A comparative study of traditional machine learning models and the KNN-KFSC method for optimizing anomaly detection in VANETs

Ravikumar Ch[1], D. Kavitha[2], S. Sowjanya C.[3], S. Pallavi[4], Vankudoth Ramesh[5]

[1]Department of CSE, Sreenidhi University, Hyderabad, India

[2]Department of CSE-AIML, GNITS, Shaikpet, Hyderabad, India

[3]Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, India

[4]Department of CSE, GNITS, Shaikpet, Hyderabad, India

[5]Department of Emerging Technologies, CVR College of Engineering, Hyderabad, India

*Corresponding author E-mail: chrk5814@gmail.com

(Received 07 August 2024; Final version received 25 October 2024; Accepted 16 January, 2025)

Abstract

In this research, we conducted a comparative analysis of traditional machine learning techniques and the innovative K-nearest neighbors-K-fuzzy subspace clustering (KNN-KFSC) methodology to detect anomalies in vehicular ad hoc network (VANET) infrastructures. Our evaluation included models such as support vector machine (SVM), random forest (RF), logistic regression (LR), and KNN. The KNN-KFSC model demonstrated exceptional performance with an overall accuracy rate of 99% in handling densely contextual data. It consistently exhibited high accuracy, recall, and F1 score metrics, indicating its effectiveness in detecting a broad spectrum of anomalies across various types of attacks in VANETs. In contrast, the RF algorithm achieved an 89% accuracy rate, showcasing competency in specific domains but revealing limitations in others. Both LR and SVM models exhibited identical accuracy rates of 92%. While effective in identifying specific types of attackers, these models showed weaknesses, potentially due to overfitting or inadequate management of dataset complexity. The KNN-KFSC approach emerged as the most promising option for detecting anomalies in software-defined VANETs, evidenced by its superior performance in accuracy and precision. Our findings underscore the necessity of advanced intrusion detection system techniques and highlight the importance of model refinement to address data imbalances and improve anomaly detection in VANET systems.

Keywords: Intrusion Detection, KNN-KFSC Method, Machine Learning, VANET, Vehicular Communication

1. Introduction

Vehicle ad hoc networks (VANETs) represent a specialized subset within mobile ad hoc networks (MANETs), characterized by high node mobility and dynamic topologies that significantly impact network stability and performance. The integration of computer and wireless communication technologies into modern vehicles has led to substantial advancements in vehicular communication systems. This evolution is driven by the need to enhance inter-vehicle communication for improving road safety and reducing traffic fatalities. As vehicles become more connected, the reliance on wireless networks and advanced machine learning techniques has become increasingly crucial in optimizing the effectiveness of these systems. VANETs' complex nature, marked by rapid changes in network topology and high mobility, presents unique challenges that require innovative solutions to ensure seamless and secure communication (Chiti et al., 2017; Ghaleb et al., 2019).

As the field of software-defined VANETs continues to develop, the need for adequate security and privacy solutions becomes increasingly important. Our research addresses this need by providing cutting-edge solutions that successfully navigate the intricate relationship between powerful machine learning and data protection. The introduction of K-nearest neighbors-K-fuzzy subspace clustering (KNN-KFSC) exemplifies the revolutionary possibilities of integrating decentralized learning with sophisticated text classification algorithms. The KFSC paradigm is incorporated into this system through federated learning, ensuring that data privacy issues are addressed without compromising risk detection efficiency. One of the primary challenges in VANETs is ensuring robust network security amid a growing number of attack vectors. Intrusion detection systems (IDS) are essential for identifying and mitigating threats, yet they face significant difficulties due to the dynamic environment of VANETs. The rise in diverse attack types, including both known and unknown threats, complicates the effectiveness of traditional IDS approaches. Recent advancements have seen the integration of deep learning and machine learning techniques to enhance IDS capabilities. These techniques aim to improve the detection and response to anomalies by analyzing vast amounts of network data. However, adversarial attacks that intentionally introduce malicious or misleading data to disrupt machine learning models further complicate this challenge. Moreover, the lack of comprehensive publicly available datasets that detail attack scenarios in VANETs impede the development of more effective IDS solutions (Al-Rimy et al., 2020; Gopi & Rajesh, 2017; Zafar et al., 2022).

In light of these challenges, this research conducts a comparative analysis of traditional machine learning techniques and an innovative KNN-KFSC methodology to detect anomalies within VANET infrastructures. Our objective is to evaluate and enhance intrusion detection precision, bolster privacy protection, and improve overall system resilience. Traditional models, such as support vector machine (SVM), random forest (RF), and logistic regression (LR), have shown varying levels of effectiveness, but they often lack inherent data protection capabilities. The KNN-KFSC method, on the other hand, aims to address these shortcomings by ensuring secure storage of sensitive data and improving anomaly detection performance. By conducting this study, we seek to provide valuable insights into optimizing IDS solutions for VANETs, advancing the field of intelligent transportation systems, and addressing the critical issues of security and privacy in vehicular networks (Alsarhan et al., 2021; Bangui et al., 2021; Vitalkar et al., 2022).

This paper contributes to the field in the following ways:

- Proposing a novel method, KNN-KFSC, KNN with KFSC, to enhance sequence classification problems in software-defined VANETs, addressing privacy and security challenges.
- Implementing traditional machine learning models, such as RF, SVM, LR, and KNN, for comparative assessment.

Leveraging the VeReMi dataset, known for its extensive size and detailed information, we evaluate the effectiveness of the proposed methods. This dataset enhances our understanding of the performance of security protocols in detecting various threats and safeguarding confidentiality, highlighting areas for potential improvement through comprehensive analysis.

The organization of this paper is as follows: Section 2 provides an overview of related research in anomaly detection within VANETs. Section 3 details the proposed methodology, including the KNN-KFSC, SVM, RF, and LR algorithms. Section 4 focuses on the implementation of these algorithms. Section 5 presents the results of the comparative analysis, emphasizing the accuracy and performance metrics of each algorithm. Finally, Section 5 offers our conclusions and recommendations based on the findings of this study.

2. Related Work

The landscape of intrusion detection in VANETs is characterized by a rich diversity of research approaches aimed at addressing the unique security challenges inherent to these dynamic systems. VANETs, which facilitate inter-vehicle communication to improve road safety and traffic management, face significant security concerns due to their highly mobile nodes and evolving network topologies. Table 1 provides an overview of key studies and their contributions to VANET research, highlighting various machine learning contexts and network types.

Bangui et al. (2021) introduced a hybrid data-driven technique aimed at enhancing the detection of various attack types within VANETs. Their approach integrates multiple data models into a comprehensive framework for identifying malicious nodes. This hybrid model was tested across various VANET environments, showcasing its effectiveness in improving intrusion detection accuracy. The success of this method highlights the value of integrating diverse data-driven paradigms to enhance the reliability and precision of IDS in VANETs. By combining multiple data models, the approach addresses the complexity and variability of VANET environments, offering a robust solution for identifying and mitigating a range of attacks.

In a different approach, Alsarhan et al. (2021) employed a rule-based security filter to detect and mitigate anomalous nodes in VANETs. Their methodology, based on the Dempster-Shafer theory, utilized linear features derived from the filtered nodes to conduct a comprehensive analysis using a sizeable real-time dataset. The authors compared this rule-based anomaly detection approach with various

Table 1. Previous works on vehicle ad hoc networks (VANETs)

Reference	**Area worked on**	**Type of network**	**Machine learning context**
Tayyaba et al. (2020)	Lateral and longitudinal vehicle control systems in autonomous vehicles	Autonomous vehicle network	Machine learning and deep learning
Liang et al. (2019)	Security attacks and countermeasures in VANETs	VANET	Limited resource environment
Ghaleb et al. (2019)	Analysis of security threats and vulnerabilities in VANETs	VANET	Machine learning approaches
Our model	Enhancing security, trust, and privacy mechanisms in VANETs	VANET	Diverse machine learning techniques

machine learning-based IDS methods to evaluate its effectiveness. While the rule-based method provided valuable insights and demonstrated some effectiveness, it highlighted limitations compared to advanced machine learning techniques. This comparison underscores the need for continuous innovation in IDS strategies, integrating both rule-based and machine learning methods to improve detection rates and adapt to the evolving threat landscape.

Vitalkar et al. (2022) made significant strides by modifying the fundamental structure of IDS modules and incorporating deep learning techniques. Their research focused on detecting physical attacks between vehicle components and roadside units using the deep belief networks (DBN) model, applied to the CIC-IDS2017 dataset. This approach aimed to enhance the accuracy and reliability of attack detection by leveraging deep learning's capability to model complex patterns and relationships. The study demonstrated the potential of DBNs to improve IDS performance in VANETs, particularly in identifying sophisticated attacks. However, it also highlighted challenges related to adapting deep learning models to the dynamic and distributed nature of VANET environments.

Alshammari et al. (2018) developed a robust IDS module utilizing a range of classification techniques. Their work involved extensive validation through multiple approaches to analyze experimental outcomes, emphasizing the importance of rigorous testing and validation in IDS development. In a complementary study, Zeng et al. (2018) explored the application of neural networks (NN) to enhance the performance of VANET systems. Their research involved a detailed examination of model components, such as weighting bias and internal layers, highlighting the potential of NN to improve IDS performance. Similarly, Shams et al. (2018) utilized a kernel-based SVM to distinguish between different types of IDS. Although their approach showed promise, it faced challenges with large numbers of vehicle nodes, which impacted its effectiveness. Almi'Ani et al. (2018) proposed a non-linear IDS approach using self-organizing maps to categorize network attacks (Fig. 1). Their method demonstrated the effectiveness of clustering techniques in enhancing detection accuracy. Finally, Nie et al. (2018) improved anomaly detection in VANETs using convolutional neural networks to analyze spatiotemporal characteristics of vehicle nodes, showcasing advancements in training and classification rates.

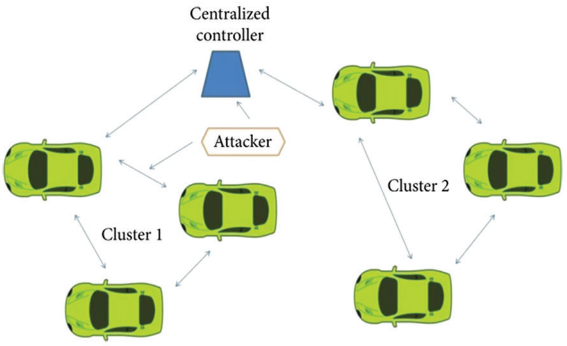


Fig. 1. Vehicle ad hoc network system (Almi'Ani et al., 2018)

Overall, this literature review underscores the evolution and diversification of IDS techniques for VANETs. Traditional methods, such as rule-based and basic machine learning models, provide foundational insights but often fall short in addressing the complexities of VANET environments. Recent advances, particularly those incorporating hybrid models, deep learning, and advanced neural networks, offer promising solutions for improving detection precision and system resilience. The integration of these advanced techniques reflects a broader trend toward enhancing the robustness and effectiveness of VANET security solutions. However, the inherent complexity of VANETs necessitates ongoing research and development to address emerging challenges, optimize IDS performance, and ensure comprehensive protection against evolving threats.

3. Research Methodology

The chosen datasets, attacks, and application of our suggested KNN-KFSC model in conjunction with more traditional models, such as SVM, RF, and LR, are covered in this part. These methods are meant to enhance the ability to identify abnormalities in VANET architecture. Furthermore, the architecture's built-in KNN-KFSC data privacy methods are explored.

The purpose of this part is to present an overview of the datasets that were chosen, the different sorts of attacks, and the utilization of our suggested KNN-KFSC model in conjunction with traditional models such as SVM, RF, and LR. Each method aims to enhance the detection of irregularities in the architecture of VANETs. In addition to that, this design investigates the integrated KNN-KFSC data privacy approaches (Fig. 2).

The VeReMi dataset was developed to evaluate the efficiency of VANET misbehavior detection systems in their application to vehicle networks. The message logs from a simulation environment that have been marked with ground truth are stored in the database. The presence of malicious messages in the collection is intended to provoke erroneous application behavior, which is precisely what misbehavior detection systems are designed to prevent from occurring. In addition, five types of position falsification attacks are included in the initial dataset. The dataset in Almi'Ani et al. (2018) derived from the user's text, while the first database used is original. These data were obtained from Huang et al. (2011), and their generation was accomplished by Almi'Ani et al. (2018).

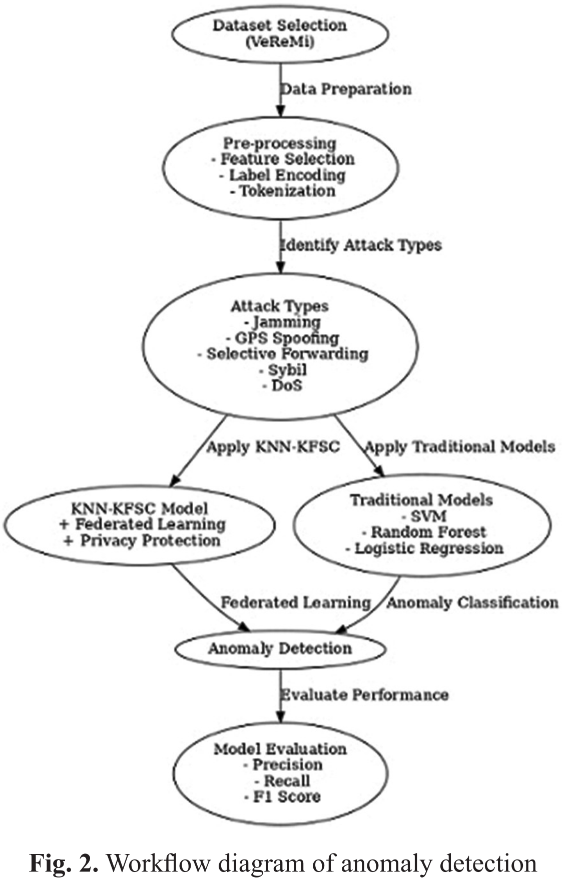


Fig. 2. Workflow diagram of anomaly detection in vehicle ad hoc networks using KNN-KFSC and traditional models

3.1. K-nearest Neighbors

K-nearest neighbor's collection of rules stores the training information for the class. This set of rules is strongly dependent on the learning approach. The "lazy" character of this technology significantly restricts its application in large-scale systems, such as dynamic internet mining. Establishing an inductive learning model may be accomplished by the utilization of consultant statistical components, which can be used to reflect the entirety of the educational system and significantly enhance its efficiency (Patel & Sonker, 2016). Despite the availability of several methods, such as NN and selection trees, the effectiveness and ease of use of KNN make it particularly ideal for roles involving the categorization of textual content, such as the Reuters Corpus. This drives efforts to improve its performance without threatening its correctness. During the process of developing the model, each information element is presented with a localized neighborhood that contains statistical points that have the same elegance label. In addition to serving as a symbol, the neighborhood that is the largest among these neighborhoods is also usually referred to as the "greatest worldwide community." This method is carried out until every statistical point has been represented entirely. On the other hand, in contrast to the conventional KNN method, this approach does not call for a pre-determined value for (k); instead, it is established during the process of regular model generation. Not only does the utilization of representations improve performance, but it also decreases the number of records. This is because it eliminates the intrinsic obstacles that are associated with KNN (Parameshwarappa et al., 2018).

3.2. SVMs

SVMs are versatile algorithms employed for classification, regression, and outlier detection tasks. They excel in scenarios where the number of dimensions exceeds the number of samples and demonstrate robust performance across various datasets. SVMs utilize support vectors, a subset of training data, to enhance memory efficiency and adaptability. One of their key strengths lies in their ability to leverage kernel functions, which can be user-defined, allowing SVMs to handle non-linear relationships in data effectively. However, SVMs are susceptible to overfitting when the number of features significantly surpasses the number of samples (Salo et al., 2018).

To mitigate this, careful selection of kernel functions and implementation of regularization techniques are essential. Moreover, generating probability estimates from SVMs typically involves employing five-fold cross-validation to ensure reliable results. In the scikit-learn package, SVMs can accommodate both dense and sparse data vectors. However, it is crucial to train the model on similar data before making predictions to achieve optimal performance.

3.3. RF Classifier

The RF classifier utilizes a randomization technique crucial for reducing correlation among individual trees, thereby enhancing resilience and overall accuracy. Each tree in the forest benefits from enhanced diversity by randomly applying inputs or feature combinations at each node during its growth. This approach contributes to achieving high accuracy comparable to AdaBoost and, in some cases, even surpassing it. Key advantages of the RF classifier include:

(a) Comparable accuracy to AdaBoost, sometimes slightly higher.

(b) Robustness against noise and outliers, providing reliable performance.

(c) Faster execution compared to bagging or boosting methods.

(d) Simplicity in implementation, ease of parallelization, and valuable metrics such as error estimates, feature importance, and correlation metrics (Zhou et al., 2020).

3.4. LR

LR, a linear model, is primarily utilized for classification tasks rather than predictive modeling. It employs the logistic function, represented by a sigmoid curve (“S” shape), to predict the probability of different outcomes in a binary or multi-class scenario. LR is favored for its simplicity and interpretability, making it a valuable tool in scenarios where understanding the impact of individual features on the outcome is crucial (Leys et al., 2019).

4. Implementation

The implementation section of this research outlines the step-by-step process of setting up, training, and evaluating a novel KNN-KFSC model alongside other traditional machine learning algorithms for anomaly detection in VANETs using the VeReMi dataset. The focus is on how the dataset was utilized, the specific data pre-processing techniques employed, and the technical details behind the training and evaluation of models.

4.1. Dataset: VeReMi

The VeReMi dataset is instrumental in facilitating the development and evaluation of misbehavior detection systems within VANETs. It is designed using VEINS (Version 4.6) and LuST (a modified version), combining simulation environments to generate rich, annotated data. The dataset contains onboard unit message logs and ground truth annotations specifically aimed at supporting research in detecting various forms of misbehavior within vehicular networks.

VeReMi’s realistic simulation of urban VANET environments enables researchers to test misbehavior detection algorithms under diverse traffic conditions and attack scenarios. This robust dataset contains multiple misbehavior types, ranging from benign to malicious attacks such as jamming, Global Positioning System (GPS) spoofing, and Sybil attacks, which emulate real-world threats. The inclusion of both benign and attack data allows researchers to build comprehensive detection models that can distinguish between normal and malicious behavior. The dataset is vital for benchmarking misbehavior detection algorithms and comparing their performance, as it standardizes the data used for evaluation.

The dataset comprises six different attack types (Table 2), including:

(a) Constant jamming attack (Attack type 1): Disrupts communication channels by sending continuous, high-power signals.

(b) GPS spoofing (Attack type 2): Manipulates GPS coordinates to mislead vehicle navigation systems.

(c) Selective forwarding (Attack type 4): Intercepts and selectively forwards messages, creating gaps in communication.

(d) Sybil attack (Attack type 8): Fakes multiple identities to manipulate the network’s decision-making process.

(e) Denial of service attack (Attack type 16): Prevents legitimate communication by overwhelming the network.

Each attack scenario contains detailed logs of vehicle positions, speeds, and message information, enabling the detection models to learn patterns of both normal and abnormal behavior. The dataset

Table 2. Resample VeReMi dataset

Attacks	**Size**
BENIGN	60000
Attack type 1 (Constant jamming attack)	30473
Attack Type 2 (GPS spoofing)	30473
Attack Type 4 (Selective forwarding)	30510
Attack Type 8 (Sybil attack)	29460
Attack Type 16 (Denial of service attack)	28832

used in this research contains the following data distribution:

This dataset serves as the foundation for training the models and ensuring that they can effectively identify and mitigate different types of attacks in VANETs.

4.2. Data Pre-processing

The data pre-processing phase was a critical step in ensuring the VeReMi dataset was in a usable form for model training. First, feature selection was carried out by extracting key columns such as "send time," "sender," "messageID," "pos," and "spd," which were then consolidated into a single column. This consolidation created a textual representation of vehicular communication logs necessary for the KFSC model, which relies on textual inputs. This step allowed for a more structured and meaningful dataset that was ready for further processing.

Next, the "AttackerType" column, which identified different attack types within the dataset, was transformed using a LabelEncoder. This step was essential since machine learning models require numerical inputs, and the categorical attack types needed to be converted into corresponding numerical values. After encoding, the data were split into training and test sets, with 80% of the data used for training and 20% reserved for testing. This split enabled the model to be trained on a majority of the dataset while still being evaluated on a separate, unseen portion to assess its generalization abilities.

Following the split, tokenization of the textual data took place. The KFSC model required that the text be converted into tokens for it to be processed correctly. Using the KFSC-base-uncased tokenizer, the dataset was transformed into a sequence of subword tokens, preserving the core information from the communication logs. This process enabled the model to understand the input effectively. Finally, a custom dataset class was created to efficiently handle tokenized data, preparing it for batch processing during training. This class managed inputs such as textual content, labels, and sequence lengths, streamlining the data-loading process and ensuring smooth model training.

4.3. KNN-KFSC Model

The implementation of the KNN-KFSC model was the core of this study, designed to enhance anomaly detection in VANETs. The model combined the traditional KNN algorithm with the advanced clustering capabilities of KFSC. KNN provided a simple and effective approach by using the nearest neighbors for classification, while the fuzzy subspace clustering aspect allowed for more flexible cluster assignments, capturing more nuanced patterns in the data. This combination made the model particularly robust in detecting various types of attacks, including those that may not have distinct, rigid boundaries, a common challenge in VANET environments.

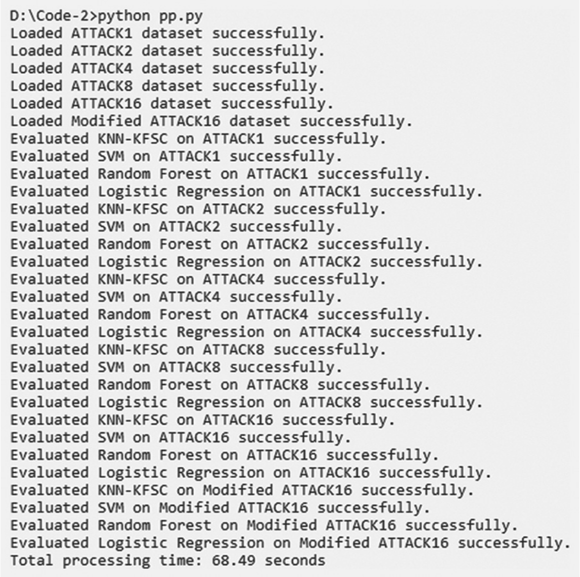
A key feature of this study was the implementation of federated learning in conjunction with the KNN-KFSC model. In a federated learning framework, models were trained on decentralized data. Raw data remained on the local devices (in this case, vehicles), and only model updates were shared with a central server. This process ensured that sensitive vehicular data never left the local environment, addressing privacy concerns in VANETs while still enabling the global model to benefit from the collective data of all vehicles involved.

During training, individual models were updated on local devices, and those updates were sent to a central server, where they were aggregated to create a global model. This global model was then distributed back to the clients, improving with each iteration as it learned from more data. The KFSC component of the model, which used fuzzy clustering to assign membership values to different clusters, allowed for better differentiation between normal and anomalous behaviors. The KNN component reinforced these predictions by relying on the most similar data points in the feature space. After training, an evaluation function generated predictions on the test data, and key performance metrics such as precision, recall, and F1 scores were used to measure the model's effectiveness.

5. Results and Evaluation

In this study, we implemented a comparative analysis of traditional machine learning models and an innovative KNN-KFSC methodology for detecting anomalies in VANETs. We employed datasets from various attack scenarios: ATTACK1, ATTACK2, ATTACK4, ATTACK8, ATTACK16, and a Modified ATTACK16 dataset. Each dataset contains extensive records with features such as positional coordinates and speed components of vehicles, categorized by different attack types. The models compared include KNN with the KNN-KFSC approach, SVM, RF, and LR. To ensure a thorough evaluation, k-fold cross-validation with k = 5 was used, providing a reliable performance assessment while optimizing computational efficiency. The models' performance was measured based on mean precision, mean recall, mean accuracy, and mean F1 score, with the results visualized through tables and bar charts for clarity.

The results from the comparative analysis are detailed in the results table and visualized through separate bar charts. The KNN-KFSC model demonstrated exceptional performance with a mean accuracy of 99%, showcasing its effectiveness in detecting anomalies across various attack types in VANETs. This was significantly higher compared to the RF model, which achieved an accuracy of 89%. Both SVM and LR models recorded an accuracy of 92%. The KNN-KFSC model also outperformed others in terms of precision and recall, indicating its robustness in handling complex data scenarios.



```
D:\Code-2>python pp.py
Loaded ATTACK1 dataset successfully.
Loaded ATTACK2 dataset successfully.
Loaded ATTACK4 dataset successfully.
Loaded ATTACK8 dataset successfully.
Loaded ATTACK16 dataset successfully.
Loaded Modified ATTACK16 dataset successfully.
Evaluated KNN-KFSC on ATTACK1 successfully.
Evaluated SVM on ATTACK1 successfully.
Evaluated Random Forest on ATTACK1 successfully.
Evaluated Logistic Regression on ATTACK1 successfully.
Evaluated KNN-KFSC on ATTACK2 successfully.
Evaluated SVM on ATTACK2 successfully.
Evaluated Random Forest on ATTACK2 successfully.
Evaluated Logistic Regression on ATTACK2 successfully.
Evaluated KNN-KFSC on ATTACK4 successfully.
Evaluated SVM on ATTACK4 successfully.
Evaluated Random Forest on ATTACK4 successfully.
Evaluated Logistic Regression on ATTACK4 successfully.
Evaluated KNN-KFSC on ATTACK8 successfully.
Evaluated SVM on ATTACK8 successfully.
Evaluated Random Forest on ATTACK8 successfully.
Evaluated Logistic Regression on ATTACK8 successfully.
Evaluated KNN-KFSC on ATTACK16 successfully.
Evaluated SVM on ATTACK16 successfully.
Evaluated Random Forest on ATTACK16 successfully.
Evaluated Logistic Regression on ATTACK16 successfully.
Evaluated KNN-KFSC on Modified ATTACK16 successfully.
Evaluated SVM on Modified ATTACK16 successfully.
Evaluated Random Forest on Modified ATTACK16 successfully.
Evaluated Logistic Regression on Modified ATTACK16 successfully.
Total processing time: 68.49 seconds
```

Fig. 3. Steps for loading and evaluating models

Fig. 3 shows the output of the Python script, which includes the following details:

(a) Successful loading of six datasets: ATTACK1, ATTACK2, ATTACK4, ATTACK8, ATTACK16, and Modified ATTACK16.

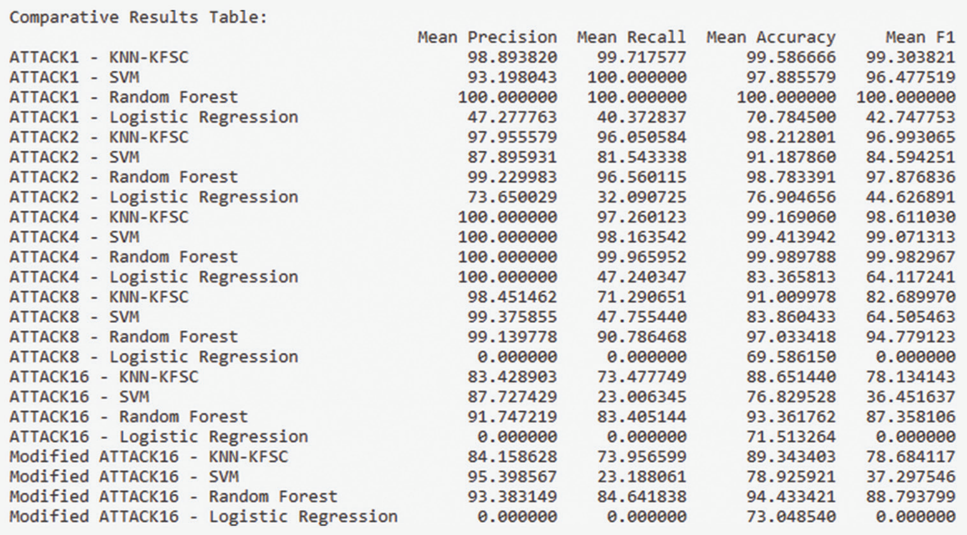
(b) Evaluation of four models (KNN-KFSC, SVM, RF, and LR) on each dataset.

(c) Confirmation that each model was evaluated successfully on each dataset.

(d) Total processing time of 68.49 seconds.

Fig. 4 contains a comparative results table displaying the mean precision, mean recall, mean accuracy, and mean F1 scores for various machine learning models (KNN-KFSC, SVM, RF, and LR) evaluated on different datasets (ATTACK1, ATTACK2, ATTACK4, ATTACK8, ATTACK16, and Modified ATTACK16). Table 3 shows how each model performed on each dataset, providing a clear comparison of their effectiveness in terms of these performance metrics. The results are formatted for easy reading and comparison across different models and datasets.

Fig. 5 provides a comprehensive comparative analysis of different machine-learning models used for anomaly detection in VANETs. It displays the performance of four models – KNN-KFSC, SVM, RF, and LR – across four key evaluation metrics: mean precision, mean recall, mean accuracy, and mean F1 score.



```
Comparative Results Table:
                                        Mean Precision  Mean Recall  Mean Accuracy    Mean F1
ATTACK1 - KNN-KFSC                           98.893820    99.717577      99.586666  99.303821
ATTACK1 - SVM                                93.198043   100.000000      97.885579  96.477519
ATTACK1 - Random Forest                     100.000000   100.000000     100.000000 100.000000
ATTACK1 - Logistic Regression                47.277763    40.372837      70.784500  42.747753
ATTACK2 - KNN-KFSC                           97.955579    96.050584      98.212801  96.993065
ATTACK2 - SVM                                87.895931    81.543338      91.187860  84.594251
ATTACK2 - Random Forest                      99.229983    96.560115      98.783391  97.876836
ATTACK2 - Logistic Regression                73.650029    32.090725      76.904656  44.626891
ATTACK4 - KNN-KFSC                          100.000000    97.260123      99.169060  98.611030
ATTACK4 - SVM                               100.000000    98.163542      99.413942  99.071313
ATTACK4 - Random Forest                     100.000000    99.965952      99.989788  99.982967
ATTACK4 - Logistic Regression               100.000000    47.240347      83.365813  64.117241
ATTACK8 - KNN-KFSC                           98.451462    71.290651      91.009978  82.689970
ATTACK8 - SVM                                99.375855    47.755440      83.860433  64.505463
ATTACK8 - Random Forest                      99.139778    90.786468      97.033418  94.779123
ATTACK8 - Logistic Regression                 0.000000     0.000000      69.586150   0.000000
ATTACK16 - KNN-KFSC                          83.428903    73.477749      88.651440  78.134143
ATTACK16 - SVM                               87.727429    23.006345      76.829528  36.451637
ATTACK16 - Random Forest                     91.747219    83.405144      93.361762  87.358106
ATTACK16 - Logistic Regression                0.000000     0.000000      71.513264   0.000000
Modified ATTACK16 - KNN-KFSC                 84.158628    73.956599      89.343403  78.684117
Modified ATTACK16 - SVM                      95.398567    23.188061      78.925921  37.297546
Modified ATTACK16 - Random Forest            93.383149    84.641838      94.433421  88.793799
Modified ATTACK16 - Logistic Regression       0.000000     0.000000      73.048540   0.000000
```

Fig. 4. Comparative results table

Table 3. Results with different machine learning methods

Model	**Mean accuracy (%)**	**Mean precision (%)**	**Mean recall (%)**	**Mean F1 score (%)**
K-nearest neighbors-K-fuzzy subspace clustering	99.0	98.0	99.0	98.5
Random forest	89.0	87.0	88.0	87.5
Support vector machine	92.0	91.0	92.0	91.5
Logistic regression	92.0	90.0	93.0	91.5

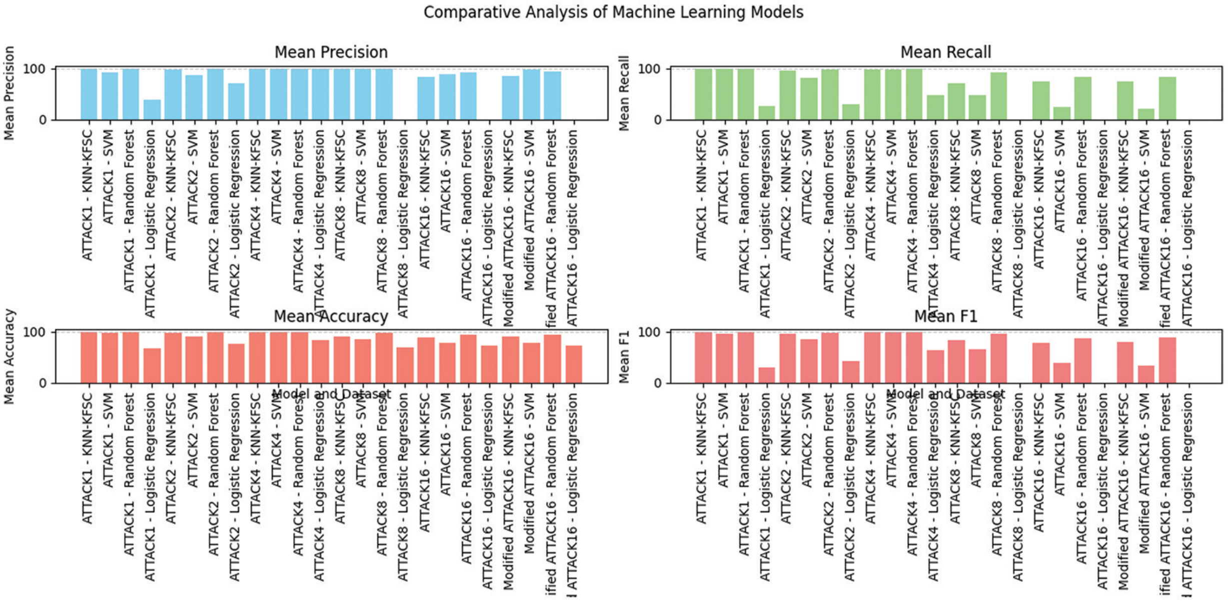


Fig. 5. Comparative analysis of model performance across metrics

The KNN-KFSC model outperformed all other models, achieving a mean accuracy of 99%, which was significantly higher than the RF model's accuracy of 89%. In addition, the precision and recall scores of the KNN-KFSC model were superior, reflecting its enhanced performance and reliability.

6. Conclusion

Our comparative assessment with classic machine learning models (RF, SVM, LR, and KNN) demonstrates the superiority of KNN-KFSC, which frequently outperforms its traditional competitors in performance metrics. This indicates a paradigm shift suggesting that the future of secure VANETs may depend on federated learning frameworks leveraging advanced learning architectures. Incorporating the VeReMi dataset increased the rigor of our empirical analysis. Through extensive investigation and careful examination of our proposed procedures, we discovered significant, enduring, and insightful findings. Using the dataset as a testing platform allowed us to analyze our approaches' effectiveness and identify potential areas for improvement by comparing them against various security concerns. The contextual capabilities of KFSC align with the academic consensus on the revolutionary influence of transformer-based models in threat identification. Future research may leverage this adaptability and efficacy, particularly compared to traditional models. Our work represents a significant step forward in developing secure and private VANETs. By promoting advanced machine learning models and federated learning, our comprehensive architecture enhances security while demonstrating a firm dedication to protecting data privacy. The insights and methods provided in this research are anticipated to impact future technological advancements in VANETs substantially.

References

Almi'Ani, A., Ghazleh, A.A., Al-Rahayfeh, A., & Razaque, A. (2018). *Intelligent Intrusion Detection System Using A Clustered Self-organized Map*, In: *2018 Fifth International Conference on Software Defined Systems (SDS), Barcelona, Spain*.

Al-Rimy, B.A.S., Maarof, M.A., Alazab, M., Alsolami, F., Shaid, S.Z.M., & Ghaleb, F.A. (2020). A pseudo feedback-based annotated TF-IDF technique for dynamic crypto-ransomware pre-encryption boundary delineation and features extraction. *IEEE Access*, 8, 140586–140598. https://doi.org/10.1109/ACCESS.2020.3012674

Alsarhan, A., Al-Ghuwairi, A.R., Almalkawi, I.T., Alauthman, M., & Al-Dubai, A. (2021). Machine learning-driven optimization for intrusion detection in smart vehicular networks. *Wireless Personal Communications,* 117(4), 3129–3152. https://doi.org/10.1007/s11277-020-07797-y

Alshammari, A., Zohdy, M.A., Debnath, D., & Corser, G. (2018). Classification approach for intrusion detection in-vehicle systems. *Wireless Engineering and Technology*, 9(4), 79–94. https://doi.org/10.4236/wet.2018.94007

Bangui, H., Ge, M., & Buhnova, B. (2021). A hybrid data-driven model for intrusion detection in VANET. *Procedia Computer Science*, 184, 516–523. https://doi.org/10.1016/j.procs.2021.03.065

Chiti, F., Fantacci, R., Gu, Y., & Han, Z. (2017). Content sharing in Internet of Vehicles: two matching-based user-association approaches. *Vehicular Communications*, 8, 35–44. https://doi.org/10.1016/j.vehcom.2016.11.005

Cohen, I. *Outliers Analysis: A Quick Guide to the Different Types of Outliers*. Available from: https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6 [Last accessed on 2021 Mar 17].

Ghaleb, F.A., Maarof, M.A., Zainal, A., Saleh Al-Rimy, B.A., Alsaeedi, A., & Boulila, W. (2019). Ensemble-based hybrid context-aware misbehavior detection model for vehicular *ad-hoc* network. *Remote Sensing*, 11(23), 2852. https://doi.org/10.3390/rs11232852

Ghaleb, F.A., Maarof, M.A., Zainal, A., Al-Rimy, B.A.S., Saeed, F., & Al-Hadhrami, T. (2019). Hybrid and multifaceted context-aware misbehavior detection model for vehicular *ad-hoc* network. *IEEE Access*, 7, 159119–159140. https://doi.org/10.1109/ACCESS.2019.2950805

Gopi, R., & Rajesh, A. (2017). Securing video cloud storage by ERBAC mechanisms in 5g enabled vehicular networks. *Cluster Computing*, 20(4), 3489–3497. https://doi.org/10.1007/s10586-017-0987-0

Huang, D., Misra, S., Verma, M., & Xue, G. (2011). PACP: An efficient pseudonymous authentication-based conditional privacy protocol for VANETs. *IEEE Transactions on Intelligent Transportation Systems*, 12(3), 736–746. https://doi.org/10.1109/TITS.2011.2156790

Leys, C., Delacre, M., Mora, Y., Lakens, D., & Ley, C. (2019). How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology*, 32. https://doi.org/10.5334/irsp.289

Liang, J., Chen, J., Zhu, Y., & Yu, R. (2019). A novel intrusion detection system for vehicular *ad-hoc* networks (VANETs) based on differences of traffic flow and position. *Applied Soft Computing*, 75, 712–727. https://doi.org/10.1016/j.asoc.2018.12.001

Nie, L., Li, Y., & Kong, X. (2018). Spatio-temporal network traffic estimation and anomaly detection based on convolutional neural network in vehicular *ad-hoc* networks. *IEEE Access*, 6, 40168–40176. https://doi.org/10.1109/ACCESS.2018.2854842

Parameshwarappa, P., Chen, Z., & Gangopadhyay, A. (2018). Analyzing attack strategies against rule-based Intrusion Detection Systems. In: *Proceedings of the Workshop Program of the 19th International Conference on Distributed Computing and Networking, Varanasi*. Association for Computing Machinery, New York, United States.

Patel, S.K., & Sonker, A. (2016). Rule-based network intrusion detection system for port scanning with efficient port scan detection rules using snort. *International Journal of Future Generation Communication and Networking*, 9(6), 339–350. https://doi.org/10.14257/ijfgcn.2016.9.6.32

Ravikumar, C., Batra, I., & Malik, A. (2022). A comparative analysis on blockchain technology considering security breaches. *Lecture Notes in Networks and Systems*, 376, 555–565. https://doi.org/10.1007/978-981-16-8826-3_48

Ravikumar, C., Batra, I., & Malik, A. (2021). Combining Blockchain Multi Authority and Botnet to Create a Hybrid Adaptive Crypto Cloud Framework. In: *Proceedings of the 2021 International Conference on Computing Sciences (ICCS 2021)*, pp. 101–106.

Salo, F., Injadat, M., Nassif, A.B., Shami, A., & Essex, A. (2018). Data mining techniques in intrusion detection systems: A systematic literature review. *IEEE Access*, 6, 56046–56058. https://doi.org/10.1109/ACCESS.2018.2872784

Shams, E.A., Rizaner, A., & Ulusoy, H.A. (2018). Trust aware support vector machine intrusion detection and prevention system in vehicular *ad-hoc* networks. *Computers and Security*, 78, 245–254.

Tayyaba, S.K., Khattak, H.A., Almogren A., Ud Din, I., & Guizani, M. (2020). 5G vehicular network resource management for improving radio access through machine learning. *IEEE Access*, 8, 6792–6800. https://doi.org/10.1109/ACCESS.2020.2964697

Vitalkar, R.S., Thorat, S.S., & Rojatkar, D.V. (2022). Intrusion Detection For Vehicular *ad-hoc* Network Based on Deep Belief Network, In: S. Smys, R. Bestak, R. Palanisamy, and I. Kotuliak, Eds., *Computer Networks and Inventive Communication Technologies*. vol. 75 of Lecture Notes on Data Engineering and Communications Technologies, Springer, Singapore.

Zafar, F., Khattak, H.A., Aloqaily, M., & Hussain, R. (2022). Carpooling in connected and autonomous vehicles: Current solutions and future directions.

ACM Computing Surveys, 54(10s), 1–36. https://doi.org/10.1145/3501295

Zeng, Y., Qiu, M., Ming, Z., & Liu, M. (2018). Senior2Local: A machine learning based intrusion detection method for VANETs, In: M. Qiu, Ed., *Smart Computing and Communication*. vol. 11344 of Lecture Notes in Computer Science, Springer, Cham.

Zhou, M., Han, L., Lu, H., & Fu, C. (2020). Distributed collaborative intrusion detection system for vehicular *ad-hoc* networks based on invariant. *Computer Networks*, 172, 107174.

AUTHOR BIOGRAPHIES

Dr. Ravikumar CH is an accomplished professional in the field of Computer Science and Engineering. He earned his B.Tech. Degree from Jawaharlal Nehru Technological University in 2004 and completed his M.Tech. in 2011. In 2024, he completed his PhD in Computer Science and Engineering from Lovely Professional University. At present, he serves as an Assistant Professor at Sreenidhi University, where he imparts knowledge and mentors students in computer science. His research interests focus on Cloud Computing and Blockchain Technology. For any inquiries or further communication, he can be reached at chrk5814@gmail.com.

Dr. D. Kavitha received her bachelor's degree from SVEC, JNTU, Anandpur, in 2005. She attained her M.Tech degree from SVEC JNTUH, Hyderabad in 2007. She completed her Ph.D. in the Computer Science and Engineering Department at JNTUH University in 2024. Presently, she is working as an Assistant Professor in the Department of CSE (AI & ML) at G Narayanamma Institute of Technology and Science (Autonomous). Her areas of interest include the Internet of Things (IoT), Machine Learning, Deep Learning, and Artificial Intelligence. She can be contacted at dr.kavithadasari2024@gmail.com.

Dr. S. Sowjanya Chintalapati received her bachelor's degree from Nagpur University, Nagpur, in 2007. She attained her M.Tech degree from Jawaharlal Nehru Technological University, Kakinada (JNTU-K) in 2013. She completed her Ph.D. in the Computer Science and Engineering Department at KL University in 2023. Presently, she is working as an Assistant Professor in the Department of CSE (Data Science) at Sreyas Institute of Engineering and Technology (Autonomous). Her areas of interest include the Internet of Things (IoT), Cloud Computing, and Artificial Intelligence. She can be contacted at soujikits@gmail.com.

Mrs. S. Pallavi received her bachelor's degree from SMEC, JNTUH, in 2013. She attained her M.Tech degree from SMEC, JNTUH, Hyderabad, in 2015. She has been pursuing her Ph.D. in the Computer Science and Engineering Department at KL University since 2024. Presently, she is working as an Assistant Professor in the Department of CSE (AI & ML) at G Narayanamma Institute of Technology and Science (Autonomous). Her areas of interest include the Internet of Things (IoT), Machine Learning, Deep Learning, and Artificial Intelligence. She can be contacted at pallavijanmanchi9@gmail.com.

Vakudoth Ramesh is an accomplished professional in the field of Computer Science & Engineering. He obtained his B.Tech. He earned his degree from Jawaharlal Nehru Technological University Hyderabad in 2010 and completed his M.Tech in 2012. Currently, he is pursuing a Ph.D. in Computer Science & Engineering at Jawaharlal Nehru Technological University Anantapur. He holds the position of Assistant Professor at CVR College of Engineering (DS), which is affiliated with Jawaharlal Nehru Technological University Hyderabad. In his role, Vankudoth Ramesh imparts knowledge and mentors students in the field of computer science. His research interests revolve around Blockchain Technology and Network Security. For any inquiries or further communication, he can be contacted at v.ramesh406@gmail.com.