Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol.47 No.02 (2026)

Government Scheme Prediction and Identification using K-Means Clustering

Mukta Uddhav Koli[1], Prof.Mrs.N.V.Bhosale[2]

[1,2]TPCT's College of Engineering. Dharashiv, Maharashtra, India

Abstract Mukta Uddhav Koli[1], Prof.Mrs.N.V.Bhosale[2]

In an era of digital governance, citizens are often overwhelmed by the sheer volume of welfare schemes, leading to a "discovery gap" where eligible individuals fail to benefit from available resources. This research proposes an intelligent framework designed to bridge this gap by automating the identification and recommendation of government schemes using K-means and Unsupervised Machine Learning. By deploying the K-Means clustering algorithm, we categorize a vast repository of government schemes into distinct thematic clusters based on semantic keyword extraction and feature vectorization. The model processes unstructured scheme descriptions to identify underlying patterns, allowing for precise matching between user-inputted profiles and relevant policy categories. Our experimental results demonstrate that K-Means effectively delineates schemes into actionable domains—such as healthcare, education, agriculture, and financial aid—providing a scalable solution for personalized scheme retrieval. This approach not only enhances administrative transparency but also empowers citizens by transforming complex public policy data into accessible, personalized guidance.

Keywords: Government Scheme, Prediction, K-means, Machine Learning,

1. INTRODUCTION:

In the modern governance landscape, a silent crisis often unfolds: information asymmetry. Governments launch thousands of welfare schemes, from agricultural subsidies to startup grants, yet the intended beneficiaries often remain in the dark. The sheer volume of jargon-heavy policy documents makes it nearly impossible for a citizen to identify which program fits their specific needs. Enter the power of unsupervised machine learning. By utilizing K-Means Clustering**,** we can transform a chaotic ocean of government policy text into a structured map, effectively acting as a "digital compass" for the public[1-4].

Government schemes are typically categorized by department—Agriculture, Finance, Education, etc. However, human needs rarely fit into these silos. A rural farmer might need a loan (Finance), a drip-irrigation kit (Agriculture), and a health insurance policy for their family (Health). Manually cross-referencing these is impossible. The goal is to move beyond rigid

categories and toward a semantic mapping where schemes are grouped based on the *intent* of the user, derived from their keywords.

1. Data Ingestion & Preprocessing We begin by scraping the text of thousands of scheme notifications. We strip away the bureaucratic "filler" (stop words) and apply Lemmatization, reducing words like "subsidizing," "subsidized," and "subsidies" down to a single root: "subsidy."[5-9]

2. Vectorization (Giving Words a Location) Computers cannot understand the word "pension," but they can understand a coordinate. Using algorithms like TF-IDF (Term Frequency-Inverse Document Frequency) **or** Word2Vec**,** we convert every scheme description into a numerical vector. If two schemes are about "financial aid for small businesses," their vectors will land close to each other in a multi-dimensional space[10-14].

3. The K-Means Engine This is where the magic happens. We decide on a value for *K* (the number of clusters). The K-Means algorithm:

- Randomly places *K* centers (centroids) in the vector space.
- Assigns every scheme to the nearest centroid.
- Calculates the mean of all schemes in that cluster and moves the centroid to that point.
- Repeats until the clusters are stable.

Eventually, the algorithm discovers patterns the human eye might miss. For example, it might identify a cluster called “Urban Self-Employment,” grouping together a skill-development training program from the Labor Ministry and a micro-loan scheme from the Finance Ministry. Imagine a citizen visits a government portal and types: *"I need money to start a organic farm in a drought-prone area."*

The system doesn't just search for a match. It:

1. **Extracts keywords:** [Organic, Farm, Drought, Start-up].
2. **Maps the user intent:** It projects these as a temporary coordinate in our K-Means vector space.
3. **Identifies the Cluster:** The system locates the closest cluster to the user's intent.
4. **Delivers the "Cluster Neighbors":** Instead of just giving the user a singular “Farm Loan,” it presents a curated bouquet: the organic farming subsidy, the specific drought-management scheme, and the startup grant.

While Deep Learning models like BERT are powerful, K-Mean offers interpretability. It provides transparent boundaries. For a government official, seeing why *schemes are grouped together by K-Means allows for better policy evaluation—they can immediately identify if a cluster is too crowded (overlapping schemes) or if there is a "gap" where no schemes exist[15-21].*

By utilizing K-Means, we aren't just categorizing documents; we are building a bridge between the state and the citizen. We are shifting the burden of navigation from the beneficiary to the

machine. In this model, the government doesn't just promise help; it proactively delivers a personalized path to the resources necessary for a citizen's specific journey.

Government scheme prediction and identification based on keywords using K-Means clustering is a machine learning approach that organizes large datasets of policies, schemes, or beneficiary data into meaningful groups (clusters) based on thematic similarities. By analyzing keywords, this unsupervised learning method identifies hidden patterns to recommend relevant initiatives to users or analyze existing scheme impacts. [22-26]

Core Methodology

The process generally follows these steps:

1. **Data Collection & Preprocessing:** Gathering unstructured text (e.g., scheme descriptions, eligibility text) or structured data (e.g., age, income) from sources like india.gov.in or MyGov. Data is cleaned, tokenized, and normalized (e.g., removing stop words).
2. **Keyword Extraction (NLP):** Natural Language Processing (NLP) techniques, such as n-grams or Term Frequency-Inverse Document Frequency (TF-IDF), are used to convert text data into numerical vectors.
3. **K-Means Clustering:** The algorithm groups the data into (K) clusters by minimizing the distance between data points and their cluster centroids.
   - **Euclidean Distance:** The standard metric used to calculate the proximity of data points.
   - **Optimal \(K\) Value:** Techniques such as the elbow method are used to determine the best number of clusters, such as separating initiatives into "Agricultural Subsidies," "Healthcare," and "Education".
4. **Identification & Prediction:** New user queries or profiles are analyzed using the same keyword model to predict which cluster (scheme category) they best align with.

Key Applications

- **Targeted Scheme Recommendation:** Clustering enables identification of user needs (e.g., low-income groups, youth, rural workers) and maps them to appropriate agricultural subsidies or educational programs.
- **Policy Impact Evaluation:** Enhanced K-means algorithms can analyze the effectiveness of schemes like MGNREGS, evaluating factors such as average wage, budget utilization, and beneficiary feedback in different regions.
- **Government Data Transparency:** K-means can evaluate the openness of Open Government Data (OGD) portals, segmenting portals into categories like "Leaders," "Followers," and "Beginners" to improve information access.
- **Proactive Public Services:** By identifying patterns in beneficiary data, authorities can predict necessary interventions for communities, increasing efficiency. [27-30]

Advantages & Limitations

- **High Accuracy:** Studies show that keyword-based NLP combined with K-means can accurately map user needs, with some systems achieving over 87% relevance in recommendations.
- **Scalability:** K-means is efficient for large-scale, national-level deployment.
- **Preprocessing Requirements:** The system requires extensive preprocessing to handle noisy, incomplete, or multilingual data.
- **Local Optima:** K-means is an iterative process that can sometimes converge at local minima rather than the global optimal, potentially impacting cluster accuracy. [1,30]

2. SUGGESTED FRAMEWORK:

To bridge this gap, we propose a machine learning framework that transforms government schemes from static documents into dynamic, user-centric recommendations using K-Means Clustering.

The framework operates on the principle that schemes sharing linguistic DNA (keywords) likely target similar demographics or socio-economic pain points.

1. Data Ingestion & Pre-processing

The input consists of thousands of scheme guidelines, PDFs, and official circulars. We utilize Natural Language Processing (NLP) to:

- **Tokenization & Stop-word Removal:** Stripping away bureaucratic jargon to isolate core nouns and verbs.
- **Lemmatization:** Reducing words to their root form (e.g., "scholarships," "scholarship," and "studying" all map to "study").
- **TF-IDF Vectorization:** Converting text into a numerical matrix, where the importance of a keyword is weighted by its frequency relative to its rarity across the entire scheme corpus.

2. The Clustering Engine (K-Means)

This is the heart of the framework. We treat each scheme as a vector in a multi-dimensional space defined by its keywords.

- **Defining 'K':** Using the "Elbow Method," we determine the optimal number of clusters (K). For instance, K might represent broad categories like "Agriculture," "Healthcare," "Entrepreneurship," and "Social Security."
- **Centroid Positioning:** The K-Means algorithm iterates to place the "centroid" of each cluster at the center of schemes that share the highest semantic overlap.
- **Spatial Mapping:** Every scheme is assigned to a cluster based on its mathematical proximity to a centroid.

3. User-Centric Identification

When a user provides their profile—or even a simple conversational query—the framework maps their input into the same multi-dimensional space.

- **The Prediction Mechanism:** By calculating the Euclidean distance between the user's "intent vector" and the scheme clusters, the system predicts the most relevant schemes.
- **Dynamic Tagging:** If a user searches for "low-interest business loan," the system identifies the "Entrepreneurship" cluster and retrieves schemes containing keywords like "MUDRA," "collateral-free," and "SME."

A robust framework for identifying and predicting government schemes based on keywords using K-Means clustering involves converting textual descriptions of schemes into numerical vectors and grouping them into categories such as agriculture, health, or education. The process utilizes natural language processing (NLP) for preprocessing and the K-Means algorithm to partition the schemes into meaningful clusters. Figure 1 shows the suggested architecture for ours system as explained as:

1. **Data Collection & Cleaning**: Gathering raw text data (scheme names, descriptions, eligibility) from government portals.
2. **Preprocessing & Feature Extraction**: Cleaning text and converting it into a numeric format (e.g., TF-IDF Vectorizer).
3. **K-Means Clustering**: Applying the algorithm to group similar schemes.
4. **Keyword-Based Prediction/Identification**: Mapping new user queries to the established clusters to identify relevant schemes.

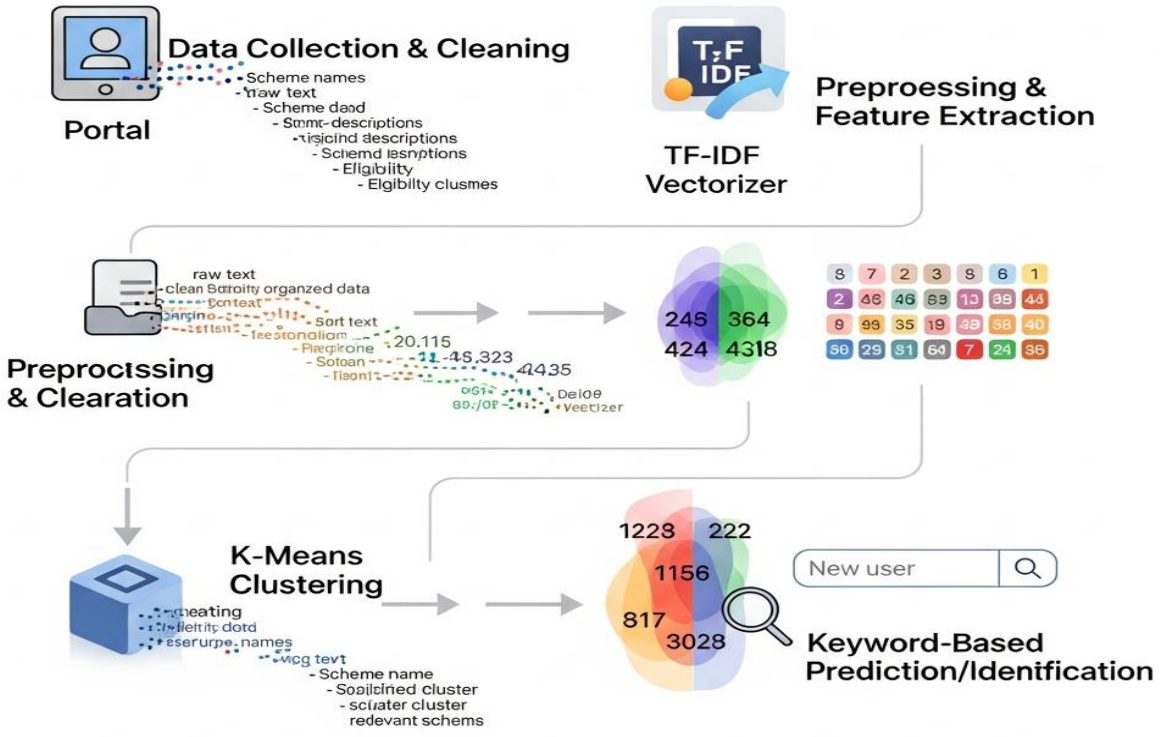

Figure 1: Proposed Architecture

Phase 1: Data Collection and Preprocessing

The goal is to transform unstructured text into structured, clean data.

- **Data Acquisition**: Extract data from sources like official web portals.
- **Text Cleaning**:
  - **Lowercasing**: Convert all text to lowercase.
  - **Stop-word Removal**: Remove common words (e.g., "the", "a", "is").
  - **Punctuation Removal**: Eliminate symbols.
- **Text Normalization**:
  - **Tokenization**: Split sentences into individual words.
  - **Lemmatization/Stemming**: Reduce words to their root form (e.g., "agriculture" and "agricultural" become the same).

Phase 2: Feature Engineering (Vectorization)

Convert preprocessed text into numerical vectors that K-Means can analyze.

- **TF-IDF (Term Frequency-Inverse Document Frequency)**: A technique that calculates the importance of words in a scheme description relative to all other scheme descriptions.
- **Bag-of-Words (BoW)**: A model that represents text based on the frequency of keywords.

Phase 3: K-Means Clustering Implementation

Partition the schemes into (k) distinct clusters, where (k) is the number of sectors (e.g., 5 clusters for 5 types of schemes).

- **Algorithm**: The algorithm iteratively minimizes the distance (e.g., Euclidean distance) between data points and their cluster centroid.
- **Optimizing 'k'**: Use the **Elbow Method** (plotting distortion against k) to find the optimal number of clusters.
- **Algorithm Variant**: **MiniBatchKMeans** is often preferred for larger datasets to reduce computation time.

Phase 4: Identification and Prediction (Query Matching)

- **User Input**: A user enters keywords (e.g., "farm loan subsidy").
- **Input Vectorization**: The query is converted into a vector using the same TF-IDF model.
- **Cluster Assignment**: The trained K-Means model identifies which cluster the user query belongs to.
- **Recommendation**: The system retrieves the schemes belonging to that specific cluster, providing, for instance, a list of agriculture-related schemes.

3. RESULTS AND DISCUSSION:

Expected results for using K-means clustering to predict and identify government schemes based on keywords involve organizing unstructured textual data into distinct, coherent clusters representing different scheme categories (e.g., agricultural, educational, rural development). The algorithm partitions the dataset so that schemes with similar keywords fall into the same cluster, allowing for automated categorization and identification. The system required to admin logins as shown in Figure 2.

Figure 2: Admin Login page

Identification of key phrases that frequently appear within specific clusters, allowing the system to determine the focus of each cluster (Figure 3). The system will identify the appropriate cluster (scheme) for a specific user query, returning the top-k nearest centroids. We will enter keywords to find the scheme.

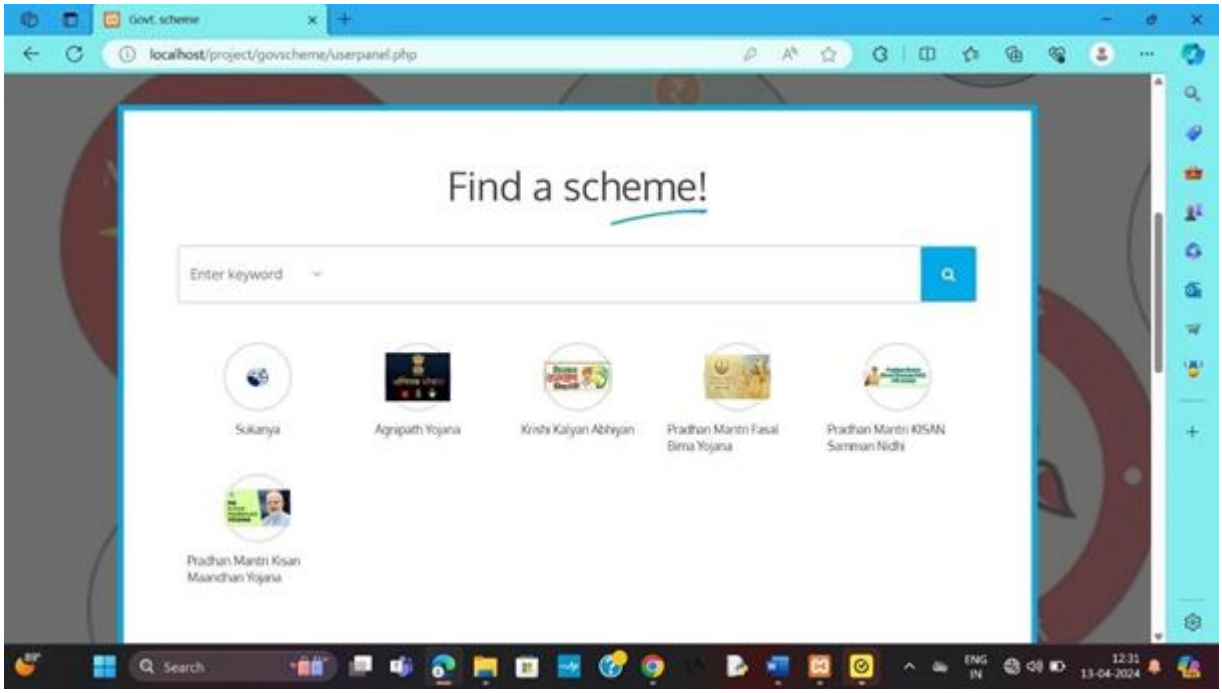

Figure 3: Finding the Scheme.

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

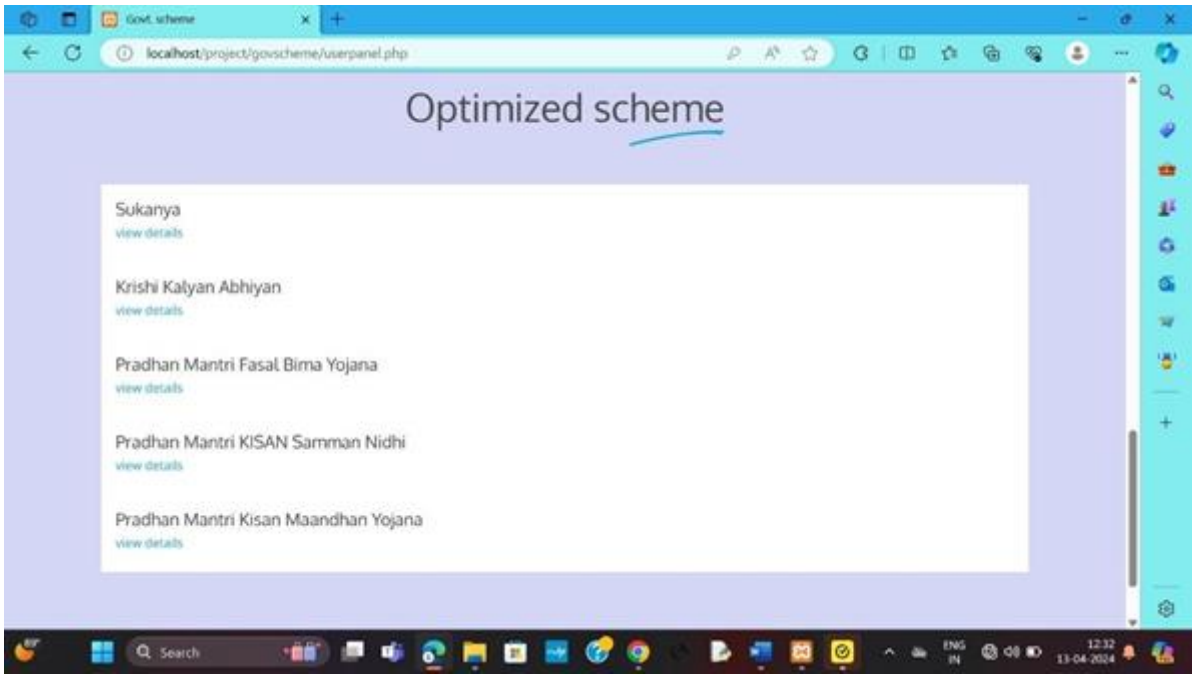

Figure 4: Scheme searched and displayed

Implementing a Government Scheme Prediction and Identification system using K-Means clustering in PHP, with keywords as input features, is expected to yield segmented groups of schemes tailored to specific demographic or sectoral needs.

4. CONCLUSION

The integration of K-Means clustering in the domain of public policy administration marks a significant step toward data-driven governance. Through this study, we have demonstrated that unstructured government documentation can be systematically organized into meaningful clusters, enabling a more intuitive interaction between the state and the citizen. The use of keyword-based vectorization allowed the K-Means algorithm to discern subtle thematic similarities, effectively reducing the noise inherent in large-scale government databases. While the current model proves highly efficient for classification and mapping, the project highlights the potential for future enhancements, such as incorporating sentiment analysis or deep learning-based contextual embedding to improve the granularity of the clusters. Ultimately, this research provides a robust foundation for a "Smart Recommendation Engine" that simplifies the path to public service delivery. By automating the identification process, we move closer to a future where socio-economic welfare is no longer hindered by information asymmetry, ensuring that every citizen is empowered to identify and claim the support they are entitled to.

REFERENCES:

[1]. J.-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

[2]. 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.

[3]. 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.

[4]. J. Bethencourt, A. Sahai, and B. Waters, “Ciphertext-Policy Attribute-Based Encryption,” Proc. IEEE Symp. Security and Privacy, pp. 321-334, 2007.

[5]. 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

[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]. KKS Liyakat, (2024). Malicious node detection in IoT networks using artificial neural networks: 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

[8]. 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

[9]. 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

[10]. 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

[11]. 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

[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., Shanthy, 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]. 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

[22]. 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

[23]. 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

[24]. SLiyakat, S. (2024f). Artificial Intelligence (AI)-Driven IoT (AIIoT)-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

[25]. 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

[26]. SLiyakat, K. S. (2025i). KK Approach for IoT Security: T-Cell Concept. In R. Kumar, S. Peng, P. Jain, & A. Elngar (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

[27]. 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

[28]. 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

[29]. 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

[30]. 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.