Machine Learning and Device's Neighborhood-Enabled Fusion Algorithm for the Internet of Things

Amal Al-Rasheed, Tahani Alsaedi, Rahim Khan, Bharati Rathore, Gaurav Dhiman, Mahwish Kundi, and Aftab Ahmad

***Abstract*—In the Internet of Things, information fusion is among the crucial problems and probably occurs due to the dense deployment of consumer electronic devices. In the literature, various methodologies have been developed to fine-tune raw data; however, consumer electronic devices' neighborhood information has been completely ignored. In this manuscript, a machine learning and neighborhood-assisted fusion approach has been developed for consumer electronic devices to ensure that captured data values have been properly refined before onward processing at the respective edge. In this approach, every server accepts member request invitations from electronic devices deployed in its coverage area. It applies the well-known K-mean and supports vector machine (SVM) algorithms to refine captured data values by consumer electronic devices. Apart from that, the server module has the built-in intelligence to compare the captured data values of those electronic devices, which reside nearby and probably have a higher redundancy ratio. Simulation results have concluded that the proposed machine learning-assisted fusion approach is an ideal solution for the IoT in general and the Artificial Intelligent-enabled IoT in particular. Additionally, the proposed algorithm was thoroughly examined via various performance evaluation metrics such as lifetime, energy efficiency, and refinement ratio, where it has shown convincing results such as 30% improvement in the fusion ratio.**

***Index Terms*—Internet of Things, machine learning, K-mean, information fusion, outliers.**

I. Introduction

THE DEVELOPMENT of a smart and intelligent automated communication infrastructure across different application areas is subjected to the careful integration of both artificial intelligence and Internet of Things. In the IoT, modules or devices are deployed, either randomly or manually, in the closed proximity of the underlined phenomenon, that is required to be monitored, and capture crucial information after a defined time intervals [1], [2]. Additionally, every module is required to transmit its captured data via a shared communication channel to the respective server module in the IoT [3], [4]. These devices could become more smarter if artificial intelligence-enabled techniques have been incorporated along with the traditional programming modules. Moreover, artificial intelligence-enabled modules have strong and well-informed decision power than traditional programming based approaches [5], [6]. However, as artificial intelligence-enabled techniques are very complex and are far beyond the processing and storage capabilities of the respective member devices in the IoT. Therefore, these approaches are developed specifically for the respective server module, which has adequate processing and storage capacities, in the IoT [7], [8]. Although with embedded AI-enabled methodologies, fusion and duplicate data values are among the critical issues associated with the IoT, which are required to be resolved on priority basis.

Information fusion has been extensively studies in the literature and excessive number of the field proven approaches have been developed to get a refined version of the raw data set in the IoT and other resources constraint networking infrastructures. However, majority of the existing fusion techniques have assumed the duplicate-insensitive functions, which has a direct proportionality ratio to the respective refinement process of the concerned data, i.e., elimination of duplicate or noisy data values [9]. Furthermore, these techniques are highly susceptible to both noisy and duplicate (outliers) data values generated either by a malfunctioning module or through interference. Periodic clustered oriented methodology has been used to enhance transmission efficiency of the underlined channel along with minimizing overall redundancy ratio. A two-tiers approach is adopted where in first phase duplicate data values are reduced locally. In second tier, a k-mean based ANOVA model with embedded statistical tests is used for the elimination of duplicate data values [10]. However, a common issue with this approach is its complexity, which makes its implementation harder in application domains where devices have to rely on the onboard batteries. Likewise, a multilevel fusion technique, known as hydra, was proposed

Received 25 June 2024; revised 6 October 2024; accepted 13 November 2024. Date of publication 24 February 2025; date of current version 12 June 2025. This work was supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, under Project PNURSP2024R235. *(Corresponding author: Rahim Khan.)*

Amal Al-Rasheed is with the Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia.

Tahani Alsaedi is with the Applied College, Taibah University, Madinah 42353, Saudi Arabia.

Rahim Khan and Aftab Ahmad are with the Computer Science Department, Abdul Wali Khan University Mardan, Mardan 25000, Pakistan (e-mail: rahimkhan@awkum.edu.pk).

Bharati Rathore is with the School of Business and Creative Industries, University of South Wales, CF37 1DL Pontypridd, U.K.

Gaurav Dhiman is with the Department of Computer Science and Engineering, Yuan Ze University, Tao Yuan, Taiwan, and also with the Centre of Research Impact and Outcome, Chitkara University, Rajpura 140417, Punjab, India.

Mahwish Kundi is with the Maynooth International Engineering College, Maynooth University, Maynooth, W23 F2H6 Ireland.

Digital Object Identifier 10.1109/TCE.2024.3500024

to enhance accuracy and precision levels of the underlined sensor modules in the agriculture domain. Additionally, this model has enhances accuracy and precision ratio of various decision taken by the concerned module [11]. Dempster-Shafer theory and adaptive weight based fusion methodology has been developed for the IoT where basic probability assignment was used for the representation of uncertain information and to further quantify similarity ratio of the captured data set [12]. As wearable sensors attached to the patient's body are likely to produce heart's vibration signals, which are required to be refined with effective fusion methodologies. For this purpose, Bayesian-enabled fusion with the embedded three interval estimators are applied to the channel along with a thorough investigation of the spectrum [13]. However, this approach has the highest vulnerability ratio to the local fusion, which is quiet common in the IoT or other resources limited domains. Additionally, cloud supported algorithm was developed to make sure that appropriate device, i.e., relay module, has been selected for the sampling and forwarding of the captured data values preferably in the refined form. To ensure this, two different parameters, that is quality of the link and temporal correlation, have been embedded in the proposed algorithm to make it energy efficient. In this approach, limited number of devices are in active state and message passing algorithm has been adopted to construct missing information, which is likely to occur as majority of the deployed devices are not active [14]. Although, these approaches have resolved the crucial issue of the information fusion either locally or globally, however, these mechanisms have neglected a important aspect of the fusion, that is device's neighborhood, which is cornerstone of the duplicate data values especially in the IoT. Therefore, fusion algorithms should consider device's neighborhood information while carrying out the data refinement process in the IoT.

In this paper, an energy and performance efficient fusion methodology has been developed to refine captured data values especially from those devices which are deployed in closed neighborhood and the similarity ratio is likely to be high. For this purpose, initially, a sophisticated neighborhood discovery process is presented to make sure that server module has accurately separated neighboring devices from other modules in the IoT. To realize this activity, media access control (MAC) address and geographic information is utilized by the respective server module. As soon as neighboring devices are identified, a k-mean clustering and Support Vector Machine (SVM) scheme based fusion methodology is used to refine captured data values preferably with the minimum possible duplicates and outliers (noise) ratio. The main contributions of this paper as given below.

1) Development of geographic information-enabled methodology to accurately separate neighboring devices from the non-neighbors;
2) SVM and K-mean Clustering methodologies are combined with the proposed fusion approach to form a sophisticated refinement scheme for IoT;
3) The proposed fusion mechanism is developed such that it is energy and performance efficient, both are very crucial for the traditional networks in general and IoT in particular.
4) Finally, the proposed fusion approach has ensured that un-necessary comparisons of data values, particularly those captured by the devices deployed far away from each other, are avoided.

The remaining paper is organized as follows: In the following section [Section II], a generalized overview of the proposed approach has been presented. In Section III, proposed information fusion algorithm has been discussed in detail. In Section IV, simulation results of the proposed and existing fusion methodologies has been presented. Finally, concluding remarks and future directions are given in Section V.

II. Proposed K-Mean Clustering and Neighborhood-Enabled Fusion Approach

In this section, the proposed neighborhood-enabled fusion approach is described in detail along with its neighbors discovery process in the IoT. Initially, every server module has the builtin capacity to pursue the proposed fusion process by thoroughly examining data sets captured by different devices in the IoT. The proposed fusion approach has enabled servers to focus on those data sets where ratio of the duplicate data values is high, i.e., data sets captured by devices deployed in closed vicinity or neighborhood. Therefore, prior to the fusion process, server modules are required to carry out the neighbors discovery process where devices reside in the respective coverage area are marked. In the next phase, fusion activity is limited to the mark devices only in the IoT. A detailed description of the various phases in the proposed scheme is given in the following subsections.

A. Internet of Things Model

Hierarchical networking infrastructure has been used for the proposed setup where every device is required to be member of the nearest possible server module, preferably with highest Received Signal Strength Indicator (RSSI) Value and should communicate its captured data via this server. Moreover, the proposed approach is equally applicable for both random and engineered deployment strategies as well as balanced and unbalanced clustered approaches in the IoT. The proposed model uses Ad-hoc On-demand Distance Vector (AODV) along with user data-gram protocol (UDP) to make sure that packets are successfully transmitted from source, i.e., device, to destination, i.e., server, module even if multiple devices are interested to communicate simultaneously. Apart from it, every device is equipped with omni directional antennas with average power consumption on the packets transmission and receiving is 0.86 and 0.5 watt respectively. Finally, every device in the IoT is assumed to relay on its on-board battery along with a fix packet size that is 512 bytes.

B. Generalized Overview of the Proposed Neighborhood-Enabled Fusion Approach in the IoT

In the proposed setup, every device captures data simultaneously after a predefined time interval, that is 60 seconds, and transmit it to the respective server module via a shared communication channel where it is thoroughly examined for

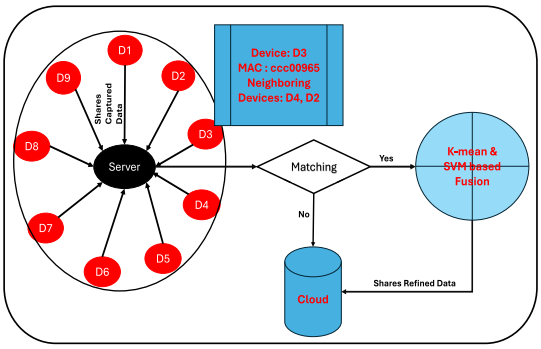

Fig. 1. Proposed K-Mean Clustering, SVM, and Neighborhood-enabled Fusion Approach.

both outliers and duplicate data values. As every server module has to serve as relaying point for the multiple devices and, thus, it has the required potentials to refine these captured readings through a sophisticated and trusted procedure in the IoT. The proposed approach is different from the existing state of the art approaches in a sense that it bounds every server module to find similarity indexes in those data sets which are captured by devices deployed in closed neighborhood as it is high likely that their captured readings will be the same. Generally in IoT and other resources constraint networks, devices are deployed randomly without considering probability of other device(s) in the coverage area or closed neighborhood. For example, deployment of devices with embedded flame or smoke sensors in a building, if an engineered approach is used, then it is high likely that every device will be deployed properly. However, if these devices are deployed using a random approach, then most probably multiple devices will be deployed at the point, which will lead to capturing of duplicate data values. For this purpose, every server module has gathered sufficient information about its member devices, that extends its decision power in separating active devices into group of matching and non-matching. A device belong to matching class mean that its capture reading will be compared against every member of the match class whereas non-matching are neglected in the refinement process. Thus, the proposed neighborhood-enabled fusion approach not only refine captured data values, it equally avoid a considerable amount of matching, i.e., refinement of those data sets which are approximately different from each other. For instance, if we assume that a particular server module receives captured data from member devices, that is nine (09) in this case as shown in Figure 1. From this Figure 1, it is evident that device-1 (D1) has two immediate closed neighbors, that is device D2 and D9 respectively. Generally, data values captured by device-1 have a highest similarity ratio with the readings captured by devices D2 and D9 respectively in the IoT. However, the correlation or similarity ratio of device-1 (D1) with other modules, that is D3, D4, D5, D6, D7, and D8, is extremely less and can be ignored. Therefore, the proposed neighborhood-enabled fusion approach has to make sure that similarity indexes could be checked among those devices which are deployed in the closed neighborhood such as device-2 (D2) data set is compared with D1 and D3 respectively.

C. Neighborhood Discovery Phase: Server Module

The proposed data fusion algorithm is primarily based on the effective utilization of the member device's neighborhood information. Neighborhood information plays a crucial role in the overall refinement process, particularly identification of the duplicate data values, of those data sets, which are collected through neighboring devices in the IoT. Moreover, as the fusion process is required to be carried out by the respective server module, therefore, every server should take appropriate measure to collect it with minimum possible efforts. For this purpose, every server module has a builtin process for the collection of neighborhood information, that is triggered as soon as the concerned IoT network become operational. A simplified message is generated by every server module and broadcast it. Every module resides in the coverage area of the particular server receives this message and update it according to local information, that is number of immediate neighbors. Additionally, it is important to note that every member module sets a back-off timer before transmission of the updated packet to ensure that its data is transmitted in collision free manner even if multiple neighboring devices are interested to send their message to the same server in the IoT. The respective back-off timer used in the proposed setup is give in equation (1).

$$Back - off\ \ Timer = Random(100 - 1500)\ ms \qquad (1)$$

where random function is used to minimize the expected ratio of collision's probability in the IoT, that is likely to occur if multiple devices, particularly those deployed in closed neighborhood, send their message to the server via shared communication medium. Let's assume that if devices D5 and D6 trigger transmission activity simultaneously, then surely collision among packet is guaranteed. However, if devices set a back-off time and waits for its expiry to transmit, then probably collision will not occur.

In the proposed communication model for the IoT, every member device is assumed to serve as a source, that is capturing of the respective data through an appropriate embedded sensor module. These modules must share their captured data with a local server, preferably that is deployed in direct communication range of the respective transceiver's module. Additionally, data values are transmitted in raw form to the server by the respective devices as their computational and storage capabilities are limited, a generalized scenario in the IoT. As server modules are likely to have the computational and storage capacities, especially those required to pursue the refinement process, removal of the duplicate data values, in the IoT. Therefore, data fusion activity is reserved for the server modules only. Secondly, as decisions are carried out the respective edge module, thus, refinement at the server level is very fruitful as duplicate data values are eliminated before sharing it with the centralized edge. In the proposed fusion process, the respective server module is smart enough to refine data set of the selected devices only, that is those deployed in closed proximity. If server module has data set of two member devices say "X and Y", then refinement process is triggered only if these devices are direct neighbors

as depicted in Figure 1. If X and Y represent D1 and D2, then surely refinement process is mandatory. However, if devices X and Y represent D1 and D7, then aggregation is not required as it is likely that data sets collected by these devices are duplicate free. Additionally, it is possible that data sets captured by those devices may contain duplicate data, but it can be neglected as it ratio is very less. Thus, the refinement process is completed in this way by every server module in the IoT and are now in position to share these refined data set with a centralized edge. Every server module transmit the refined data set to the edge directly if applicable. However, multi-hop communication methodology is adopted if any of the server module do not have the capacity to transmit directly in the IoT. Secondly, during the fusion process, two data values, particularly those captured by different devices, should be considered as equal if their distance is within the specified bound, i.e., threshold value. For the computation of distance, Euclidean distance measure is used as depicted in Equation (2).

$$D(C_i, C_j) = \sqrt{(x_i - y_j)^2} \tag{2}$$

where C_i and C_j represent different neighboring devices have data value x_i and y_j respectively. If distance between two values falls within the defined bounds, that is 0.15 in this case, then one of these values is retained whereas second is discarded. This process is repeatedly applied by every server module in the IoT to get the required refined version of data. Moreover, K-mean clustering algorithm is used to form a balanced hierarchical deployed of IoT's devices where server module serves as a centralized interaction point, that is every member module shares the captured data with it. Generally in the resources limited networks, it is not always possible to get balanced clustering, a methodology where every cluster has equal number of member devices. To address this, the proposed fusion approach has builtin support for both balanced and unbalanced clustering approaches in the IoT. Additionally, then various machine learning models have been used to enhance effectiveness of the proposed scheme particularly in terms of accuracy and precision ratios.

III. Proposed Device's Neighborhood-Based Fusion Approach for the Internet of Things

In this section, a detailed description of the proposed fusion methodology along with how a member device gathers information about its closed neighbors (Neighbors discovery), especially those where correlation among captured data sets are very high. As device's neighborhood information forms basis of the proposed fusion methodology, therefore, neighborhood discovery process should be carried out in a systematic order. Furthermore, the proposed neighborhood-enabled fusion algorithm is suitable for the server modules only as ordinary devices are limited in terms of their computational and storage powers. If we assume that $C_1, C_2, C_3, \ldots, C_n$ and $S_1, S_2, S_3, \ldots\ldots, m$ are used to represent ordinary devices and server modules respectively in the IoT where n & m represent total numbers, then according to the proposed approach, K-mean clustering must ensure that both S_1 & S_2 have equal number of member devices in the IoT. Secondly, every device should become member of the nearest server module, that is ensured through RSSI value. Initially, every server module generates a message by setting hop-count value to zero and broadcast it that is likely to be received by those devices deployed in the coverage area. Secondly, a timer is set, in which it expects replies from devices reside in its closed neighborhood or coverage area. Every device updates message credentials according to its neighborhood hop-count information. This updated packet is transmitted to the respective server preferably before expiry of the defined time interval. Secondly, if a device receives a message where hop-count value is one (01), then it is discarded as in proposed setup, member device must be in the direct coverage area of at least one server module. Thirdly, if a device has received two message with zero hop count value, then it will send a response message to that server, which has maximum value of the RSSI. For example, a message sent by a server module of cluster-1 is most probably received by devices belong to that cluster only. However, it is possible that this message can be received by those devices belong to other clusters which are deployed on the boundary area of the entire IoT. Therefore, those devices should become member of that server which has the maximum possible value of RSSI. After completion of this process, every device has become a member of the nearest server module in the IoT. For example, messages sent, preferably through broadcast, by server module in Figure 1 is received by D1, D2,..D9. These modules will update the message contents and resend it to the server. However, in addition to the server, this updated message is received by neighboring devices, those reside in the respective coverage area. Message sent by D2 is received by the server as well as D1 and D9, due to closed neighborhood. These module will discard such messages as its hop-count value is greater than zero (0). Moreover, it is possible that every server module will not be able to communicate directly with the concerned edge. In this case, a multiple hope communication methodology is adopted where another server module, particularly that resided in the coverage area, acts as a relay module. Thus, it has additional responsibility of packets forwarding along with its own dedicated task, that is collection and refinement of data collected by the member devices in the IoT.

A. Proposed Information Fusion Algorithm at the Respective Server in the Internet of Things

In the proposed fusion scheme, every server module receives data sets captured by the member devices in the IoT. These data sets are required to be refined, i.e., duplicate and noise free, through a sophisticated fusion process before sharing. For this purpose, various distance measures (such as K-mean clustering, Euclidean, SVM, and Decision Tree) have been used to find the best possible refinement scheme, that is not only effective, but should be energy efficient as well. Generally in the IoT, a sever module passes every data sets, those captured by its member devices without utilizing neighborhood information, through the defined fusion process, which is time consuming as non-correlated data sets are also matched with each other.

In proposed fusion process, neighborhood information plays a crucial role in differentiating data sets into two groups. (i) Correlated and (ii) not Correlated. The former data sets are captured by those devices, which reside in the closed neighborhood and thus, their collected data sets have the highest possible ratio of the similarity indexes. These data sets are the potential candidates require to be passed through the respective fusion process. Non-correlated data sets are captured values of those devices, which are deployed far-away from each other. Therefore, these data sets are excluded from the fusion process as the expected similarity ratio among these data sets is very low and can be neglected. Excluding these non-correlated data sets from the refinement process is very effective, especially in terms of power and other resources consumption.

For finding similarity indexes, a well-known distance measure, i.e., Euclidean, is used with an embedded threshold value, that is 0.15 in this case. Thus, if two values are matched with the defined distance measure and the computed difference is less than or equal to the allowable threshold, then these readings are assumed be to similar. Thus, one of these readings is stored whereas other copies are deleted. Secondly, the proposed fusion approach is smart enough to separate accurate data values from the outliers or noisy, those generate either through interference or malfunctioning of the embedded sensor module in the particular device. In case of false readings, every outlier is replaced with the average value, that is computed using equation (3).

$$Val_i = Avg(Val_{i-1}, Val_{i+1}) \tag{3}$$

However, as Euclidean distance measure is highly vulnerable to noise or outliers, therefore, the proposed scheme has adopted a two-levels fusion methodology, that is initially refine data with the Euclidean distance measure by the respective device and then with other machine learning algorithms such as SVM and K-mean at the server module in the IoT. Generally, machine learning algorithms are computationally expensive and power hungry, therefore, the refinement process, especially among data sets captured through different devices is carried out at the server module. Initially, K-mean clustering algorithm partitioned these data sets into clusters where every cluster consists of highly correlated data values. As K-mean clustering is based on the centriod conception, therefore, every value is thoroughly evaluated to confirm its feasibility to respective centriod. In the next step, K centriod are required to be re-computed such that datasets are accurately partitioned into cluster where each clustered has only those values which falls within the defined bounds. This re-computation of k centriod is repeated until non of these centriods have significant changes in the resultant clusters formation. After successful completion of this step, server module initiates the respective refinement process such that both outliers and duplicate data values are identified and make sure that exactly one copy of the duplicate data values is retained or stored whereas other are permanently deleted. As soon as the refinement process is complete, then refined datasets are sent to the respective edge module for the onward processing.

As Support vector machine (SVM) is primarily used for the classification or regression related problem in different domain, however, in this paper, it is used for the classification purpose only, i.e., separate datasets with duplicates and duplicates-free. To classify datasets as duplicate-oriented (Positive) and duplicate free (Negative), the following Equations (4), (5), and (6) are used.

$$\overrightarrow{X}.\overrightarrow{w} = c \tag{4}$$

$$\overrightarrow{X}.\overrightarrow{w} > c \tag{5}$$

$$\overrightarrow{X}.\overrightarrow{w} < c \tag{6}$$

where X and w represent vectors such that c lies on the decision tree boundary, positive (duplicate in this case), and negative sample respectively. For accurate classification, decision rules are required to be defined such as those given in Equation (7).

$$\overrightarrow{X}.\overrightarrow{w} - c >= 0 \tag{7}$$

If we assumed that value of -c equal to + b in the revised version of Equation (7), then it becomes as given below in Equation (8)

$$\overrightarrow{X}.\overrightarrow{w} + b >= 0 \tag{8}$$

and hence, we have the finalized version of the equation as given in Equation (9) where positive (duplicates) and negative are separated.

$$y = \begin{cases} 1\ +1, \text{if } \overrightarrow{X}.\overrightarrow{w} + \text{b¿} = 0 \\ 2\ -1, \text{if } \overrightarrow{X}.\overrightarrow{w} + \text{b¡0} \end{cases} \tag{9}$$

For simplicity, we have assumed the following Equation (10).

$$f(x) \ = \sigma * c + MAC(C_i) \tag{10}$$

where C_i and σ represent the respective device and threshold values respectively. Furthermore, we have assumed that X and w are two vectors, then these are represented by Equations (11) and (12), where separation of data values captured through the deployed devices is classified as given below.

$$\overrightarrow{X} = \{f(x) \leq 0\} \tag{11}$$

$$\overrightarrow{w} = \{f(x) > 0\} \tag{12}$$

Let us further assume that we have five (05) support vectors, i.e., V_1,V_2, V_3, V_4, and V_5 respectively. By careful application of the integration process on the respective coordinates of devices and server modules in IoT, the following equality is formed, which is depicted in equation (13).

$$V_1 = \begin{pmatrix} A_1 \\ B_1 \end{pmatrix}, V_2 = \begin{pmatrix} A_2 \\ B_2 \end{pmatrix}, V_3 = \begin{pmatrix} A_3 \\ B_3 \end{pmatrix} V_4 = \begin{pmatrix} A_4 \\ B_4 \end{pmatrix} V_5 = \begin{pmatrix} A_5 \\ B_5 \end{pmatrix} \tag{13}$$

Furthermore, if the bias value is assumed as equal to one (01), then equation (13) becomes as given below in equation (14)

$$V_1 = \begin{pmatrix} A_1 \\ B_1 \\ 1 \end{pmatrix}, V_2 = \begin{pmatrix} A_2 \\ B_2 \\ 1 \end{pmatrix}, V_3 = \begin{pmatrix} A_3 \\ B_3 \\ 1 \end{pmatrix} V_4 = \begin{pmatrix} A_4 \\ B_4 \\ 1 \end{pmatrix} V_5 = \begin{pmatrix} A_5 \\ B_5 \\ 1 \end{pmatrix} \tag{14}$$

As we have assumed three support vector, therefore, at least three parameters are required for these in the realistic environment of the IoT, which are given below in Equations (15), (16), and (17) respectively.

$$\alpha_1\left[\overrightarrow{V_1}.\overrightarrow{V_1}\right]+\alpha_2\overrightarrow{V_2}.\overrightarrow{V_1}+\alpha_3\overrightarrow{V_3}.\overrightarrow{V_1}+\alpha_4\overrightarrow{V_4}.\overrightarrow{V_1}+\alpha_5\overrightarrow{V_5}.\overrightarrow{V_1}=1 \quad (15)$$

$$\alpha_1\overrightarrow{V_1}.\overrightarrow{V_2}+\alpha_2\overrightarrow{V_2}.\overrightarrow{V_2}+\alpha_3\overrightarrow{V_3}.\overrightarrow{V_2}+\alpha_4\overrightarrow{V_4}.\overrightarrow{V_2}+\alpha_5\overrightarrow{V_5}.\overrightarrow{V_2}=1 \quad (16)$$

$$\alpha_1\overrightarrow{V_1}.\overrightarrow{V_3}+\alpha_2\overrightarrow{V_2}.\overrightarrow{V_3}+\alpha_3\overrightarrow{V_3}.\overrightarrow{V_3}+\alpha_4\overrightarrow{V_4}.\overrightarrow{V_3}+\alpha_5\overrightarrow{V_5}.\overrightarrow{V_3}=1 \quad (17)$$

After applying usual mathematics, the following values for vector $\overrightarrow{V_1}$, $\overrightarrow{V_2}$, $\overrightarrow{V_3}$, $\overrightarrow{V_4}$, and $\overrightarrow{V_5}$ are computed, which are given in equation (18), (19), (20), and (21) respectively.

$$\overrightarrow{V_1}=\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}\overrightarrow{V_2}=\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}\overrightarrow{V_3}=\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}\overrightarrow{V_4}=\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}\overrightarrow{V_5}=\begin{pmatrix}A_5\\B_5\\1\end{pmatrix} \quad (18)$$

$$\alpha_1\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}+\alpha_2\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}+\alpha_3\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}+\alpha_4\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}+\alpha_5\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}=1 \quad (19)$$

$$\alpha_1\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}.\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}+\alpha_2\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}+\alpha_3\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}+\alpha_4\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}+\alpha_5\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}=1 \quad (20)$$

$$\alpha_1\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}+\alpha_2\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}+\alpha_3\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}+\alpha_4\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}+\alpha_5\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}=1 \quad (21)$$

$$\alpha_1\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}+\alpha_2\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}+\alpha_3\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}+\alpha_4\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}+\alpha_5\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}=1 \quad (22)$$

$$\alpha_1\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}+\alpha_2\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}+\alpha_3\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}+\alpha_4\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}+\alpha_5\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}=1 \quad (23)$$

To find approximate values for the α_1, α_2, α_4, α_4, and α_5, the aforementioned inequalities have been simplified through usual approaches. These values play a significant role in the actual separation process classes, and utilized in the given equation (24)

$$\overrightarrow{f}=\sum_{i=1}^{5}\alpha_i\overrightarrow{V_i} \quad (24)$$

Algorithm 1 Classification of Member Devices Into Two Disjoint Groups

Require: Classification of Member Devices in the Internet of Things

Ensure: Group-I (Fusion is Skipped) OR Group-II (Fusion is Mandatory)

for every Server $S_j \leftarrow 0$ to m **do**
  Generate $Msg_{Neighbor\ Discovery(ND)}$
  set *Neighboring Devices List* $\leftarrow 0$
  broadcast Msg_{ND}
for $C_i \leftarrow 0$ to n **do**
  if (ND != 0) **then**
  set *Neighboring Devices List* $\leftarrow$ k
  broadcast Msg_{ND}
  else then
  wait for T_b to expire
  end if
end for
S_i Separates Devices into Group-I and Group-II based on ND.
  return Group-I and Group-II
end for

To compute the threshold value for the separation of values in to disjoint classes, the following equation (25) is used, that is obtained through a comprehensive simplification process (addition in this case) of the aforementioned equations.

$$\alpha_1\begin{pmatrix}A_1\\B_1\\1\end{pmatrix}+\alpha_2\begin{pmatrix}A_2\\B_2\\1\end{pmatrix}+\alpha_3\begin{pmatrix}A_3\\B_3\\1\end{pmatrix}+\alpha_4\begin{pmatrix}A_4\\B_4\\1\end{pmatrix}+\alpha_5\begin{pmatrix}A_5\\B_5\\1\end{pmatrix}=\begin{pmatrix}\alpha_1A_1+\alpha_2A_2+\alpha_3A_3+\alpha_4A_4+\alpha_5A_5\\\alpha_1B_1+\alpha_2B_2+\alpha_3B_3+\alpha_4B_4+\alpha_5B_5\\\alpha_1+\alpha_2+\alpha_3+\alpha_4+\alpha_5\end{pmatrix} \quad (25)$$

Algorithm for the Classification of member devices into two disjoint groups is presented in Algorithm 1. Additionally, the proposed machine learning-enabled fusion is presented below as Algorithm 2, wherel every device, i.e., both ordinary and server modules, must find detailed information about their vicinity through a sophisticated neighbor discovery process. Furthermore, a smart machine learning-enabled fusion methodology is proposed where server modules classify member devices into to disjoint groups, i.e., (i) group-I consists of devices where refinement is necessary and (ii) group-II where fusion process is skipped.

IV. Simulation Results and Discussion

In this section, a detailed and comprehensive analysis of the proposed and existing schemes performance in term of various evaluation metrics is presented particularly with reference to the average packet delivery ratio, lifetime, throughput etc. For this purpose, these schemes were implemented in NS-2 simulation environment by providing completely similar topological infrastructures. Furthermore, performance of the proposed scheme is compared with existing state-of-the-art

Algorithm 2 A Machine Learning-Enabled Smart Algorithm for the Refinement of Data in the Internet of Things

Require: Raw Data Captured through Electronic Devices
Ensure: Refined Data (Outliers and Duplicate Free)

```
Group − I
Group − II
Sampling Interval (SI) ← 30 seconds
Adversary ← Zero
N ← Module in IoT
while SI Expires do
    Call Algorithm 1
    if C_i ∈ Group − I then
        Fusion is Not Required.
        Broadcast Data in Current Form.
    elseif C_i ∈ Group − II then
        Initiate Fusion Process
        Fused Data through SVM or K-Mean
        Broadcast Refined Data
    end if
end while
return Refined Data
```

approach such as [15], [16], [17], [18]. Secondly, as the proposed scheme has adopted a smart machine learning-enabled approach to distinguish neighboring devices from non-neighbors, which play a significant role in avoiding refinement of datasets with minimum possible duplicate ratio.

A. Fusion of Data at the Respective Server Module

Every server module has a group of connected ordinary devices, which are deployed in its coverage area and are able to communicate directly. The proposed approach enabled these servers to carried out the respective fusion process on the datasets of those devices, which are deployed in closed neighborhood and most probably their captured data has a slightly higher ratio of duplicates. The proposed approach is not only helpful in saving the resources by reducing a slight portion of the fusion process, it also performs exceptionally well than its counterpart schemes as shown in Figure 2. We have noted that when K-mean and SVM is embedded proposed smart and neighborhood-enabled fusion scheme, then its refinement ratio of the captured data values is up to 66.23% and 45.28% respectively. Other machine learning approaches like IDK, Euclidean & cosine distances and proposed schemes are 69.56%, 56.24%, 57.56% and 70.23% respectively. Moreover, duplicates elimination ratio of the proposed machine learning and neighborhood-enabled scheme is approximately 98.4%.

B. Sever Module's Energy Efficiency Through the Proposed Approach

Generally in IoT and other resource limited networks, every device has to rely on its onboard battery and, thus, effective utilization of these batteries is very significant for a prolonged lifetime of the underlined IoT. Therefore, the performance of the proposed scheme in terms of energy efficiency has been thoroughly examined both at the device and sever levels.

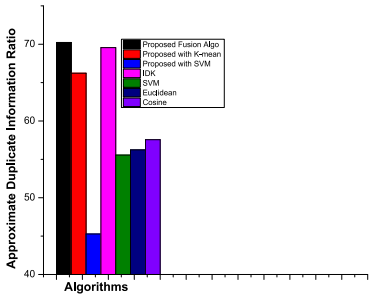

Fig. 2. Fusion Ratio of the Proposed and Existing Algo.

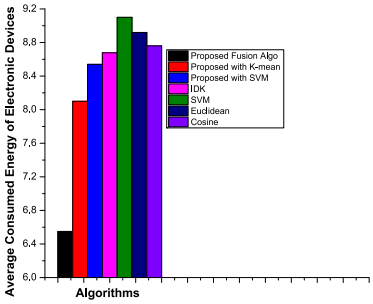

Fig. 3. Energy consumption of the Proposed and Existing Algo.

Moreover, as server module has to carry out the refinement process after a defined interval of time and its prolonged lifetime is very important for the member devices in IoT. Simulation results show that proposed scheme has consumed relatively less power than existing state of the art approaches while carrying out the respective fusion process. Furthermore, when proposed scheme is not embedded with the k-means and SVM algorithms, it has consumed 6.55 J of the available power. However, when it is embedded with K-mean and SVM, then its consumption ratio is approximately 8.1 & 8.54 J respectively as shown in Figure 3. However, when the proposed scheme is embedded with IDK, SVM, cosine and Euclidean distances, its consumption ratios are 8.6789 J, 9.1 J, 8.92 J, and 8.76 J respectively. Apart from it, the proposed and k-mean schemes are integrated to form a hybrid, then its performance is exceptionally well as shown in Figure 3. The proposed approach has minimum possible energy consumption when it is integrated with the k-mean, i.e., 5.15%, whereas with other schemes such as IDK, SVM, Euclidean, and Cosine distances are 6.77%, 10.98%, 7.53%, and 9.19% respectively.

C. End-to-End Delay in the Internet of Things

The approximate end-to-end delay ratio of the proposed approach is 1.532 ms, which is less than existing schemes as shown in Figure 4. Secondly, when the proposed smart fusion scheme is integration with k-mean and SVM, then its end-to-end delay ratio is 2.92 ms and 1.87 ms respectively. However, the approximate end-to-end delay ratios of the SVM, IDK, cosine, and Euclidean distances are 2.11 ms, 3.41 ms, 1.96, and 2.96 ms respectively. Likewise, the approximate ratio of end to end delay metric is slightly reduced by 45%, when the proposed scheme is integrated with the K-mean approach. Similarly, for SVM, IDK, cosine, and Euclidean distances,

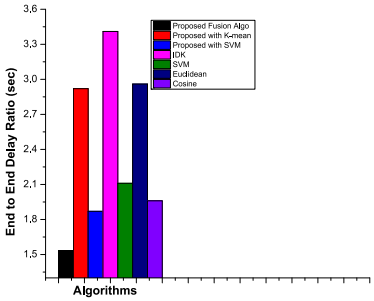

Fig. 4. End-to-End delay of the Proposed and Existing Algo.

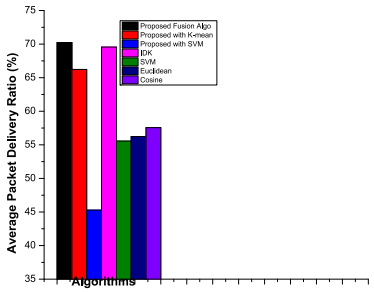

Fig. 5. APDR of the Proposed and Existing Algorithms.

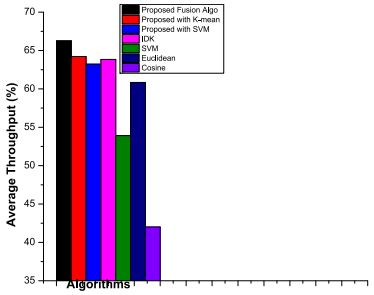

Fig. 6. Throughput of the Proposed and Existing Algorithms.

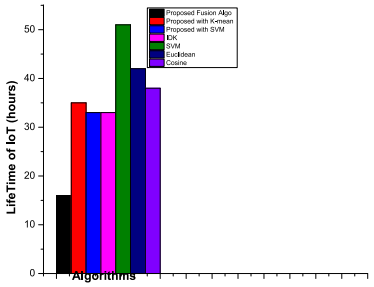

Fig. 7. Dead and Active Device's Ratio.

its reduction ratio is approximately 35%, 37%, 12%, and 6% respectively.

D. Average Packet Delivery Ratio (APDR)

During the simulation setup, it is observed that the proposed machine learning and neighborhood-enabled approach has a higher APDR ratio than existing field proven approaches as shown in Figure 5. Furthermore, when the proposed scheme is integrated with SVM, then it APDR ratio is 90.21%, which is highest among all machine learning approaches when integrated with the proposed scheme. Similarly, when it is integrated or embedded with K-mean, IDK, cosine, and Euclidean distances, then it APDR ratios are 87.21%, 68.58%, 65.15, and 86.24% respectively. From Figure 5, we can conclude that the maximum APDR ratio is 93.22%, that is achieved by the proposed scheme. Moreover, performance of the proposed fusion is approximately 3.44% better than SVM algorithm. Likewise, it has improve APDR ratio by 31.53%, 3.5%, 4.60%, and 38.46% than k-mean, IDK, Euclidean, and cosine distances respectively.

E. Throughput of IoT Devices When Embedded With the Proposed Scheme

A graphical view of the respective throughput of the proposed and existing schemes are shown in Figure 6, where the former approach has outperformed the later approaches. Secondly, if the proposed approach is integrated with the K-mean approach, then its performance is further improved. The maximum possible value of the respective through is approximately 66.57%, that is achieved by the IoT when the proposed machine learning and neighborhood-enabled scheme is embedded in both devices and server modules. Additionally, packet loss ratio is controlled if proposed scheme is integrated with k-mean or SVM approaches. The packet loss ratios are reduced by 1.54% than K-mean algorithm. Similarly, when it is embedded with SVM, IDK, Euclidean and cosine, then its reduction ratios are 12.84%, 19.12%, 5.54%, and 52.54% respectively.

F. Lifetime of IoT

A graphical representation of the proposed neighborhood-enabled refinement and existing schemes is shown in Figure 7, which concludes that lifetime of the IoT is prolonged if the proposed approach is integrated with the SVM. Additionally, ratio of the dead devices, those completely exhausted, was observed to be very less, that is 16 and 23, when proposed and SVM are utilized to form a hybrid scheme. Similarly, when k-mean is embedded with the proposed algorithm, then ratio of the dead devices was observed to be 35, that is slightly higher as compared to the SVM. Moreover, when integration is made with IDK, Euclidean, and Cosine, the observed ratios are 33, 51 and 42 respectively. Finally, ratio of the dead and alive devices, when the proposed scheme is integrated with different machine learning approach are given as 34.28%, 30.30%, 53.06%, 45.38%, and 39.47% from k-mean, SVM, IDK, Euclidean, and cosine distances respectively.

G. Normalized Communication Overhead

As shown in Figure 8, the proposed scheme when integrated with SVM has the minimum possible overhead, i.e., 1.015. Moreover, existing schemes have a relatively higher overhead than the proposed approach. Furthermore, the normalized overhead of the proposed fusion approach when integrated with k-mean, IDK, Euclidean, and cosine distances are 1.19, 1.015, 1.45, 1.16, and 1.56 respectively. Finally, an extensive

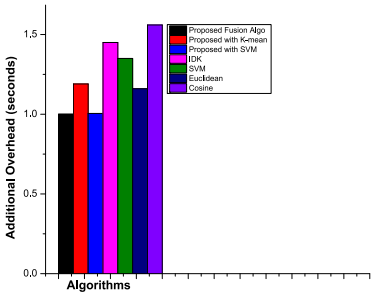

Fig. 8. Overhead Ratio of the Proposed and Existing Algo.

TABLE I
PERFORMANCE EVALUATION OF THE PROPOSED AND EXISTING ALGORITHMS

Algo	Fusion	Power	delay	overhead
Prop Algo	70.23%	6.55 J	1.532 ms	1.001
Pro+SVM	45.28%	8.54 J	1.87 ms	1.015
Pro+k-mean	66.23%	8.1 J	2.92 ms	1.19
IDK	69.56%	8.678 J	3.41 ms	1.45
SVM	55.56%	9.1 J	2.11 ms	1.35
Euclidean	56.24%	8.92 J	2.96 ms	1.16
Cosine	57.6%	8.76 J	1.96 ms	1.56

TABLE II
PERFORMANCE EVALUATION OF THE PROPOSED AND EXISTING ALGORITHMS

Algo	APDR	Throughput	Dead Devices
Prop Algo	93.22%	66.26 Mbps	16
Pro+SVM	90.21%	63.22 Mbps	23
Pro+k-mean	87.21%	64.22 Mbps	35
IDK	68.58%	63.84 Mbps	33
SVM	87.11%	53.91 Mbps	51
Euclidean	86.24%	60.84 Mbps	42
Cosine	65.15%	42.01 Mbps	38

comparative analysis of the existing and proposed machine learning based scheme, i.e., individually and when integrated with other machine learning algorithms, is presented in Table I & II.

V. CONCLUSION AND FUTURE WORK

In this paper, a hybrid fusion algorithm, that is based on Machine Learning and neighborhood information, was presented for the refinement of data, that is captured through consumer electronic devices in the Internet of Things. Initially, every server module gathers neighborhood information about every electronic device deployed in its coverage area and are divided into two disjoint groups, (i) Group-I where fusion process is mandatory and (ii) Group-II where refinement is optional. In next phase, machine learning algorithms such as k-mean and SVM were utilized to refine those datasets which are captured through electronic devices deployed in closed neighborhood and, thus, have a slightly higher ratio of duplicate data values. Simulation results have verified the exceptional performance of the proposed machine learning and neighborhood-enabled fusion algorithm particularly in terms of various evaluation metrics such as minimum ratio of duplicates, APDR, throughput, network's lifetime, and overhead ratio.

REFERENCES

[1] C. K. Wu, C.-T. Cheng, Y. Uwate, G. Chen, S. Mumtaz, and K. F. Tsang, "State-of-the-art and research opportunities for next-generation consumer electronics," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 937–948, Nov. 2023.
[2] D. Javeed, M. S. Saeed, I. Ahmad, P. Kumar, A. Jolfaei, and M. Tahir, "An intelligent intrusion detection system for smart consumer electronics network," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 906–913, Nov. 2023.
[3] J. Pei, S. Li, Z. Yu, L. Ho, W. Liu, and L. Wang, "Federated learning encounters 6G wireless communication in the scenario of Internet of Things," *IEEE Commun. Stand. Mag.*, vol. 7, no. 1, pp. 94–100, Mar. 2023.
[4] M. Li, "A lightweight architecture for query-by-example keyword spotting on low-power IoT devices," *IEEE Trans. Consum. Electron.*, vol. 69, no. 1, pp. 65–75, Feb. 2023.
[5] A. Li, B. Zheng, and L. Li, "Intelligent transportation application and analysis for multi-sensor information fusion of Internet of Things," *IEEE Sensors J.*, vol. 21, no. 22, pp. 25035–25042, Nov. 2021.
[6] X. Wang and Y. Wu, "Fog-assisted Internet of Medical Things for smart healthcare," *IEEE Trans. Consum. Electron.*, vol. 69, no. 3, pp. 391–399, Aug. 2023.
[7] S. Hadzovic, S. Mrdovic, and M. Radonjic, "A path towards an Internet of Things and artificial intelligence regulatory framework," *IEEE Commun. Mag.*, vol. 61, no. 7, pp. 90–96, Jul. 2023.
[8] F. Pescador and S. P. Mohanty, "Machine learning for smart electronic systems," *IEEE Trans. Consum. Electron.*, vol. 67, no. 4, pp. 224–225, Nov. 2021.
[9] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Multi-level feature fusion for multimodal human activity recognition in Internet of Healthcare Things," *Inf. Fusion*, vol. 94, pp. 17–31, Jun. 2023.
[10] M. Rida, A. Makhoul, H. Harb, D. Laiymani, and M. Barhamgi, "EK-means: A new clustering approach for datasets classification in sensor networks," *Ad Hoc Netw.*, vol. 84, pp. 158–169, Mar. 2019.
[11] S. Almaghrabi, M. Rana, M. Hamilton, and M. S. Rahaman, "Multivariate solar power time series forecasting using multilevel data fusion and deep neural networks," *Inf. Fusion*, vol. 104, Apr. 2024, Art. no. 102180.
[12] F. Belmahdi, M. Lazri, F. Ouallouche, K. Labadi, R. Absi, and S. Ameur, "Application of Dempster-Shafer theory for optimization of precipitation classification and estimation results from remote sensing data using machine learning," *Remote Sens. Appl. Soc. Environ.*, vol. 29, Jan. 2023, Art. no. 100906.
[13] F. Ghazipour and N. Mahjouri, "A multi-model data fusion methodology for seasonal drought forecasting under uncertainty: Application of Bayesian maximum entropy," *J. Environ. Manag.*, vol. 304, Feb. 2022, Art. no. 114245.
[14] Y. Chen, S. Yang, J.-F. Martínez, L. López, and Z. Yang, "A resilient group-based multisubset data aggregation scheme for smart grid," *IEEE Internet Things J.*, vol. 10, no. 15, pp. 13649–13661, Aug. 2023.
[15] S. Pattamaset and J. S. Choi, "Irrelevant data elimination based on a k-means clustering algorithm for efficient data aggregation and human activity classification in smart home sensor networks," *Int. J. Distrib. Sens. Netw.*, vol. 16, no. 6, 2020, Art. no. 1550147720929828.
[16] H. Harb, A. Makhoul, S. Tawbi, and R. Couturier, "Comparison of different data aggregation techniques in distributed sensor networks," *IEEE Access*, vol. 5, pp. 4250–4263, 2017.
[17] V. Chandran and Nikesh, "Elimination of data redundancy and latency improving in wireless sensor networks," *Int. J. Eng. Res. Technol.*, vol. 3, no. 8, pp. 435–439, 2014.
[18] S. R. U. Jan, R. Khan, and M. A. Jan, "An energy-efficient data aggregation approach for cluster-based wireless sensor networks," *Ann. Telecommun.*, vol. 76, pp. 321–329, Jun. 2021.