.

3.2. Results for Smart Home Application

The Smart Home dataset consists of 503911 records with 30 features. The PCA technique was applied to the data which reduced the variance by using FDM for clustering both time and information sensitivity data. The output of the PCA technique determined the importance of each feature. It can be observed from the results that both best and weak feature clusters were obtained using K-Mode clustering. The results for Smart Home application using the proposed methodology are depicted in Figure 11.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

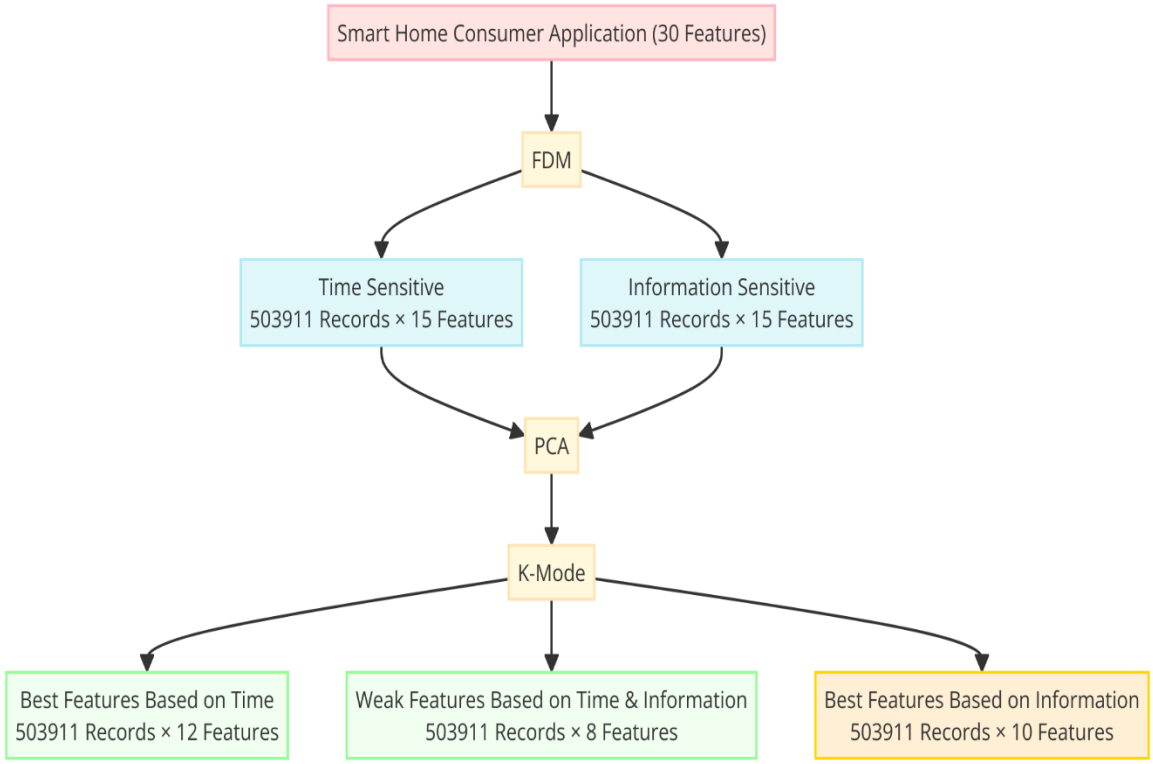

Figure 11. Smart Home Commercial Application

The K-Mode algorithm categorizes the 30 features into Best features based on time that contains 12 features, Weak Features Based on Time and Information containing 8 features, and Best features based information containing 10 features.

3.3 Results for Smart City Application

The Smart City dataset consists of 10000 records with 23 features. The PCA technique was applied to the data which reduced the variance by using FDM for clustering both time and information sensitivity data. The output of the PCA technique determined the importance of each feature. It can be observed from the results that both best and weak feature clusters were obtained using K-Mode clustering. The results for Smart City application using the proposed methodology are depicted in Figure 12.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

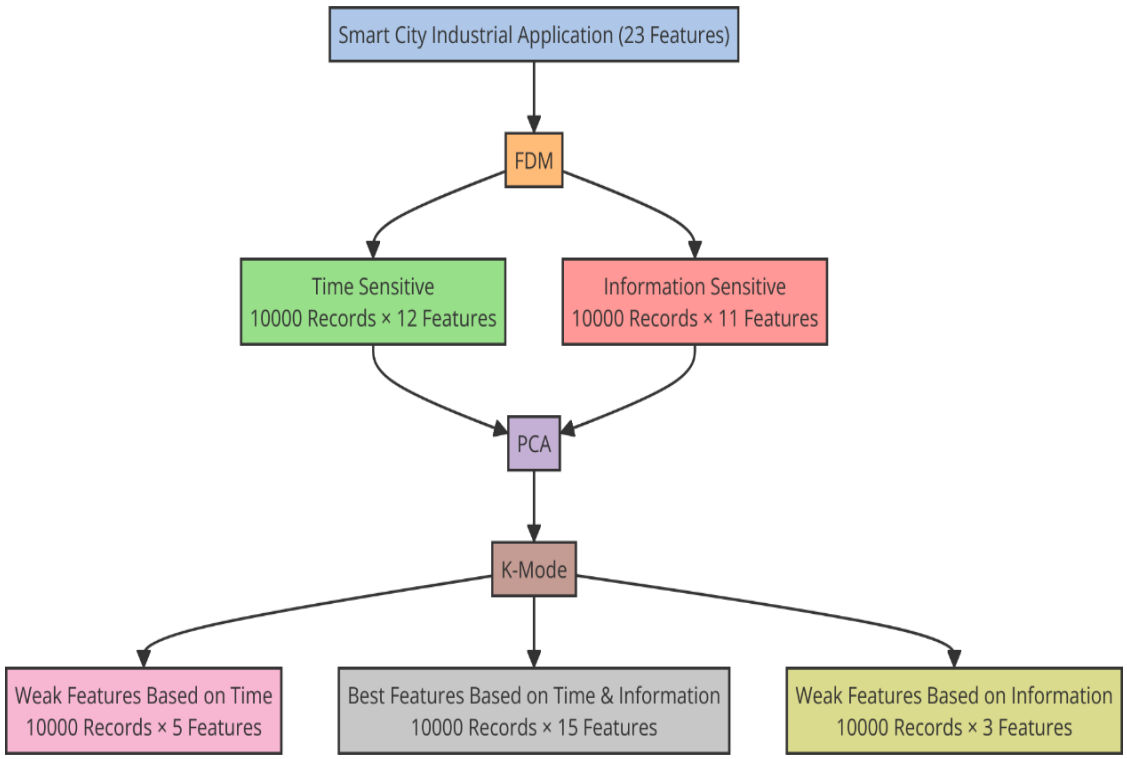

Figure 12. Smart City Industrial Application

The K-Mode algorithm categorizes the 23 features into Weak Features Based on Time that contains 5 features, Best Features Based on Time and Information containing 15 features, and weak features-based information containing 3 features.

4 Conclusion and Future Scope of Research:

4.1. Conclusion

The study successfully demonstrates the impact of dimensionality reduction on the clustering of IoT stream data. Module 1 of the proposed framework, helps in reducing data dimensions and selecting essential features that enhances the efficiency of IoT stream data processing. The FDM is effective in categorizing data into time-sensitive and information-sensitive groups, which was crucial for improving the clarity and utility of the data. PCA successfully identified critical features, enhancing the accuracy of K-Mode clustering, which in turn led to a better classification of features based on their sensitivity. The methodology in Module 1 was applied to various IoT smart applications, including Healthcare, Smart Homes, and Smart City, demonstrating its versatility and effectiveness across different domains. Overall, the study shows that dimensionality reduction techniques could significantly reduce data size, making it easier to manage and analyze IoT stream data while improving storage and communication costs.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

4.2. Future Scope of Research

Future research can build on these findings by exploring advanced dimensionality reduction techniques such as t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP), which may further enhance data processing efficiency. Implementing and testing these methodologies in real-time IoT systems can provide deeper insights into their practical applicability and performance, ensuring that the techniques can handle the dynamic nature of IoT stream data. Additionally, extending the study to a broader range of IoT applications, such as smart agriculture and industrial automation, can validate the robustness of the proposed methods. Investigating the integration of these clustering techniques with advanced machine learning models for predictive analysis and decision-making can offer more comprehensive solutions for IoT stream data management. Finally, future work should address the security and privacy aspects of handling sensitive IoT stream data, ensuring that data reduction and clustering methods do not compromise data integrity and confidentiality. Focusing on these areas, future research can further enhance the management and analysis of IoT stream data, leading to more efficient and effective smart services across various domains.

References

[1] Ejaz, W., & Ibnkahla, M. (2015, October). Machine-to-machine communications in cognitive cellular systems. In *2015 IEEE international conference on ubiquitous wireless broadband (ICUWB)* (pp. 1-5). IEEE.

[2] Ericsson, A. B. (2018). Internet of Things forecast. *URL: https://www. ericsson. Com/en/mobility-report/internet-of-things-forecast (visited on 02/06/2019)*.

[3] Stolpe, M. (2016). The internet of things: Opportunities and challenges for distributed data analysis. *Acm Sigkdd Explorations Newsletter*, *18*(1), 15-34.

[4] Yang, G., Tan, W., Jin, H., Zhao, T., & Tu, L. (2019). Review wearable sensing system for gait recognition. *Cluster Computing*, *22*, 3021-3029.

[5] Shi, X., Zheng, Z., Zhou, Y., Jin, H., He, L., Liu, B., & Hua, Q. S. (2018). Graph processing on GPUs: A survey. *ACM Computing Surveys (CSUR)*, *50*(6), 1-35.

[6] Lin, Y., Zhu, X., Zheng, Z., Dou, Z., & Zhou, R. (2019). The individual identification method of wireless device based on dimensionality reduction and machine learning. *The journal of supercomputing*, *75*(6), 3010-3027.

[7] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and information systems*, *34*, 483-519.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

[8] Jader, R. and Aminifar, S., 2022. Fast and Accurate Artificial Neural Network Model for Diabetes Recogni-tion. *NeuroQuantology*, 20(10), pp.2187-2196.

[9] Hira, Z. M., & Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, *2015*.

[10] Hoffman, J. D., & Frankel, S. (2018). Numerical methods for engineers and scientists. *CRC press*.

[11] Pajouh, H. H., Javidan, R., Khayami, R., Dehghantanha, A., & Choo, K. K. R. (2016). A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks. *IEEE Transactions on Emerging Topics in Computing*, *7*(2), 314-323.

[12] Zhang, T., & Yang, B. (2016, November). Big data dimension reduction using PCA. In *2016 IEEE international conference on smart cloud (SmartCloud)* (pp. 152-157). IEEE.

[13] Guo, K., Tang, Y., & Zhang, P. (2017). CSF: Crowdsourcing semantic fusion for heterogeneous media big data in the internet of things. *Information Fusion*, *37*, 77-85.

[14] Rehman, M. H., Chang, V., Batool, A., & Wah, T. Y. (2016). Big data reduction framework for value creation in sustainable enterprises. *International journal of information management*, *36*(6), 917-928.

[15] He, S., Shin, D. H., Zhang, J., Chen, J., & Sun, Y. (2015). Full-view area coverage in camera sensor networks: Dimension reduction and near-optimal solutions. *IEEE Transactions on Vehicular Technology*, *65*(9), 7448-7461.

[16] Papageorgiou, A., Cheng, B., & Kovacs, E. (2015, November). Real-time data reduction at the network edge of Internet-of-Things systems. In *2015 11th international conference on network and service management (CNSM)* (pp. 284-291). IEEE.

[17] Cheng, S., Cai, Z., Li, J., & Gao, H. (2016). Extracting kernel dataset from big sensory data in wireless sensor networks. *IEEE Transactions on Knowledge and Data Engineering*, *29*(4), 813-827.

[18] Greu, V. (2020). Using the information and communications technology data deluge from a semantic perspective of a dynamic challenge: what to learn and what to ignore?-part 3. *Romanian Distribution Committee Magazine*, 11(1), 16-29.

[19] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future generation computer systems*, *29*(7), 1645-1660.

Power System Technology
ISSN:1000-3673
Received: 16-04-2025 Revised: 05-05-2025 Accepted: 22-06-2025

[20] Jader, R.F., Aminifar, S. and Abd, M.H.M., 2022. Diabetes detection system by mixing supervised and unsupervised algorithms. *Journal of Studies in Science and Engineering*, 2(3), pp.52-65.

[21] Jader, R. and Aminifar, S., 2022. Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach. *Applied Computational Intelligence and Soft Computing*, 2022.

[22] Song, X. F., Zhang, Y., Gong, D. W., & Gao, X. Z. (2021). A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data. *IEEE Transactions on Cybernetics*, 52(9), 9573-9586.

[23] Haq, A. U., Zhang, D., Peng, H., & Rahman, S. U. (2019). Combining multiple feature-ranking techniques and clustering of variables for feature selection. *Ieee Access*, 7, 151482-151492.

[24] Jader, R.F., Abd, M.H.M. and Jumaa, I.H., 2022. Signal Modulation Recognition System Based on Different Signal Noise Rate Using Artificial Intelligent Approach. *Journal of Studies in Science and Engineering*, 2(4), pp.37-49.

[25] Jader, R. and Aminifar, S., 2023. An Intelligent Gestational Diabetes Mellitus Recognition System Using Machine Learning Algorithms. *Tikrit Journal of Pure Science*, 28(1), pp.82-88.

[26] Talabani, H. S., & Jumaa, I. H. (2024). A Review of Various Machine Learning Techniques and its Application on IoT and Cloud Computing. Tikrit Journal of Pure Science, 29(1), 185-195.

[27] Dai, J., & Xu, Q. (2013). Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing*, *13*(1), 211-221.

[28] G. H. Lee, J. Han, and J. K. Choi, "MPdist-based missing data imputation for supporting big data analyses in IoT-based applications," *Future Generation Computer Systems,* vol. 125, pp. 421-432, 2021.

[29] B. Agbo, H. Al-Aqrabi, R. Hill, and T. Alsboui, "Missing data imputation in the Internet of Things sensor networks," *Future Internet,* vol. 14, no. 5, p. 143, 2022.