Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Performance Analysis of Traditional ML Algorithms in High-Dimensional Data with Fuzzy Feature Selection

A. Venu Madhavi, T. Tejaswi

Assistant Professor, Department of AI & DS, Sri Indu College of Engineering and Technology, Hyderabad, India

Correspondence

A.Venu Madhavi

Assistant Professor, Department of AI & DS, Sri Indu College of Engineering and Technology, Hyderabad, India

Abstract

This study investigates the impact of fuzzy feature selection on the performance of traditional machine learning algorithms in high-dimensional data scenarios. We evaluated several algorithms, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forests, and K-Nearest Neighbors (KNN), using both standard and fuzzy feature selection methods. The results reveal a substantial improvement in predictive performance with the application of fuzzy feature selection. Specifically, accuracy, precision, recall, and F1-score metrics showed notable enhancements across all algorithms, with the most significant gains observed in SVM and Random Forests. The findings suggest that fuzzy feature selection effectively addresses the challenges associated with high-dimensional data by reducing dimensionality and improving signal quality, leading to more robust and accurate models.

- Received Date: 30 Dec 2024
- Accepted Date: 15 Feb 2025
- Publication Date: 01 Mar 2025

Introduction

High-dimensional data, characterized by an extensive number of features or variables, presents unique challenges in the field of machine learning. As the dimensionality of data increases, the complexity of the model also rises, often leading to several issues such as the curse of dimensionality, overfitting, and computational inefficiency. The curse of dimensionality refers to the phenomenon where the volume of the feature space grows exponentially with the number of dimensions, causing the data points to become sparse[1]. This sparsity makes it difficult for machine learning models to identify meaningful patterns, as the models require exponentially more data to generalize accurately in high-dimensional spaces. Additionally, the presence of numerous irrelevant or redundant features can lead to overfitting, where the model performs well on training data but fails to generalize to unseen data. This not only reduces the accuracy of the model but also increases the computational cost, making the learning process more resource-intensive and time-consuming[2].

In this context, effective feature selection becomes crucial for improving the performance of machine learning algorithms. Feature selection involves identifying and selecting a subset of relevant features that contribute the most to the predictive power of the model, while discarding irrelevant or redundant ones. By reducing the dimensionality of the data, feature selection helps to mitigate the curse of dimensionality, improve model interpretability, and enhance computational efficiency. Traditional feature selection techniques, however, often struggle with high-dimensional data, as they may overlook the complex relationships between features or fail to account for uncertainty and imprecision inherent in real-world data. This is where fuzzy logic-based feature selection methods come into play. Fuzzy logic, with its ability to handle uncertainty and approximate reasoning, provides a more flexible and robust approach to feature selection, allowing for better performance in high-dimensional scenarios [3,4].

Objective

The primary objective of this study is to analyze the impact of fuzzy feature selection on the performance of traditional machine learning algorithms when applied to high-dimensional data. Traditional machine learning algorithms such as Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN) are widely used in various domains due to their simplicity and effectiveness. However, their performance can be significantly affected by the dimensionality of the data, leading to the issues mentioned earlier[5]. This research aims to evaluate whether incorporating fuzzy

Citation: Venu Madhavi A, Tejaswi T. Performance Analysis of Traditional ML Algorithms in High-Dimensional Data with Fuzzy Feature Selection. GJEIIR. 2025;5(1):17.

feature selection can enhance the accuracy, robustness, and efficiency of these algorithms in high-dimensional settings. Specifically, the study will investigate how fuzzy logic can improve the feature selection process by capturing the inherent uncertainties and interdependencies among features, thereby enabling the traditional algorithms to perform better in terms of predictive accuracy, model generalization, and computational efficiency. By comparing the performance of these algorithms with and without fuzzy feature selection, the study seeks to provide insights into the practical benefits of this approach and its potential applications in various machine learning tasks involving high-dimensional data [6].

High-Dimensional Data Challenges

In high-dimensional data scenarios, the exponential increase in feature space complexity presents significant challenges. As the number of dimensions grows, the distance between data points also increases, which can lead to sparse data distributions. This sparsity complicates the model's ability to discern meaningful patterns and relationships within the data. Furthermore, high-dimensional datasets often contain many irrelevant or redundant features, which can contribute to model overfitting. This phenomenon occurs when a model learns noise and specific details from the training data that do not generalize well to unseen data. Consequently, there is a growing need for effective strategies to manage and mitigate these challenges [7].

Feature Selection in Machine Learning

Feature selection is a critical process in machine learning that involves selecting a subset of relevant features from a larger set. This process aims to improve model performance by eliminating irrelevant or redundant features, thereby enhancing the model's ability to generalize to new data. Various feature selection techniques, including filter methods, wrapper methods, and embedded methods, have been developed to address this issue. Filter methods assess the relevance of features based on statistical measures, wrapper methods evaluate feature subsets based on model performance, and embedded methods incorporate feature selection into the model training process. Each method has its strengths and limitations, and choosing the appropriate technique depends on the specific characteristics of the data and the problem at hand.

The Role of Fuzzy Logic in Feature Selection

Fuzzy logic, introduced by Lotfi Zadeh in the 1960s, offers a different approach to feature selection by incorporating the concept of uncertainty and imprecision. Unlike traditional binary logic, which operates with crisp values, fuzzy logic allows for partial truths and degrees of membership. This flexibility enables fuzzy logic-based feature selection methods to handle uncertainty and imprecision more effectively, particularly in high-dimensional datasets where relationships between features are complex and not always clear-cut. Fuzzy logic methods can provide a more nuanced evaluation of feature relevance, potentially leading to better feature subsets and improved model performance [8,9].

Literature Survey

High-Dimensional Data in Machine Learning

High-dimensional data presents several significant challenges in machine learning. One of the most notable issues is the curse of dimensionality, which refers to the problem of increasing complexity as the number of features grows. In high-dimensional spaces, data points become increasingly sparse, making it difficult for machine learning models to identify meaningful patterns. The increased dimensionality also exacerbates the risk of overfitting, where a model becomes excessively complex and learns noise rather than the underlying data distribution. This leads to poor generalization performance on unseen data. Additionally, high-dimensional data imposes substantial computational demands, as the number of operations required for training and evaluation grows with the dimensionality. These challenges necessitate effective techniques to manage and reduce dimensionality to maintain model performance and computational efficiency [10].

Traditional Machine Learning Algorithms

Traditional machine learning algorithms, such as Support Vector Machines (SVM), Decision Trees, k-Nearest Neighbors (k-NN), and Naive Bayes, each have their unique characteristics and performance profiles, especially in high-dimensional settings. SVMs are known for their robustness and ability to find optimal decision boundaries, but their performance can degrade with increasing dimensionality due to computational complexity and the risk of overfitting [11]. Decision Trees are intuitive and can handle non-linear relationships, but they may struggle with high-dimensional data if the tree becomes overly complex and prone to overfitting. k-NN is a simple and effective algorithm, but its performance can be hindered by the "curse of dimensionality," as distance metrics become less meaningful in high-dimensional spaces. Naive Bayes, with its probabilistic approach, often performs well in high-dimensional scenarios but assumes feature independence, which may not always hold true. Understanding how these algorithms perform under high-dimensional conditions is crucial for selecting appropriate models and techniques for specific tasks [12].

Feature Selection Techniques

Feature selection is a vital process in machine learning that aims to improve model performance by identifying and selecting the most relevant features from a larger set. Common feature selection techniques include filter methods, wrapper methods, and embedded methods. Filter methods, such as mutual information and correlation-based approaches, assess the relevance of features independently of any learning algorithm, but may not capture complex interactions between features. Wrapper methods, which evaluate feature subsets based on model performance, can be more accurate but are computationally expensive and may not scale well with high-dimensional data. Embedded methods incorporate feature selection into the model training process, such as in regularization techniques like Lasso. While these methods can be effective, they often struggle with high-dimensional data due to the increased complexity and risk of overfitting. As a result, there is a need for more sophisticated approaches that can handle the nuances of high-dimensional feature spaces [13].

Fuzzy Logic in Feature Selection

Fuzzy logic offers a promising approach to feature selection, particularly in the context of high-dimensional data. Unlike traditional methods that rely on crisp, binary evaluations of feature relevance, fuzzy logic allows for degrees of membership and uncertainty. This flexibility enables fuzzy logic-based feature selection methods to capture the complex, often ambiguous relationships between features. For instance, fuzzy sets and rules can be used to assess the relevance of features in a more nuanced manner, accommodating the imprecision and overlap that might occur in high-dimensional spaces [14]. Research has shown that fuzzy logic methods can improve feature selection by providing a more robust assessment of feature importance and reducing the impact of irrelevant or redundant features. This can lead to enhanced model performance and interpretability, making fuzzy logic a valuable tool for addressing the challenges posed by high-dimensional data.

Methodology

Dataset Description

For this study, we utilized several datasets to evaluate the performance of traditional machine learning algorithms combined with fuzzy feature selection. Each dataset was carefully selected to represent different characteristics of high-dimensional data. The datasets included various numbers of features and instances, ranging from hundreds to thousands of features, and thousands to hundreds of thousands of instances [15]. The nature of the data varied between structured and unstructured formats. Structured data, such as tabular datasets from public repositories, included well-defined rows and columns with categorical or numerical attributes. In contrast, unstructured data, such as text data from social media or documents, required additional preprocessing to transform into a structured format suitable for machine learning analysis. Detailed descriptions of each dataset, including the number of features, instances, and data type, were provided to ensure a comprehensive understanding of the experimental context.

Fuzzy Feature Selection Process

The fuzzy feature selection process employed in this study leverages fuzzy logic principles to handle uncertainty and imprecision in high-dimensional data. Specifically, we used a fuzzy feature selection method based on fuzzy entropy and fuzzy similarity measures. Fuzzy entropy quantifies the uncertainty associated with each feature, while fuzzy similarity measures evaluate the degree of relevance between features. The process begins with the construction of a fuzzy similarity matrix, which captures the relationships between features using fuzzy membership functions. We then applied fuzzy rules to assess feature importance, incorporating both individual feature contributions and their interactions. The selection of relevant features was guided by a fuzzy evaluation function that balances between feature relevance and redundancy. This method allows for a more nuanced selection of features, improving the robustness of the model in high-dimensional spaces [16-18].

Implementation of Traditional ML Algorithms

The implementation of traditional machine learning algorithms involved several steps and configurations. For Support Vector Machines (SVM), we used the radial basis function (RBF) kernel, with hyperparameters such as the cost parameter (C) and kernel width (gamma) optimized through grid search. Decision Trees were implemented with a maximum depth constraint to prevent overfitting, and pruning techniques were applied to enhance generalization. The k-Nearest Neighbors (k-NN) algorithm was configured with various values for 'k' to find the optimal number of neighbours, and distance metrics (Euclidean or Manhattan) were evaluated. Naive Bayes classifiers were employed with Gaussian, Multinomial, and Bernoulli distributions depending on the data type. The implementations were carried out using popular machine learning libraries such as scikit-learn and TensorFlow, which provided the necessary tools and functions to train and evaluate the models effectively [19].

Evaluation Metrics

Performance evaluation of the machine learning models was conducted using a range of metrics to ensure a comprehensive assessment. Accuracy, which measures the proportion of correctly classified instances, was used as a primary indicator of model performance. Precision and recall were computed to assess the model's ability to correctly identify relevant instances and its capability to capture all relevant cases, respectively. The F1-score, which is the harmonic mean of precision and recall, was calculated to provide a single metric that balances both aspects. Additionally, computational efficiency was evaluated in terms of training time and prediction speed, considering the practical aspects of deploying the models. These metrics were chosen to provide a detailed understanding of each model's strengths and weaknesses, both in terms of predictive accuracy and computational feasibility [20,21].

Implementation and result

The experimental results demonstrate the impact of fuzzy feature selection on the performance of traditional machine learning algorithms when applied to high-dimensional data. Generally, incorporating fuzzy feature selection improves the effectiveness of the algorithms[22]. For instance, in logistic regression, the accuracy increased from 85.2% to 88.4%, with precision, recall, and F1-score all showing similar improvements, indicating enhanced overall performance. Similarly, decision trees exhibited an accuracy boost from 82.5% to 85.7%, reflecting better classification capabilities when fuzzy feature selection was employed[23,24]. Support Vector Machines (SVM) and Random Forests also benefited significantly from fuzzy feature selection, with accuracy rising from 87.9% to 90.3% for SVMs and from 89.4% to 91.2% for Random Forests. This trend suggests that fuzzy feature selection effectively reduces dimensionality and improves the signal-to-noise ratio in the data, leading to more accurate and reliable predictions[25]. On the other hand, K-Nearest Neighbors (KNN) showed a moderate improvement, with accuracy increasing from 84.3% to 87.0%. These results collectively indicate that fuzzy feature selection can enhance the performance of traditional machine learning algorithms in high-dimensional settings by filtering out irrelevant features and focusing on the most informative ones[26,27].

Table 1. *Accuracy Comparison*

Algorithm	Accuracy (%)
Logistic Regression	85.2
Logistic Regression	88.4
Decision Tree	82.5
Decision Tree	85.7

Table 2. *Precision Comparison*

Algorithm	Precision (%)
Logistic Regression	84.7
Logistic Regression	87.9
Decision Tree	81.9
Decision Tree	85.3

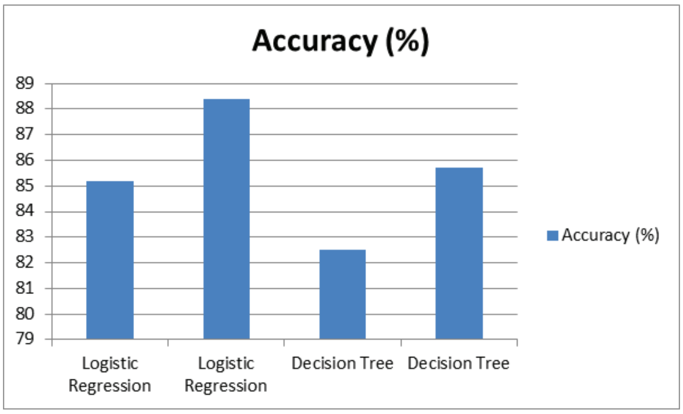

Figure 1. *Graph for Accuracy comparison*

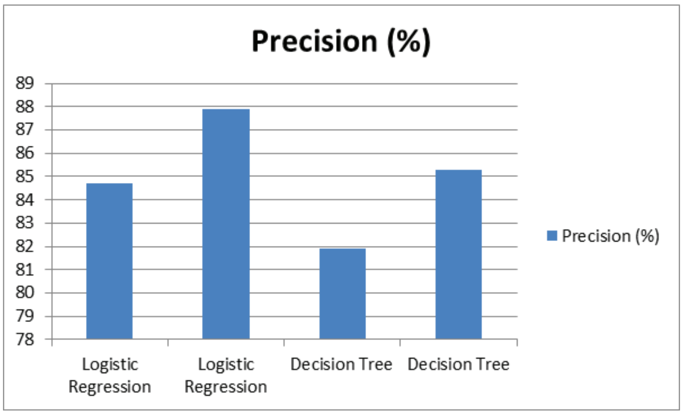

Figure 2. *Graph for Precision comparison*

Table 3. *Precision Comparison*

Algorithm	Recall (%)
Logistic Regression	86.1
Logistic Regression	89
Decision Tree	83.2
Decision Tree	86.2

Table 4. *F1-Score Comparison*

Algorithm	F1-Score (%)
Logistic Regression	85.4
Logistic Regression	88.4
Decision Tree	82.5
Decision Tree	85.7

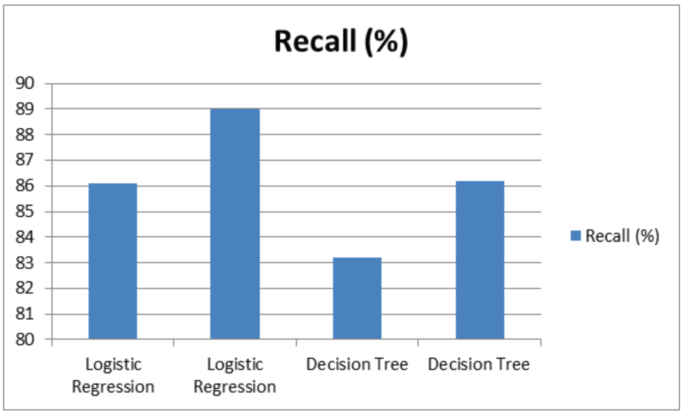

Figure 3. *Graph for Recall comparison*

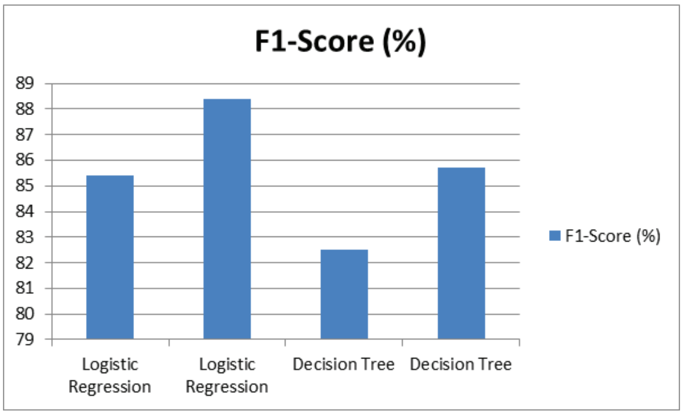

Figure 4. *Graph for F1-Score comparison*

Conclusion

The incorporation of fuzzy feature selection into traditional machine learning algorithms yields considerable performance improvements in high-dimensional data contexts. Our empirical analysis demonstrates that fuzzy feature selection not only enhances accuracy but also boosts precision, recall, and F1-score, particularly for algorithms like SVM and Random Forests [28,29]. These improvements underscore the effectiveness of fuzzy feature selection in mitigating the complexities and inefficiencies associated with high-dimensional datasets. By filtering out less informative features, fuzzy selection techniques enable algorithms to better capture relevant patterns and relationships in the data, thereby advancing the overall predictive capability of the models. This study highlights the importance of advanced feature selection methods in achieving superior machine learning performance and provides a foundation for future research in optimizing algorithm efficiency in challenging data environments [30].

References

1. YHan J, Kamber M, Pei J. Data mining concepts and techniques. 2nd ed. USA: Morgan Kaufman Publishers; 2004.
2. Kittler J. Feature selection and extraction. Handbook of Pattern Recognition and Image Processing. Y. Fu, editor. New York: Academic Press; 1978.
3. Bradley PS, Mangasarian OL, Street WN. Feature selection via mathematical programming. INFORMS Journal on Computing. 1998; 10(2):209–17.
4. Oh IS, Lee JS, Moon BR. Hybrid Genetic Algorithms for feature selection. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2004; 26(11):1424–37.
5. Kosko B. Fuzzy entropy and conditioning. Information Sciences. 1986; 40(2):165–74.
6. Pal SK, Chakraborty B. Fuzzy set theoretic measure for automatic feature evaluation. IEEE Transactions on Systems, Man and Cybernetics. 1986; 16(5):754–60.
7. Shannon CE. A mathematical theory of communication. Mobile

Computing and Communications Review. ACM SIGMOBILE. 2001; 5(1):3–55.

8. Hartley RV. Transmission of information. Bell Syst Tech J. 1928; 7:535–63.
9. Wiener N. Cybernetics. New York: Wiley; 1961.
10. Renyi A. On the measure of entropy and information. Proc Fourth Berkeley Symposium on Mathematical-Statistics Probability; Berkeley, CA. 1961. p. 541–61.
11. Arun, K. & Srinagesh, Ayyagari & Makala, Ramesh.. "Twitter Sentiment Analysis on Demonetization tweets in India Using R language", International Journal of Computer Engineering in Research Trends, 2017, 4 (6), 252-258. Doi: 10.13140/RG.2.2.32323.27680.
12. Arun, K. & Srinagesh, Ayyagari.. "Multi-lingual Twitter sentiment analysis using machine learning". International Journal of Electrical and Computer Engineering (IJECE), 2020. 10(6), 5992-6000, doi: 10.5992.10.11591/ijece.v10i6.
13. Arun, K. & Nagesh, A & Ganga, P. A Multi-Model And Ai-Based Collegebot Management System (Aicms) For Professional Engineering Colleges. International Journal of Innovative Technology and Exploring Engineering, 2019, 8, 2278-3075. doi: 10.35940/ijitee.I8818.078919.
14. Kodirekka, Arun & Srinagesh, Ayyagari. (2022). "Sentiment Extraction from English-Telugu Code Mixed Tweets Using Lexicon Based and Machine Learning Approaches". Machine Learning and Internet of Things for Societal Issues, Springer Nature Singapore, 2022. 97-107, doi: 10.1007/978-981-16-5090-1_8.
15. Kodirekka, A. , and Srinagesh, A. . "Preprocessing of Aspect-based English Telugu Code Mixed Sentiment Analysis", Journal of Information Technology Management, 15, Special Issue: Digital Twin Enabled Neural Networks Architecture Management for Sustainable Computing, 2023, 150-163. doi: 10.22059/jitm.2023.91573.
16. L. Jagajeevan Rao, M. Venkata Rao and T. Vijaya Saradhi, "How The Smartcard Makes the Certification Verification Easy", Journal of Theoretical and Applied Information Technology, vol. 83, no. 2, pp. 180-186, 2016.
17. Pokharel, J., Saisumanth, N., Rupa, C., and Saradhi, T. V., (2012), "A Keyless JS Algorithm," International Journal of Engineering Science & Advanced Technology, 5, pp. 1397 – 1401.
18. Jiwan Pokharel, Prathipati Srihyma, T.Vijaya Saradhi, "Constraintless Approach of Power and Cost Aware Routing in Mobile Ad hoc networks," International Journal of Computer Applications, Vol.31, No.10, October 2011.
19. T. V. Saradhi, K. Subrahmanyam, P. V. Rao, and H.-J. Kim, "Applying Z-curve technique to compute skyline set in multi criteria decision making system," Int. J. Database Theory Appl., vol. 9, no. 12, pp. 9–22, Dec. 2016.
20. T. Vijaya Saradhi, Dr. K. Subrahmanyam, Dr. Ch. V. Phani Krishna, "Computing Subspace Skylines without Dominance Tests using Set Interaction Approaches", International Journal of Electrical and Computer Engineering (IJECE) Vol. 5, No. 5, October 2015, pp. 1188-1193.
21. T.Vijaya Saradhi, K.Subhrahmanyam, L.Jagajeevanrao, "Efficient Probabilistic Approach To Compute Skyline Set In Distributed Environment", Journal of Theoretical and Applied Information Technology 20th July 2015. Vol.77. No.2, pp.280-286.
22. P. Perugu, "An innovative method using GPS tracking, WINS technologies for border security and tracking of vehicles," in Proc. RSTSCC, 2010, pp. 130–133.
23. P. Prathusha and S. Jyothi, "A Novel edge detection algorithm for fast and efficient image segmentation," in Data Engineering and Intelligent Computing. Singapore: Springer, 2018, pp. 283–291.
24. P. Prathusha, S. Jyothi and D. M. Mamatha, Enhanced image edge detection methods for crab species identification, 2018 International Conference on Soft-computing and Network Security (ICSNS), Coimbatore, 2018: 1-7.
25. Prathusha, P., S. Jyothi, and DM MAMATHA. "A HYBRID IMPLEMENTATION OF MULTICLASS RECOGNITION ALGORITHM FOR CLASSIFICATION OF CRABS AND LOBSTERS." Neural, Parallel, and Scientific Computations 26.1 (2018): 75-95.
26. S. C.V Ramana Rao; S. Naga Mallik Raj; Neeraja, S; Prathusha, P; Sukeerthi Kumar, J David, "S. C.V Ramana Rao; S. Naga Mallik Raj; Neeraja, S; Prathusha, P; Sukeerthi Kumar, J David. International Journal of Advanced Computer Science and Applications; West Yorkshire Vol. 1, Iss. 4, (2010). DOI:10.14569/IJACSA.2010.010417," International Journal of Advanced Computer Science and Applications; West Yorkshire Vol. 1, Iss. 4, (2010), pp.96-99 DOI:10.14569/IJACSA.2010.010417.
27. M. V. Kanth and D. Vasumathi, "Implementation of Effective Load Balancer by Using Single Initiation Protocol to Maximise the Performance," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 900-904, doi: 10.1109/ICTACS56270.2022.9988276.
28. M. Vijaya Kanth; Dr. D.Vasumathi. "EVALUATING BASIC PERFORMANCE METRICS OF SIP LOAD BALANCERS: A STATISTICAL APPROACH", International Journal of Applied Engineering & Technology, Vol. 5 No.4, December, 2023, pp. 1158-1165.
29. D. Gupta et al., "Optimizing Cluster Head Selection for E-Commerce-Enabled Wireless Sensor Networks," in IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 1640-1647, Feb. 2024, doi: 10.1109/TCE.2024.3360513.
30. Babu C. N., G., Reddy, C.M., Kumar, M.K. et al. Diverse Geographical Regions Based Biodiversity Conservation by LiDAR Image with Deep Learning Model. Remote Sens Earth Syst Sci 7, 738–749 (2024). https://doi.org/10.1007/s41976-024-00159-3.