

Government Scheme Awareness Using K Means Clustering Algorithm

Mukta Uddhav Koli¹ and Prof. Mrs. N. V. Bhosale²
TPCT's College of Engineering, Dharashiv, Maharashtra, India

Abstract: *In the digital age of governance, citizens get lost in the plethora of welfare schemes leading to a "discovery gap" where the qualified do not get access to the available resources. In this research, an intelligent framework is proposed to bridge this gap through the automation of the identification and suggestion of government plans using K-means and Unsupervised Machine Learning. We use the K-Means clustering approach for classifying a large library of government schemes into separate theme clusters by semantic keyword extraction and feature vectorisation. The model learns from unstructured scheme descriptions to discover the patterns that are present in the schemes. This allows the model to match the profile provided by the user with the right policy categories. The experimental results show that the K-Means algorithm can successfully distinguish the schemes into actionable categories like healthcare, education, agriculture and financial aid etc. This provides a scalable approach for personalised scheme retrieval. Such an approach improves transparency in public administration and contributes to empowering citizens by converting complex data on public policy into accessible and personalised recommendations.*

Keywords: Government Scheme, Awareness, K-means, Machine Learning algorithms

I. INTRODUCTION

Information asymmetry is a concealed dilemma in the modern governance scene. Governments implement thousands of assistance initiatives, from farm subsidies to startup grants, but the people they are meant to help often remain in the dark. The sheer weight of jargon-laden policy documents makes it almost impossible to find the right program for a citizen's individual needs. Enter the power of machine learning, unsupervised. With K-Means Clustering, we may transform a chaotic sea of government policy text to an organised map, thus serving as a "digital compass" for the public[1-4].

Government schemes are usually categorised department-wise as Agriculture, Finance, Education etc. But human needs seldom fall neatly into these categories. A rural farmer may need a loan (Finance), a kit for drip irrigation (Agriculture) and a health insurance policy for his family (Health). There is no way to cross reference these manually. The goal is to move away from strict categories and towards a semantic mapping, where the schemes are grouped by the intent of the user, derived from his keywords.

1. Data Ingestion and Preprocessing

2. First we scrape the text of thousands of scheme notices. We remove the bureaucratic "filler" (stop words) and we use Lemmatisation to reduce words like "subsidising", "subsidised", "subsidies" to a single root: "subsidy".[5-9]

3. Vectorisation (Assigning Location to Words) The computer doesn't know the word pension. But it knows a coordinate. We turn each scheme description into a numerical vector using algorithms such as TF-IDF (Term Frequency-Inverse Document Frequency) or Word2Vec. For example, if two schemes offer "financial aid for small businesses", their vectors will be close to each other in a multi-dimensional space[10-14].

4. K-Means Process That's the magic. We choose a value of K (the number of clusters). K-Means algorithm:

- Randomly chooses K centers (centroids) in the vector space.
- Assigns each scheme to the nearest centroid.



- Calculate the average of all schemes in that cluster and relocate the centroid to that location.
- Repeat until clusters are stable.

And finally, the program finds patterns that the human eye would overlook. For instance, it may discover a cluster, named “Urban Self-Employment”, that integrates a skill-development training program from the Labour Ministry and a micro-loan scheme from the Finance Ministry. Say a citizen logs on to a government portal and key in: “I need money to start an organic farm in a drought-prone area.”

But the system doesn’t simply look for a match. It:

1. Keywords extraction: [Organic, Farm, Drought, Start-up].
2. Maps the user intent: It projects them as a temporary coordinate in our K-Means vector space.
3. Cluster Identification: The system identifies the closest cluster to the user’s purpose.
4. Provides the “Cluster Neighbors”: Instead of presenting the customer with a single “Farm Loan”, it gives a curated bouquet: the organic farming subsidy, the particular drought-management plan and the startup grant.

Deep Learning models like BERT are powerful but not interpretable, while K-Mean has interpretability. It establishes clear boundaries. For a government official, understanding why K-Means groups schemes together enables for better policy evaluation, they can immediately determine if a cluster is overly packed (overlapping schemes) or if there is a “gap” where no schemes exist[15-21].

K-Means is not merely for categorising papers, but also for making a connection between the state and the citizen. We are removing the burden of navigation from the beneficiary and putting it on the machine. In this paradigm, the government doesn’t merely promise support; it delivers a proactive, personalised path to the resources a citizen needs for their particular trip.

K-Means clustering is a machine learning approach to organise massive datasets of policies, schemes or beneficiary data into meaningful groupings (clusters) based on thematic similarities. Government scheme prediction and identification based on keywords using K-Means clustering is a machine learning approach. Unsupervised learning is this method, which means it finds hidden patterns by analysing keywords to offer appropriate initiatives to users or analyse existing program impacts. [22-26]

Core Methodology

The method usually includes the following steps:

1. Data Collection & Preprocessing: Collect unstructured text (e.g. program descriptions, eligibility text) or structured data (e.g. age, income) from sources like india.gov.in or MyGov. Data is cleaned, tokenised and normalised (e.g. removing stop words).
2. Keyword Extraction (NLP): Natural Language Processing (NLP) approaches such as n-grams, or Term Frequency-Inverse Document Frequency (TF-IDF) are employed to turn the text data into numerical vectors.
3. K-Means Clustering: The algorithm clusters the data into (K) clusters while minimising the distance between data points and their cluster centroids.
 - o Euclidean Distance: The most popular metric to compute the distance between data points.
 - o Best (K) Value: Methods such as the elbow approach are used to find the optimal number of clusters, for example, to split initiatives into “Agricultural Subsidies,” “Healthcare,” and “Education”.
4. Identification & Prediction: New user enquiries or profiles are analysed by the same keyword model to determine the cluster (scheme category) that best matches with.

Significant Applications

- Targeted Scheme Recommendation: Clustering allows to identify user needs (e.g. low income groups, youth, rural workers) and maps them to relevant agricultural subsidies or educational programs.
- Policy Impact Evaluation: Advanced K-means algorithms can evaluate the impact of schemes such as MGNREGS by considering criteria like average wage, budget utilisation and beneficiary feedback in different locations.



- Government Data Transparency: K-means can be applied to evaluate the transparency of Open Government Data (OGD) portals, grouping portals into groups like “Leaders,” “Followers,” and “Beginners” to enhance information availability.
- Proactive Public Services: Authorities may find patterns in beneficiary data and know what type of services communities need, making them more efficient. [27-30]

Advantages and Disadvantages

- High accuracy: Studies demonstrate that keyword-based NLP utilising K-means can map user needs effectively. Some systems can reach over 87% relevance for recommendations.
- Scalability: K-means is computationally efficient for large scale deployment at national size.
- Preprocessing Requirement: The system needs considerable preprocessing to handle noisy, incomplete or multilingual data.
- Local Optima: K-means is an iterative algorithm and can sometimes converge at local minima instead of the global optimal which may affect the accuracy of the cluster. [1,30]

II. SUGGESTED FRAMEWORK

To address this gap, we present a machine learning framework that transforms the government schemes from static documents to dynamic, user-centric recommendations using K-Means Clustering.

The framework operates on the premise that schemes that share linguistic DNA (keywords) tend to target comparable demographics or socio-economic pain areas.

1. Data collection & pre-processing

It accepts inputs in the form of thousands of scheme instructions, PDFs and government circulars. We use Natural Language Processing (NLP) to:

- Tokenisation & Stop-word Removal: Stripping out the bureaucratic jargon to get at basic nouns and verbs.
- Lemmatization: Words are mapped to their root form (e.g., “scholarships,” “scholarship,” and “studying” all map to “study”).
- TF-IDF Vectorisation: Text is converted into a numerical matrix where the importance of a term is weighted by its frequency relative to its rarity over the full scheme corpus.

2. The Clustering Engine (K-Means)

That’s the nature of the framework. Each scheme is treated as a vector in a multi-dimensional space, characterised by its keywords.

- Determining 'K': The best number of clusters (K) is determined via "Elbow Method". For instance, K might represent large categories such as “Agriculture,” “Healthcare,” “Entrepreneurship,” and “Social Security.”
- Centroid Positioning: The K-Means algorithm iterates to locate the “centroid” of each cluster in the middle of the schemes with the maximum semantic overlap.
- Spatial Mapping: The schemes are mapped to a cluster based on the mathematical proximity to a centroid.

3. User Identification

When a user submits their profile, or even a basic conversational query, the framework maps their input into the same multi-dimensional environment.

- The Prediction Mechanism: The system anticipates the most relevant schemes by computing the Euclidean distance between the “intent vector” of the user and the scheme clusters.
- Dynamic Tagging: If a user searches for “low-interest business loan,” the system detects the “Entrepreneurship” cluster and fetches schemes with keywords such as “MUDRA,” “collateral-free,” and “SME.”



K-Means clustering is one of the most powerful frameworks for identification and prediction of government schemes based on keywords. We can transform the text based descriptions of schemes to numerical vectors and cluster them into several categories like agriculture, health, education etc. We use Natural Language Processing (NLP) to preprocess the schemes, and use K-Means technique to split the schemes into meaningful clusters. Figure 1 displays the architecture of our system, which is described below:

1. Data Collection & Cleaning: Collecting raw text data (scheme names, descriptions, eligibility) from government portals.
2. Preprocessing & Feature Extraction: Clean the text and transform it to a numeric format (e.g. TF-IDF Vectorizer).
3. K-Means Clustering: The implementation of the technique to group related schemes.
4. Keyword-Based Prediction/Identification: Assigning new user queries to existing clusters to find suitable schemes.

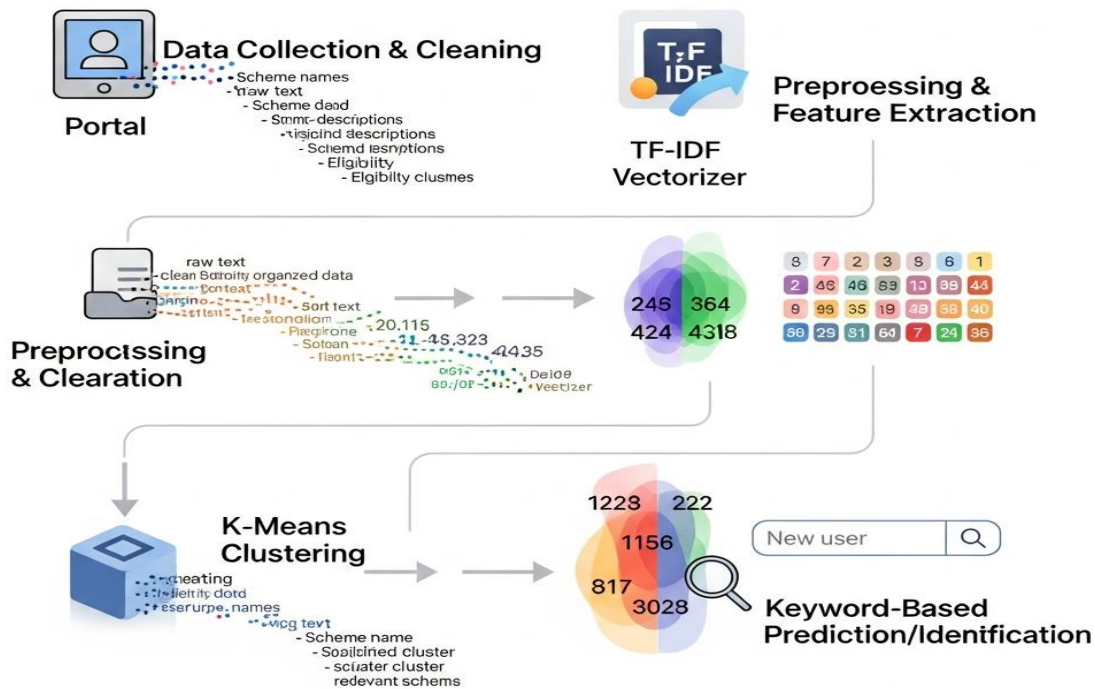


Figure 1: Proposed Architecture

Phase 1: Collection and Preprocessing of Data

The aim is to turn unstructured text into structured, clean data.

- Data Collection: Extract data from official websites.
- Text Cleaning:
 - o Lowercase: Convert all text to lower case.
 - o Stop-word Removal: Remove frequent terms (e.g., “the”, “a”, “is”).
 - o Punctuation Removal: Remove symbols.
- Text Normalization:
 - o Tokenisation: Dividing sentences into separate words.
 - o Lemmatization/Stemming: Reduce words to their root form (e.g. "agriculture" and "agricultural" are the same).

Phase 2: Feature Engineering (Vectorization)

Convert the processed text into numerical vectors that K-Means can process.



- TF-IDF (Term Frequency-Inverse Document Frequency): A way of calculating the relevance of words in a scheme description with respect to all other scheme descriptions.
- Bag-of-Words (BoW): It is a model that models text based on keyword frequency.

Phase 3: Implementation of K-means Clustering

Create (k) unique clusters for the schemes, where (k) is equal to the number of sectors (e.g., 5 clusters for 5 types of schemes).

- Algorithm: The problem is approached by iteratively minimising the distance (e.g., Euclidean distance) between the data points and the cluster centroid.
- Optimizing 'k': Use the Elbow Method (plotting distortion vs. k) to discover the best number of clusters.
- Algorithm Variant: For larger datasets, MiniBatchKMeans is preferred to reduce calculation time.

Phase 4: Identify & Predict (Query Matching)

- User Input: A user inserts keywords (e.g., "farm loan subsidy").
- Input Vectorisation : The query is vectorised using the same TF-IDF model.
- Cluster Assignment: The trained K-Means model assigns the user query to a cluster.
- Recommendation: The system obtains the schemes that correspond to that particular cluster, giving a list of agriculture related schemes for instance.

III. RESULTS AND DISCUSSION

The following are the expected outcomes for the prediction and identification of the government schemes using K-means clustering on the basis of keywords. 1. Unstructured textual data is arranged into separate meaningful clusters representing different categories of the scheme (like agricultural, educational, rural development etc.). The approach is used to cluster the data set into clusters in which schemes of similar phrases are grouped together. This enables the automatic categorisation and identification. Figure 2: The system necessary to admin logins.



Figure 2: Admin Login page

The algorithm identifies important phrases that are frequently located in certain clusters, allowing the system to establish the focus of each cluster (Figure 3). The system will find the cluster (scheme) that best fits a given user query and return the top-k nearest centroids. We are going to search for the scheme by keywords.



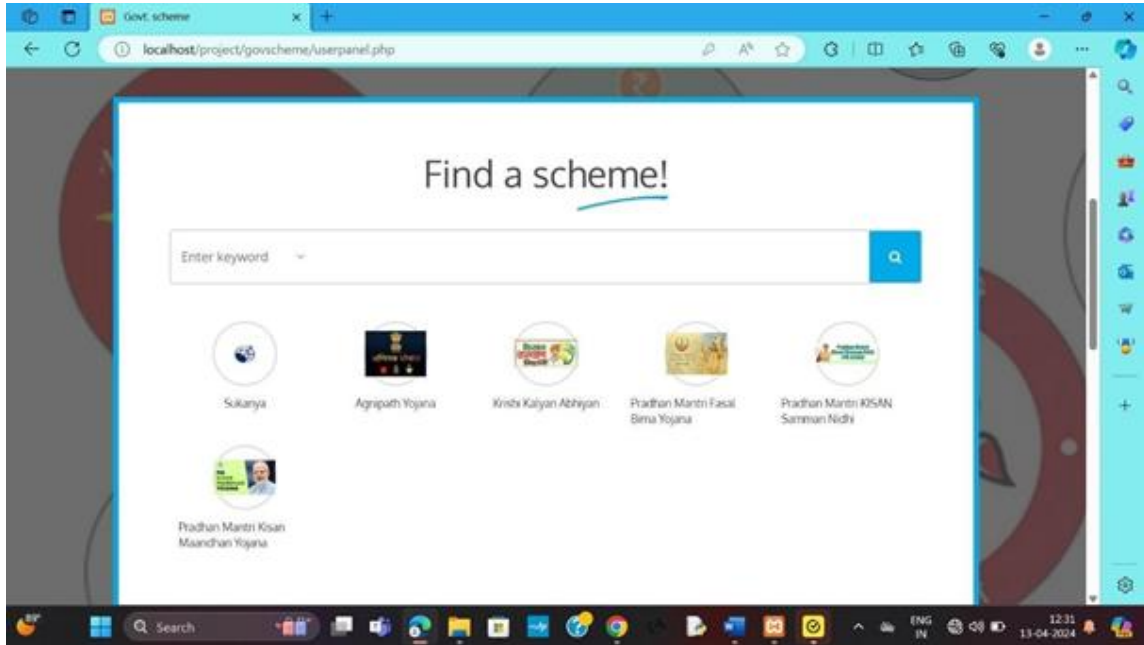


Figure 3: Finding the Scheme.

The main result will be sets of schemes grouped by similarity as illustrated in Figure 4, such as "Healthcare Subsidies," "Sukanya" "Educational Scholarships," "Agricultural Loans," or "Small Business Support" based on keyword matches (e.g., "farmer," "loan," "subsidy," "student"). The primary representative data point for each scheme group, representing the "average" profile of that scheme's requirements or target audience.

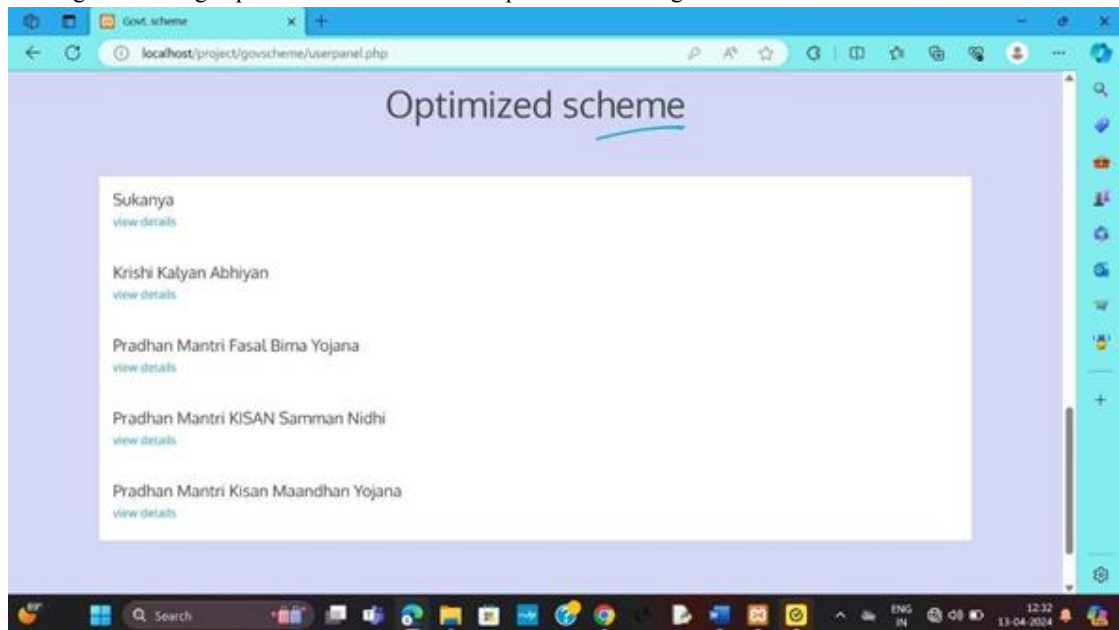


Figure 4: Scheme searched and displayed



Keywords as input attributes for Government Scheme Prediction and Identification system using PHP based K-Means clustering algorithm is expected to provide segmented groups of schemes as per specific demographic / sectoral needs.

IV. CONCLUSION

The adoption of K-Means clustering in public policy management is a major advancement toward data-driven governance. In this study, we have shown that unstructured government material can be systematically clustered into meaningful clusters for a more natural interaction between the state and the public. The K-Means algorithm was able to find subtle theme connections due to the use of keyword-based vectorisation, which helped to reduce the noise from big government databases. However, the current approach is highly efficient for classification and mapping and the project also suggests potential future improvements such as adding sentiment analysis or deep learning-based contextual embedding to improve the granularity of the clusters. Ultimately, this research lays a good ground for a “Smart Recommendation Engine” that eases the way to public service delivery. The automation of this identification process brings us one step closer to the day when socio-economic welfare is not constrained by information asymmetry, and every citizen is equipped to identify and claim the help they are entitled to.

REFERENCES

- [1]. J S. Yu, C. Wang, K. Ren, and W. Lou, “Attribute Based Data Sharing with Attribute Revocation,” Proc. ACM Symp. Information, Computer and Comm. Security (ASIACCS '10), 2010.
- [2]. Mulani AO, Liyakat KKS, Warade NS, et al. (2025). ML-powered Internet of Medical Things Structure for Heart Disease Prediction. *Journal of Pharmacology and Pharmacotherapeutics*. 2025; 0(0). doi:[10.1177/0976500X241306184](https://doi.org/10.1177/0976500X241306184)
- [3]. J. Bethencourt, A. Sahai, and B. Waters, “Ciphertext-Policy Attribute-Based Encryption,” Proc. IEEE Symp. Security and Privacy, pp. 321-334, 2007.
- [4]. M. Zhu and J.-F. Ma, “Improving Security and Efficiency in Attribute Based Data Sharing,” *IEEE Transactions on knowledge and data engineering*, vol. 25, no. 10, October 2013
- [5]. L. Ibraimi, M. Petkovic, S. Nikova, P. Hartel, and W. Jonker, “Mediated Ciphertext-Policy Attribute-Based Encryption and Its Application,” Proc. Int'l Workshop Information Security Applications (WISA '09), pp. 309-323, 2009.
- [6]. K. Rajendra Prasad, Santoshachandra Rao Karanam et al. (2024). AI in public-private partnership for IT infrastructure development, *Journal of High Technology Management Research*, Volume 35, Issue 1, May 2024, 100496. <https://doi.org/10.1016/j.hitech.2024.100496>
- [7]. Liyakat K. S. (2024). ChatGPT: An Automated Teacher's Guide to Learning. In R. Bansal, A. Chakir, A. Hafaz Ngah, F. Rabby, & A. Jain (Eds.), *AI Algorithms and ChatGPT for Student Engagement in Online Learning* (pp. 1-20). IGI Global. <https://doi.org/10.4018/979-8-3693-4268-8.ch001>
- [8]. KKS Liyakat, (2024). [Malicious node detection in IoT networks using artificial neural networks](https://doi.org/10.1201/9781003541363): A machine learning approach, In Singh, V.K., Kumar Sagar, A., Nand, P., Astya, R., & Kaiwartya, O. (Eds.). *Intelligent Networks: Techniques, and Applications* (1st ed.). CRC Press. <https://doi.org/10.1201/9781003541363>
- [9]. Keerthana, R., K. V., Bhagyalakshmi, K., Papinaidu, M., V. V., & Liyakat, K. K. S. (2025). Machine learning based risk assessment for financial management in big data IoT credit. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.5086671>
- [10]. KKS Liyakat, (2024b). Machine Learning (ML)-Based Braille Lippi Characters and Numbers Detection and Announcement System for Blind Children in Learning, In Gamze Sart (Eds.), *Social Reflections of Human-Computer Interaction in Education, Management, and Economics*, IGI Global. <https://doi.org/10.4018/979-8-3693-3033-3.ch002>
- [11]. Liyakat, K.K.S. (2023a). Machine Learning Approach Using Artificial Neural Networks to Detect Malicious Nodes in IoT Networks. In: Shukla, P.K., Mittal, H., Engelbrecht, A. (eds) *Computer Vision and Robotics*.



- CVR 2023. Algorithms for Intelligent Systems. Springer, Singapore.* https://doi.org/10.1007/978-981-99-4577-1_3
- [12]. Liyakat. (2024a). Machine Learning Approach Using Artificial Neural Networks to Detect Malicious Nodes in IoT Networks. In: Udgata, S.K., Sethi, S., Gao, XZ. (eds) *Intelligent Systems. ICMIB 2023. Lecture Notes in Networks and Systems*, vol 728. Springer, Singapore. https://doi.org/10.1007/978-981-99-3932-9_12 available at: https://link.springer.com/chapter/10.1007/978-981-99-3932-9_12
- [13]. Odnala, S., Shanthi, R., Bharathi, B., Pandey, C., Rachapalli, A., & Liyakat, K. K. S. (2025). Artificial Intelligence and Cloud-Enabled E-Vehicle Design with Wireless Sensor Integration. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.5107242>
- [14]. P. Neeraja, R. G. Kumar, M. S. Kumar, K. K. S. Liyakat and M. S. Vani. (2024), DL-Based Somnolence Detection for Improved Driver Safety and Alertness Monitoring. *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, 2024, pp. 589-594, doi: 10.1109/IC2PCT60090.2024.10486714. Available at: <https://ieeexplore.ieee.org/document/10486714>
- [15]. S. B. Khadake, A. B. Chounde, A. A. Suryagan, M. H. M. and M. R. Khadatare, (2024). AI-Driven-IoT(AIIoT) Based Decision Making System for High-Blood Pressure Patient Healthcare Monitoring. *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*, Theni, India, 2024, pp. 96-102, doi: 10.1109/ICSCNA63714.2024.10863954.
- [16]. Sayyad (2025b). AI-Powered IoT (AI IoT) for Decision-Making in Smart Agriculture: KSK Approach for Smart Agriculture. In S. Hai-Jew (Ed.), *Enhancing Automated Decision-Making Through AI* (pp. 67-96). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-6230-3.ch003>
- [17]. Sayyad (2025c). KK Approach to Increase Resilience in Internet of Things: A T-Cell Security Concept. In D. Darwish & K. Charan (Eds.), *Analyzing Privacy and Security Difficulties in Social Media: New Challenges and Solutions* (pp. 87-120). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-9491-5.ch005>
- [18]. Sayyad, (2025). KK Approach for IoT Security: T-Cell Concept. In Rajeev Kumar, Sheng-Lung Peng, & Ahmed Elngar (Eds.), *Deep Learning Innovations for Securing Critical Infrastructures*. IGI Global Scientific Publishing. DOI: 10.4018/979-8-3373-0563-9.ch022
- [19]. Sayyad (2025d). Healthcare Monitoring System Driven by Machine Learning and Internet of Medical Things (MLIoMT). In V. Kumar, P. Katina, & J. Zhao (Eds.), *Convergence of Internet of Medical Things (IoMT) and Generative AI* (pp. 385-416). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-6180-1.ch016>
- [20]. Shinde, S. S., Nerkar, P. M., SLiyakat, S. S., & SLiyakat, V. S. (2025). Machine Learning for Brand Protection: A Review of a Proactive Defense Mechanism. In M. Khan & M. Amin Ul Haq (Eds.), *Avoiding Ad Fraud and Supporting Brand Safety: Programmatic Advertising Solutions* (pp. 175-220). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-7041-4.ch007>
- [21]. SLiyakat, K. S. (2025j). Hydrogen Energy: Adaptation and Challenges. In J. Mabrouki (Ed.), *Obstacles Facing Hydrogen Green Systems and Green Energy* (pp. 205-236). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-8980-5.ch013>
- [22]. SilpaRaj M, Senthil Kumar R, Jayakumar K, Gopila M, Senthil kumar S. (2025). Scalable Internet of Things Enabled Intelligent Solutions for Proactive Energy Engagement in Smart Grids Predictive Load Balancing and Sustainable Power Distribution, In S. Kannadhasan et al. (eds.), *Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 24), Advances in Computer Science Research 120*, https://doi.org/10.2991/978-94-6463-718-2_85
- [23]. SLiyakat, S. (2025l). AI-Driven-IoT (AIIoT)-Based Decision Making in Drones for Climate Change: KSK Approach. In S. Aouadni & I. Aouadni (Eds.), *Recent Theories and Applications for Multi-Criteria Decision-Making* (pp. 311-340). IGI Global. <https://doi.org/10.4018/979-8-3693-6502-1.ch011>



- [24]. Upadhyaya, A. N., Surekha, C., Malathi, P., Suresh, G., Suriyan, K., & Liyakat, K. K. S. (2025). Pioneering cognitive computing for transformative healthcare innovations. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.5086894>.
- [25]. SLiyakat, S. (2024d). Computer-Aided Diagnosis in Ophthalmology: A Technical Review of Deep Learning Applications. In M. Garcia & R. de Almeida (Eds.), *Transformative Approaches to Patient Literacy and Healthcare Innovation* (pp. 112-135). IGI Global. <https://doi.org/10.4018/979-8-3693-3661-8.ch006>
Available at: <https://www.igi-global.com/chapter/computer-aided-diagnosis-in-ophthalmology/342823>
- [26]. SLiyakat, S. (2024e). IoT Driven by Machine Learning (MLIoT) for the Retail Apparel Sector. In T. Tarnanidis, E. Papachristou, M. Karypidis, & V. Ismyrlis (Eds.), *Driving Green Marketing in Fashion and Retail* (pp. 63-81). IGI Global. <https://doi.org/10.4018/979-8-3693-3049-4.ch004>
- [27]. SLiyakat, S. (2024f). Artificial Intelligence (AI)-Driven IoT (AIoT)-Based Agriculture Automation. In S. Satapathy & K. Muduli (Eds.), *Advanced Computational Methods for Agri-Business Sustainability* (pp. 72-94). IGI Global. <https://doi.org/10.4018/979-8-3693-3583-3.ch005>
- [28]. SLiyakat, K. S. (2025h). KK Approach to Increase Resilience in Internet of Things: A T-Cell Security Concept. In M. Almaiah & S. Salloum (Eds.), *Cryptography, Biometrics, and Anonymity in Cybersecurity Management* (pp. 199-228). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-8014-7.ch010>
- [29]. SLiyakat, K. S. (2025i). KK Approach for IoT Security: T-Cell Concept. In R. Kumar, S. Peng, P. Jain, & A. Elnagar (Eds.), *Deep Learning Innovations for Securing Critical Infrastructures* (pp. 369-390). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3373-0563-9.ch022>
- [30]. SLiyakat, K. S. (2025k). Roll of Carbon-Based Supercapacitors in Regenerative Breaking for Electrical Vehicles. In M. Mhadhbi (Ed.), *Innovations in Next-Generation Energy Storage Solutions* (pp. 523-572). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-9316-1.ch017>

