ENSEMBLE
Journal of Applied Mathematics, Statistics, and Data Science
eISSN: xxxx-xxxx
Vol. 1, No. 1, June 2026
Pages: 1-9

Clustering Criminal Offense Areas in the Kolaka Police Jurisdiction Using the K-Means Method

Siska Nur Fajriani[1], Rabiah Adawiyah[2], Yuwanda Purnamasari Pasrun[3*]

[1,2,3] Department of Information System, Universitas Sembilanbelas November, Kolaka 93517, Indonesia

E-mail Correspondence Author: *yuwandapurnamasari@gmail.com

: xxxxxx

ARTICLE INFO

Article history

Received: January 17, 2026
Revised: March 15, 2026
Accepted: April 10, 2026

Keywords:
Clustering;
Crime Mapping;
Euclidean Distance;
K-Means;
Web Information System

Abstract

Criminal offenses are a common problem that occurs in daily life, including in Kabupaten Kolaka, Indonesia. Different types of crime occur at different places, times, and frequencies, so that the annual crime rate fluctuates and community members - and the police in particular - find it difficult to determine which areas require closer supervision. This study designs and builds a web-based information system that groups the sub-jurisdictions of Polres Kolaka according to their level of criminality using the K-Means clustering algorithm. The system was developed with the waterfall model, covering requirement analysis, design (data flow diagrams, an entity-relationship diagram, and flowcharts), PHP/MySQL implementation, and black-box testing. A dataset of 550 recorded cases across 8 sub-jurisdictions and 8 offense categories (theft, immorality, armed robbery, assault, fraud, illegal liquor, gambling, and narcotics) was clustered into three groups - low, moderate, and high criminality - using Euclidean distance with randomly selected initial centroids. The clustering process converged after four iterations, when the same cluster membership was obtained in two consecutive iterations. The results show that Polres Kolaka forms its own high-criminality cluster, Polsek Wundulako, Polsek Pomalaa, and Polsek Watubangga form a moderate cluster, while the remaining four sub-jurisdictions form a low-criminality cluster. All seven black-box test scenarios were valid, and the clustering results produced by the system matched the manual computation for all 8 sub-jurisdictions. The system provides the public and, in particular, the police with quick, accurate information on crime-prone areas to support patrol planning and resource allocation.

1. INTRODUCTION

Crime prevention and control is one of the principal mandates of the Indonesian National Police (POLRI), which is responsible for maintaining public order, upholding the law, and protecting the community while respecting human rights. Within POLRI, the Criminal Investigation Agency (Bareskrim) carries out investigation, monitoring, and the management of national criminal information in support of these duties. Despite continuing efforts by the government and the community, criminal acts keep recurring with different modes of operation, and law enforcement is not always felt to be fully effective by the public it serves.

Kabupaten Kolaka, located in Southeast Sulawesi, is no exception. Cases of theft, immorality, armed robbery, assault, fraud, illegal-liquor consumption, gambling, and narcotics abuse are recorded every year at Polres Kolaka and its subordinate sub-jurisdictions (Polsek), and

the proportion of each offense type fluctuates from year to year. This fluctuation makes it difficult for the police and the public to know, at a glance, which sub-jurisdictions require closer supervision and patrol. Grouping sub-jurisdictions by their level of criminality would let the police prioritize preventive patrols and let the public stay informed about relatively unsafe areas.

Data-mining techniques, and clustering in particular, have been used extensively to turn raw crime records into actionable groupings. K-Means remains the most widely adopted clustering algorithm for this purpose because it is conceptually simple, computationally inexpensive, and scales well to the modest, structured datasets that are typical of a single-district police record system [3], [4], [5]. Pratama et al. used K-Means together with a Geographic Information System to group districts by crime and traffic-violation intensity into three vulnerability levels, feeding the result into a public web portal [1]. Maharrani et al. combined K-Means with principal component analysis to cluster criminal acts and reduce the dimensionality of offense-type attributes before clustering [2]. More generally, distance-based partitioning methods such as K-Means continue to be refined - for example, through alternative distance measures instead of the plain Euclidean metric - to reduce their sensitivity to the random initialization of centroids [6]. These studies confirm that K-Means is a practical and interpretable tool for prioritizing high-crime areas, but most of them either stop at a standalone analysis or focus on the algorithm itself rather than on an end-to-end, verifiable operational system that a police sector office could use directly.

This study contributes a complete pipeline that (i) formulates the criminality-level grouping problem for a real police sector (Polres Kolaka) using 8 official offense categories recorded across 8 sub-jurisdictions, (ii) documents a manual, step-by-step K-Means computation so that the clustering logic is fully auditable, and (iii) implements the same logic as a web-based information system whose output is cross-checked against the manual computation for every sub-jurisdiction. The objective of the research is therefore to design, build, and verify a web-based system that groups criminal-offense areas within the Polres Kolaka sector using the K-Means clustering method, so that both the police and the public can quickly obtain accurate information about crime-prone areas.

2. METHOD

2.1 Research Object and Data Collection

The research was carried out at Polres Kolaka over a period of three months, covering system analysis, system design, programming, and system testing. Data were collected through direct observation at Polres Kolaka, interviews with police officers who handle criminal case records, and a literature study of previous K-Means clustering research. A total of 550 criminal case records were obtained, spanning 8 sub-jurisdictions (Table 1) and 8 offense-type attributes (Table 2).

Table 1. Sub-Jurisdiction Codes

Code	Sub-jurisdiction
Wil-001	Polres Kolaka
Wil-002	Polsek Kolaka
Wil-003	Polsek Kawasan Pelabuhan
Wil-004	Polsek Samaturu
Wil-005	Polsek Wolo
Wil-006	Polsek Wundulako
Wil-007	Polsek Pomalaa
Wil-008	Polsek Watubangga

Table 2. Offense-Type Attribute Codes

Code	Offense type
A01	Theft
A02	Immorality
A03	Armed robbery (sharp weapon/firearm)
A04	Assault
A05	Fraud
A06	Illegal liquor
A07	Gambling
A08	Narcotics

The number of recorded cases per sub-jurisdiction and offense type, which forms the clustering dataset, is presented in Table 3.

Table 3. Number of Cases per Sub-Jurisdiction and Offense Type

Region	A01	A02	A03	A04	A05	A06	A07	A08
Wil-001	86	25	21	73	28	14	11	50
Wil-002	2	1	1	13	1	0	2	0
Wil-003	1	1	1	5	0	0	0	0
Wil-004	13	0	1	11	0	0	0	1
Wil-005	6	4	0	16	4	0	1	2
Wil-006	17	1	2	21	1	0	0	0
Wil-007	25	7	1	25	0	0	0	3
Wil-008	18	3	3	27	1	0	0	0

2.2 System Development Method

Explaining the research chronologically, including the research design, research The information system was developed using the waterfall model, consisting of four sequential stages [10], [12], [13], [14]. In the analysis stage, functional and data requirements were derived from observation, interviews, and the literature study described above. In the design stage, a context diagram and a level-1 data flow diagram were used to describe the flow of data into and out of the system [9]; an entity-relationship diagram was used to model the wilayah (region), atribut (attribute), kasus (case), and hasil (result) tables and their relationships [8]; and flowcharts were used to describe the control flow of each page, including the core K-Means clustering page. In the programming stage, the system was implemented in PHP with MySQL as the database, accessed through phpMyAdmin, and served locally through XAMPP. In the testing stage, black-box testing was used to verify that each function of the system produced the expected output without requiring knowledge of the internal code structure [11], [15].

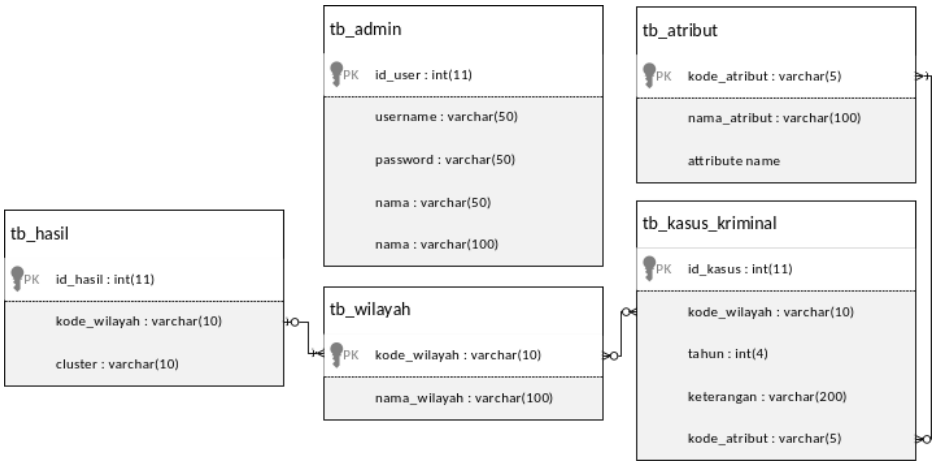


Figure 1. Entity-Relationship Diagram of the Clustering Information System

Figure 2. Context Diagram (Level-0 Data Flow Diagram) of the System

As shown in Figure 2, the admin performs the input of region data, attribute data, and criminal-case data, while the output of the system - the clustering results for each region - can be viewed by both the admin and the general public through the website.

2.3 K-Means Clustering Procedure

The K-Means algorithm partitions the eight sub-jurisdictions into k = 3 clusters, labelled C1 (low criminality), C2 (moderate criminality), and C3 (high criminality), following the general procedure described by MacQueen [4] and later reviewed extensively by Jain [5] and Ikotun et al. [3]. The steps applied in this study are as follows.

1) Determine the number of clusters k and select k initial centroids. In this study the initial centroids were selected directly from three representative rows of the case-count data in Table 3 (Table 4).
2) Compute the Euclidean distance from every region's case-count vector to each of the three centroids, using Equation (1) [4], [5], [6].

$$d(X_i, C_i) = \sqrt{\sum_{i=n}^{n} (X_i - C_i)^2}$$

where d is the Euclidean distance, X_i is the case-count value of region i for a given offense attribute, and C_i is the corresponding value of centroid i.

3) Assign each region to the cluster whose centroid gives the shortest distance.
4) Recompute each cluster's centroid as the arithmetic means of the regions currently assigned to it.
5) Repeat steps 2-4 until the cluster membership of every region is identical in two consecutive iterations, at which point the algorithm has converged.

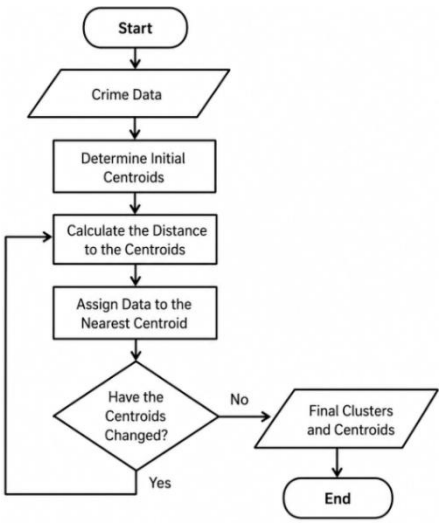


Figure 3. Flowchart of the K-Means Clustering Procedure

3. RESULTS AND DISCUSSION

3.1 Initial Centroids and Iterative Computation

Following the procedure in Section 2.3, three regions were selected as the initial centroids: the region with the smallest total case count (Wil-003) as C1-Low, a mid-range region (Wil-006) as C2-Moderate, and the region with the largest total case count (Wil-001) as C3-High (Table 4).

Table 4. Initial Centroids

Cluster	A01	A02	A03	A04	A05	A06	A07	A08
C1 - Low	1	1	1	5	0	0	0	0
C2 - Moderate	17	1	2	21	1	0	0	0
C3 - High	86	25	21	73	28	14	11	50

The Euclidean distance of every region to C1, C2, and C3 was computed with Equation (1) and the shortest distance was used to assign a temporary cluster in every iteration. Table 5 shows the shortest-distance assignment obtained in the fourth iteration, at which the algorithm converged because the third and fourth iterations produced identical cluster memberships.

Table 5. Shortest Euclidean Distance, Iteration 4 (Convergence)

Region	C1	C2	C3	Nearest cluster
Wil-001	121.24	104.91	0.00	3
Wil-002	4.22	21.58	123.04	1
Wil-003	7.89	27.28	128.16	1
Wil-004	7.80	15.55	117.05	1
Wil-005	6.23	16.81	117.08	1
Wil-006	15.18	5.32	109.31	2
Wil-007	24.63	6.48	100.30	2
Wil-008	20.32	3.70	105.33	2

Table 6 summarizes the final cluster membership produced by the manual computation.

Table 6. Final Cluster Membership

Code	Sub-jurisdiction	Cluster
Wil-001	Polres Kolaka	C3 - High
Wil-002	Polsek Kolaka	C1 - Low
Wil-003	Polsek Kawasan Pelabuhan	C1 - Low
Wil-004	Polsek Samaturu	C1 - Low
Wil-005	Polsek Wolo	C1 - Low
Wil-006	Polsek Wundulako	C2 - Moderate
Wil-007	Polsek Pomalaa	C2 - Moderate
Wil-008	Polsek Watubangga	C2 - Moderate

The final clustering shows that Polres Kolaka alone forms the high-criminality cluster, reflecting the fact that it records the largest number of cases in seven of the eight offense categories. Polsek Wundulako, Polsek Pomalaa, and Polsek Watubangga form the moderate-criminality cluster, while Polsek Kolaka, Polsek Kawasan Pelabuhan, Polsek Samaturu, and Polsek Wolo form the low-criminality cluster. This grouping is consistent with the pattern reported in comparable regional crime-clustering studies, where the central or main police unit of a district typically records substantially more cases than its subordinate sub-jurisdictions [1], [2], [7].

3.2 Web-Based System Implementation

The clustering logic verified manually in Section 3.1 was implemented as a PHP/MySQL web application. The public-facing page displays the clustering result for every sub-jurisdiction and lets a visitor open a detail page that shows the case count of each offense type together with a bar chart, as shown in Figure 4.

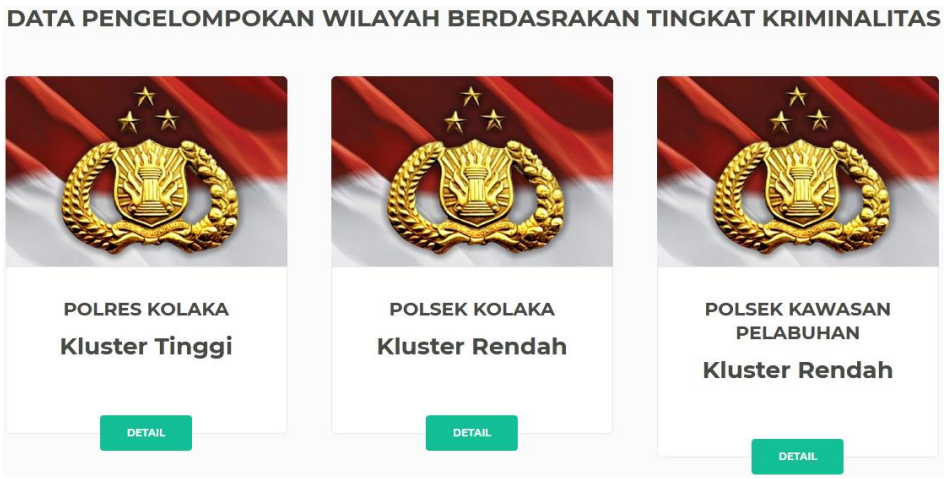


Figure 4. Public Page Showing Clustering Results by Region

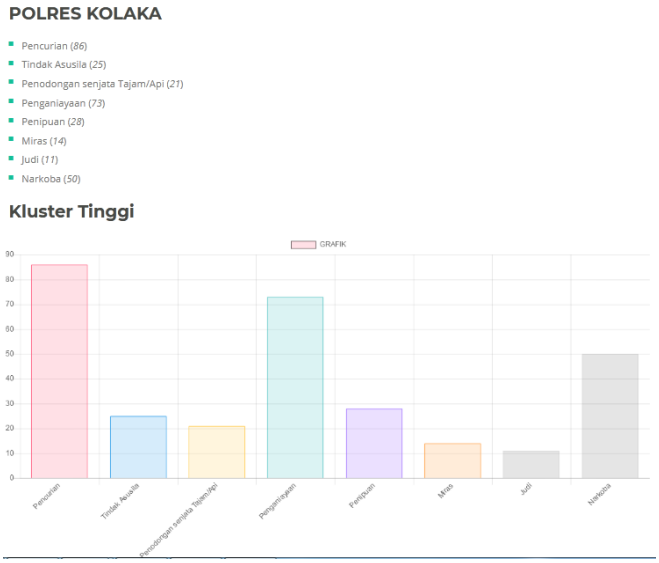
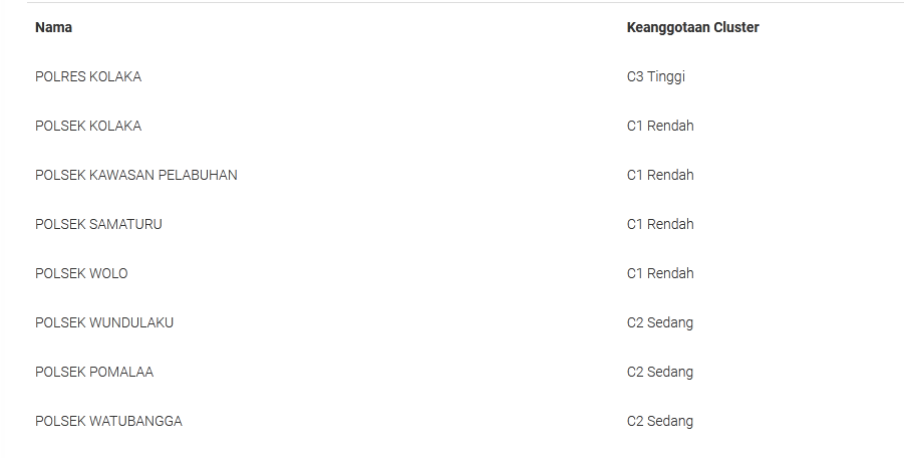


Figure 5. Detail Page Showing the Case-Count Chart of a Selected Region

On the administrative side, after the admin enters the initial centroids, the system automatically executes the Euclidean-distance computation, iterates until convergence, and displays the resulting cluster for every region on a dedicated results page (Figure 6), matching the manual procedure described in Section 2.3.



Nama	Keanggotaan Cluster
POLRES KOLAKA	C3 Tinggi
POLSEK KOLAKA	C1 Rendah
POLSEK KAWASAN PELABUHAN	C1 Rendah
POLSEK SAMATURU	C1 Rendah
POLSEK WOLO	C1 Rendah
POLSEK WUNDULAKU	C2 Sedang
POLSEK POMALAA	C2 Sedang
POLSEK WATUBANGGA	C2 Sedang

Figure 6. System-Generated Clustering Result Page

3.3 Black-Box Testing

Black-box testing was carried out on the main functions of the system: viewing cluster detail, admin login (invalid and valid credentials), region/attribute/case data entry with correct and empty input, and running the clustering process with and without initial centroids. Table 7 summarizes a representative sample of the seven test scenarios.

Table 7. Black-Box Testing Summary

No.	Test scenario	Expected result	Status
1	Click the cluster detail button	System displays the case detail of the selected region	Valid
2	Log in with an incorrect username/password	System rejects access and shows an error message	Valid
3	Log in with a correct username/password	System redirects to the admin main page	Valid
4	Submit a data-entry form with valid input	System shows a data-saved confirmation message	Valid
5	Submit a data-entry form with empty input	Input fields are highlighted and a "required field" message is shown	Valid
6	Run clustering without entering initial centroids	Input fields are highlighted and a "required field" message is shown	Valid
7	Run clustering after entering initial centroids	System displays the K-Means clustering result	Valid

All seven scenarios were valid, i.e., the actual system behavior matched the expected behavior in every case, indicating that the system is functionally ready for operational use [11].

3.4 Verification of Manual versus System Computation

To confirm that the system correctly reproduces the K-Means logic verified manually in Section 3.1, the cluster assigned to each of the eight sub-jurisdictions by the system was compared with the manual result (Table 8).

Table 8. Comparison of Manual and System Clustering Results

No.	Sub-jurisdiction	Manual result	System result	Comparison
1	Polsek Kolaka	C1 - Low	C1 - Low	Match
2	Polsek Kawasan Pelabuhan	C1 - Low	C1 - Low	Match
3	Polsek Samaturu	C1 - Low	C1 - Low	Match
4	Polsek Wolo	C1 - Low	C1 - Low	Match

No.	Sub-jurisdiction	Manual result	System result	Comparison
5	Polsek Wundulako	C2 - Moderate	C2 - Moderate	Match
6	Polsek Pomalaa	C2 - Moderate	C2 - Moderate	Match
7	Polsek Watubangga	C2 - Moderate	C2 - Moderate	Match
8	Polres Kolaka	C3 - High	C3 - High	Match

The system reproduced the manual computation exactly for all eight sub-jurisdictions, confirming that the implemented K-Means procedure is computationally correct. This agreement between manual and automated computation is an important reliability check that is often omitted in prior web-based crime-clustering systems, which typically report only the final clusters without demonstrating that the underlying implementation matches a verifiable reference computation [1], [2].

3.5 Discussion

The results confirm that K-Means clustering, despite its well-known sensitivity to the choice of initial centroids and to the Euclidean distance metric [3], [6], remains a practical tool for a small, well-structured municipal crime dataset such as the one used in this study: the algorithm converged after only four iterations and produced a three-level grouping that is directly interpretable by non-technical police staff. Compared with GIS-based approaches that overlay clustering results on a spatial map [1], the present system instead emphasizes transparency and verifiability of the clustering computation itself, which is valuable when the output is intended to justify operational decisions such as patrol allocation. As in dimensionality-reduction-based studies of crime clustering [2], future extensions of this work could incorporate additional contextual attributes - such as population density, time of day, or distance between sub-jurisdictions - and could explore alternative distance measures or automatic centroid initialization strategies to reduce dependence on the arbitrary selection of initial centroids [3], [6].

4. CONCLUSION

A web-based information system that groups the sub-jurisdictions of Polres Kolaka by criminality level using the K-Means clustering algorithm has been successfully designed and built with the waterfall development model. Using 550 case records spanning 8 sub-jurisdictions and 8 offense-type attributes, the manual K-Means computation converged after four iterations and produced three clusters: Polres Kolaka as the sole member of the high-criminality cluster; Polsek Wundulako, Polsek Pomalaa, and Polsek Watubangga as the moderate-criminality cluster; and Polsek Kolaka, Polsek Kawasan Pelabuhan, Polsek Samaturu, and Polsek Wolo as the low-criminality cluster. Black-box testing showed that all seven tested scenarios behaved as expected, and the clustering results generated by the system matched the manual computation for every sub-jurisdiction, confirming that the system is both functionally reliable and computationally correct. The system enables both the police and the general public to quickly obtain accurate, up-to-date information on crime-prone areas, supporting more targeted patrol planning and resource allocation. Future work should extend the dataset to a longer time span and to the entire jurisdiction of Southeast Sulawesi Regional Police, compare K-Means against alternative clustering algorithms, and consider a mobile-platform implementation.

Acknowledgments

The authors thank the Kolaka Resort Police (Polres Kolaka), in particular the Criminal Investigation Unit and the Narcotics Unit, for granting access to the criminal case data used in this study.

REFERENCES

[1] M. Sigala, A. Beer, L. Hodgson, and A. O'Connor, *Big Data for Measuring the Impact of Tourism Economic Development Programmes: A Process and Quality Criteria Framework for Using Big Data*. 2019.

[2] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.

[3] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J.*

Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.
[4] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," *IEEE Access*, vol. 7, pp. 41525–41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
[5] K. Sivaraman, R. M. V. Krishnan, B. Sundarraj, and S. Sri Gowthem, "Network failure detection and diagnosis by analyzing syslog and SNS data: Applying big data analysis to network operations," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 9 Special Issue 3, pp. 883–887, 2019, doi: 10.35940/ijitee.I3187.0789S319.
[6] A. D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A decentralized privacy-preserving healthcare blockchain for IoT," *Sensors (Switzerland)*, vol. 19, no. 2, pp. 1–17, 2019, doi: 10.3390/s19020326.
[7] F. Al-Turjman, H. Zahmatkesh, and L. Mostarda, "Quantifying uncertainty in internet of medical things and big-data services using intelligence and deep learning," *IEEE Access*, vol. 7, pp. 115749–115759, 2019, doi: 10.1109/ACCESS.2019.2931637.
[8] S. Kumar and M. Singh, "Big data analytics for healthcare industry: Impact, applications, and tools," *Big Data Min. Anal.*, vol. 2, no. 1, pp. 48–57, 2019, doi: 10.26599/BDMA.2018.9020031.
[9] L. M. Ang, K. P. Seng, G. K. Ijemaru, and A. M. Zungeru, "Deployment of IoV for Smart Cities: Applications, Architecture, and Challenges," *IEEE Access*, vol. 7, pp. 6473–6492, 2019, doi: 10.1109/ACCESS.2018.2887076.
[10] B. P. L. Lau *et al.*, "A survey of data fusion in smart city applications," *Inf. Fusion*, vol. 52, no. January, pp. 357–374, 2019, doi: 10.1016/j.inffus.2019.05.004.
[11] Y. Wu *et al.*, "Large scale incremental learning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2019-June, pp. 374–382, 2019, doi: 10.1109/CVPR.2019.00046.
[12] A. Mosavi, S. Shamshirband, E. Salwana, K. wing Chau, and J. H. M. Tah, "Prediction of multi-inputs bubble column reactor using a novel hybrid model of computational fluid dynamics and machine learning," *Eng. Appl. Comput. Fluid Mech.*, vol. 13, no. 1, pp. 482–492, 2019, doi: 10.1080/19942060.2019.1613448.
[13] V. Palanisamy and R. Thirunavukarasu, "Implications of big data analytics in developing healthcare frameworks – A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 31, no. 4, pp. 415–425, 2019, doi: 10.1016/j.jksuci.2017.12.007.
[14] J. Sadowski, "When data is capital: Datafication, accumulation, and extraction," *Big Data Soc.*, vol. 6, no. 1, pp. 1–12, 2019, doi: 10.1177/2053951718820549.
[15] J. R. Saura, B. R. Herraez, and A. Reyes-Menendez, "Comparing a traditional approach for financial brand communication analysis with a big data analytics technique," *IEEE Access*, vol. 7, pp. 37100–37108, 2019, doi: 10.1109/ACCESS.2019.2905301.