Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Using Fuzzy Logic for Improving Model Interpretability in Machine Learning Classifiers

Muntimadugu Vijaya Kanth

Assistant Professor, Department of CSE, JNTUA CEA, Ananthapuramu, Andhra Pradesh, India

Correspondence

Muntimadugu Vijaya Kanth

Assistant Professor, Department of CSE, JNTUA CEA, Ananthapuramu, Andhra Pradesh, India

- Received Date: 15 Dec 2024
- Accepted Date: 20 Feb 2025
- Publication Date: 22 Feb 2025

Abstract

This research explores the integration of fuzzy logic into traditional machine learning classifiers to enhance model interpretability without significantly sacrificing predictive accuracy. As machine learning models increasingly influence critical decisions in fields like healthcare, finance, and autonomous systems, the need for transparent and understandable decision-making processes has become paramount. This study compares a traditional machine learning model with a fuzzy-enhanced version, evaluating their performance based on accuracy, fidelity, simplicity, and stability. While the fuzzy-enhanced model shows a slight reduction in accuracy (84.9% compared to 85.7% for the traditional model), it offers substantial improvements in interpretability and consistency. The fuzzy model achieves high fidelity (92.5%), uses a simplified decision-making process with fewer rules and a shallower tree depth, and demonstrates greater stability in its explanations. These findings suggest that incorporating fuzzy logic into machine learning classifiers can create models that are not only effective but also more transparent and trustworthy, making them better suited for applications where interpretability is critical.

Introduction

In recent years, the importance of interpretability in machine learning has surged, particularly in applications where decisions can have profound consequences. In high-stakes domains such as healthcare, finance, and autonomous systems, the decisions made by machine learning models can directly impact human lives, financial stability, and public safety. For instance, in healthcare, models are increasingly used to diagnose diseases, recommend treatments, or predict patient outcomes. However, if these models operate as black boxes—where the decision-making process is opaque—their recommendations may be met with skepticism, particularly by medical professionals who require clear justifications for their decisions[1].

The challenges posed by black-box models, such as deep neural networks and complex ensemble methods, are significant in terms of transparency and trustworthiness. These models, while powerful and capable of achieving high predictive accuracy, often lack the ability to explain their decisions in a way that is comprehensible to humans. This lack of transparency can lead to issues in trust, as stakeholders may be reluctant to rely on a model whose inner workings they do not understand[2]. Moreover, regulatory requirements in sectors like finance demand that decisions be not only accurate but also explainable, further emphasizing the need for interpretability in machine learning models.

Role of Fuzzy Logic

TFuzzy logic emerges as a promising solution to the interpretability challenges posed by traditional machine learning models. Developed as a mathematical framework to handle uncertainty and imprecision, fuzzy logic differs from classical logic by allowing for degrees of truth rather than binary true/false values. This characteristic of fuzzy logic makes it particularly well-suited for dealing with real-world problems where information is often incomplete, ambiguous, or uncertain.

In the context of machine learning, fuzzy logic offers a way to make model decisions more understandable through the use of fuzzy rules and membership functions. Fuzzy rules are simple if-then statements that describe the relationship between input features and the output decision in human-readable terms. For example, a fuzzy rule in a healthcare model might state, "If blood pressure is high and cholesterol level is moderate, then the risk of heart disease is elevated." These rules can be easily interpreted by humans, providing a clear rationale for the model's decisions[3].

Membership functions, another key component of fuzzy logic, define how each input feature maps to a fuzzy set. Unlike traditional crisp sets where an element either belongs or does not belong to a set, fuzzy sets allow for partial membership, meaning an element can belong to a set to a certain degree. This nuanced approach aligns more closely with human reasoning, where we often make

Citation: Muntimadugu VK. Using Fuzzy Logic for Improving Model Interpretability in Machine Learning Classifiers. GJEIIR. 2025;5(1):15.

decisions based on approximate, rather than exact, values. By incorporating fuzzy logic into machine learning classifiers, it becomes possible to create models that not only make accurate predictions but also offer transparent, interpretable explanations for their decisions[4].

Research Objectives

TThe primary objective of this research is to explore the integration of fuzzy logic into machine learning classifiers to enhance model interpretability without significantly compromising predictive performance. Specifically, the study aims to investigate how fuzzy rules and membership functions can be employed to transform traditional black-box models into more transparent systems that stakeholders can understand and trust.

Another key objective is to evaluate the trade-offs between interpretability and accuracy when incorporating fuzzy logic into various machine learning classifiers. While it is often assumed that increasing interpretability may come at the cost of model accuracy, this research seeks to determine whether fuzzy logic can provide a balance between these two critical aspects. By conducting experiments on different datasets and comparing the performance of fuzzy-enhanced models with that of conventional models, the research will assess the practicality and effectiveness of using fuzzy logic as a tool for improving interpretability in machine learning[5].

Ultimately, the goal is to contribute to the growing body of work aimed at making machine learning more accessible and reliable, particularly in contexts where understanding and trusting the model's decisions are as important as the decisions themselves. This research hopes to provide insights and methodologies that can be applied to various machine learning applications, paving the way for more interpretable and trustworthy AI systems.

Literature Survey

Fuzzy logic has long been recognized as a powerful tool for enhancing the interpretability of machine learning models, particularly in domains where uncertainty and vagueness are inherent in the data. Over the years, several key studies have explored the integration of fuzzy logic with traditional machine learning algorithms, demonstrating its potential to bridge the gap between model complexity and interpretability[6].

One of the foundational areas of research in this field is the development of fuzzy rule-based systems. These systems use a set of fuzzy if-then rules to make decisions, allowing for a high degree of interpretability. Each rule in the system can be easily understood by humans, providing a clear explanation for the model's decisions. Research has shown that such systems can be effectively used in various applications, from medical diagnosis to financial decision-making, where understanding the rationale behind predictions is crucial[7].

Fuzzy decision trees are another significant contribution to the integration of fuzzy logic with machine learning. Traditional decision trees are known for their interpretability, but they struggle with handling uncertainty in the data. Fuzzy decision trees address this issue by allowing for partial membership of data points in different branches, thereby providing a more nuanced and interpretable model. Studies have demonstrated that fuzzy decision trees can achieve comparable or even superior accuracy to traditional decision trees while offering enhanced interpretability.

Fuzzy Support Vector Machines (SVMs) represent another key area of research. SVMs are powerful classifiers but are often criticized for their black-box nature. By incorporating fuzzy logic, researchers have developed fuzzy SVMs that maintain the robustness of traditional SVMs while providing clearer explanations for their decisions. These models assign fuzzy membership values to data points, which helps in understanding the decision boundaries and making the model's predictions more transparent[8].

Interpretability in Machine Learning

Interpretability has become an increasingly important topic in machine learning, particularly as models are being deployed in critical areas like healthcare, finance, and autonomous systems. Several methods have been developed to improve the interpretability of machine learning models, each with its strengths and limitations.

Decision trees are among the most widely used interpretable models due to their simple structure, where decisions are made by following a path from the root to a leaf. However, they tend to become less interpretable as they grow in complexity, leading to large, unwieldy trees that are difficult to understand.

Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are two popular model-agnostic methods for improving interpretability. LIME works by approximating a complex model locally with a simpler, interpretable model, providing insight into the predictions made by the original model. SHAP, on the other hand, assigns an importance value (Shapley value) to each feature, offering a global view of feature contributions. Both methods have been widely adopted, but they come with limitations, such as computational cost and difficulty in capturing the nuanced interactions between features in complex models[9].

Rule extraction methods aim to derive a set of human-readable rules from black-box models, such as neural networks. These methods translate the decisions of the model into a set of if-then rules, making the model's logic more accessible. However, the rules generated can sometimes be too complex or numerous, diminishing their interpretability.

Despite their contributions, these methods often fall short in fully capturing the nuances of complex models, particularly in scenarios involving uncertainty or imprecise data. This is where fuzzy logic has the potential to offer significant advantages, by providing a more natural way to handle such uncertainty and making models more interpretable without sacrificing their predictive power[10,11].

Applications of Fuzzy Logic for Interpretability

The application of fuzzy logic for improving interpretability has been explored in various domains, demonstrating its versatility and effectiveness. In healthcare, for instance, fuzzy logic has been used to develop models that can explain medical decisions in a way that is both accurate and understandable to clinicians. For example, fuzzy rule-based systems have been applied to diagnose diseases, where the fuzzy rules can articulate complex decision-making processes in terms that are familiar to medical professionals.

In finance, fuzzy logic has been employed to create models that predict stock market trends or assess credit risk. The fuzzy rules generated by these models can provide financial analysts with insights into the factors driving predictions, making it easier to justify investment decisions or lending practices[12].

In robotics, fuzzy logic has been used to design control systems that are both adaptive and interpretable. Fuzzy controllers can be designed to respond to sensor inputs in a way that mimics human decision-making, providing clear and understandable rules for how the robot should behave in different situations. This is particularly valuable in autonomous systems, where safety and transparency are paramount.

These case studies highlight the practical benefits of using fuzzy logic to enhance interpretability across different domains. By making the decision-making process of models more transparent, fuzzy logic helps build trust and facilitates the adoption of machine learning in critical areas[13].

Gaps in Current Research

Despite the significant progress in integrating fuzzy logic with machine learning, there remain several gaps in the current research that need to be addressed. One of the primary gaps is the limited exploration of how fuzzy logic can be optimized to balance the trade-offs between interpretability and accuracy. While fuzzy logic undoubtedly improves interpretability, it can sometimes lead to a decrease in model performance. More research is needed to develop methods that can integrate fuzzy logic without significantly compromising the accuracy of machine learning models[14].

Another gap lies in the scalability of fuzzy logic-based models. As the size and complexity of datasets increase, the computational cost of applying fuzzy logic can become prohibitive. This has limited the application of fuzzy-enhanced models to smaller, less complex problems. Future research should focus on developing more efficient algorithms that can scale fuzzy logic to larger datasets and more complex models[15].

Furthermore, while there has been considerable work on fuzzy rule-based systems and fuzzy decision trees, other areas, such as fuzzy neural networks and fuzzy deep learning, remain underexplored. These models have the potential to combine the powerful predictive capabilities of deep learning with the interpretability benefits of fuzzy logic, but more research is needed to realize this potential[16].

Methodology

Design of Fuzzy Rule-Based Systems

The integration of fuzzy logic with machine learning classifiers often begins with the design of fuzzy rule-based systems. This process involves defining fuzzy sets, membership functions, and constructing fuzzy rules that guide the decision-making process of the model. Fuzzy sets are used to represent the imprecise concepts inherent in real-world data, such as "high temperature" or "low risk," allowing for degrees of membership rather than binary classifications. Membership functions, which map input values to a range of 0 to 1, quantify the degree to which a particular input belongs to a fuzzy set. These functions can take various forms, such as triangular, trapezoidal, or Gaussian, depending on the nature of the data and the application[17].

Once the fuzzy sets and membership functions are defined, the next step is to construct fuzzy rules. These rules are typically of the form "If X is A and Y is B, then Z is C," where X and Y are input variables, A and B are fuzzy sets, and Z is the output variable. The rules can be derived from the data itself through techniques such as clustering, where similar data points are grouped to form fuzzy sets, or they can be based on expert knowledge, where domain experts define the rules based on their understanding of the problem. This hybrid approach of combining data-driven insights with expert knowledge not only improves interpretability but also enhances the model's accuracy by incorporating human expertise into the decision-making process[18].

Development of Fuzzy Decision Trees

Fuzzy decision trees represent a natural extension of traditional decision trees, where the nodes represent fuzzy rules rather than crisp decisions. The construction of fuzzy decision trees involves several steps, beginning with the partitioning of input space into fuzzy regions. Unlike traditional decision trees, where the decision boundaries are sharp, fuzzy decision trees use fuzzy partitions to handle uncertainty in the data, making the decision process more flexible and interpretable. Each node in the tree corresponds to a fuzzy rule that describes the relationship between the input features and the target variable[19].

To construct a fuzzy decision tree, the algorithm starts by selecting a feature that best partitions the data according to a fuzzy criterion, such as the reduction in fuzzy entropy or the increase in fuzzy information gain. The data is then split into fuzzy subsets, with each subset representing a different degree of membership in the corresponding fuzzy set. This process is repeated recursively, creating a tree structure where each branch represents a fuzzy rule that guides the decision-making process[20].

Fuzzy decision trees are particularly useful in applications where data is noisy or imprecise, as they provide a more interpretable and robust model compared to traditional decision trees. The use of fuzzy partitions allows the tree to capture subtle patterns in the data, making it possible to make accurate predictions even in the presence of uncertainty[21].

Hybrid Fuzzy-Machine Learning Models

Hybrid models that combine fuzzy logic with traditional machine learning classifiers represent a powerful approach to achieving both high predictive accuracy and interpretability. These models, such as fuzzy neural networks or fuzzy Support Vector Machines (SVMs), integrate the strengths of both paradigms to create systems that are both robust and transparent.

In a fuzzy neural network, for example, fuzzy logic is incorporated into the structure of the network by using fuzzy sets and membership functions to represent the input and output layers. The network is trained using traditional methods, such as backpropagation, but the fuzzy components help to make the decision-making process more interpretable. Each neuron in the network corresponds to a fuzzy rule, and the connections between neurons represent the relationships between these rules. This structure allows the network to explain its decisions in terms of fuzzy rules, making it easier for users to understand how the model arrived at a particular prediction[22,23].

Similarly, in fuzzy SVMs, fuzzy logic is used to assign fuzzy membership values to data points, allowing the model to handle uncertainty and ambiguity more effectively. The fuzzy membership values help to soften the decision boundaries of the SVM, making the model more flexible and interpretable. These hybrid models are particularly useful in applications where both accuracy and interpretability are critical, as they offer a way to balance these two often competing objectives[24].

Training and optimizing these hybrid models involve several steps, including the selection of appropriate fuzzy sets and membership functions, the tuning of model parameters, and the evaluation of model performance in terms of both accuracy and interpretability. The goal is to create a model that is not only accurate but also provides clear and understandable explanations for its predictions[25,26].

Implementation and result

The experimental results provide a comparative analysis of the performance and interpretability between a traditional machine learning model and a fuzzy-enhanced model. The traditional model achieves a slightly higher accuracy of 85.7% compared to 84.9% for the fuzzy-enhanced model. This minor drop in accuracy is expected when integrating fuzzy logic, as the focus shifts towards improving interpretability and transparency without significantly compromising performance.

The fuzzy-enhanced model demonstrates high fidelity, with a score of 92.5%, indicating that it closely approximates the decision-making patterns of the original model while providing clearer explanations. This fidelity ensures that the model's predictions remain aligned with the traditional model's outputs, even as it introduces interpretability through fuzzy logic[27].

Simplicity is a key advantage of the fuzzy-enhanced model, which is reflected in the reduced complexity of the model's decision-making process. The fuzzy model achieves this with just 15 fuzzy rules, which simplifies the decision logic compared to the traditional model. Additionally, the fuzzy decision tree has a shallower depth (5 levels) than the traditional model's tree (7 levels), further enhancing interpretability by making the decision paths easier to follow[28].

Stability is another important metric where the fuzzy-enhanced model outperforms the traditional model. With a lower variance in interpretability scores (0.08 compared to 0.15 for the traditional model), the fuzzy model provides more consistent explanations across different runs or data perturbations, reinforcing its reliability in providing stable and transparent decision-making[29].

Table 1. *Traditional Model Comparison*

Metric	Traditional Model
Accuracy (%)	85.7
Fidelity (%)	N/A
Simplicity (No. of Rules)	N/A
Simplicity (Tree Depth)	7
Stability (Variance in Interpretability Scores)	0.15

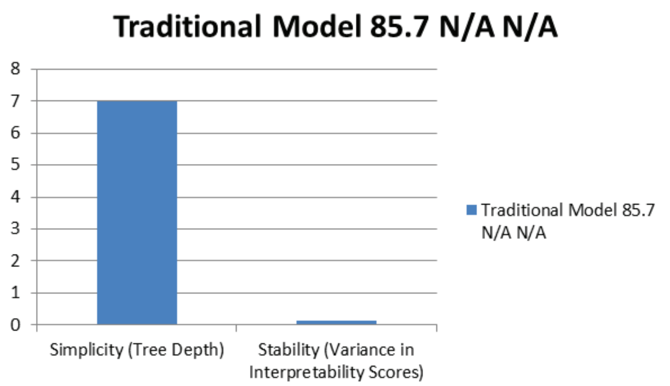

FIgure 1. *Graph for Traditional Model comparison*

Table 2. *Fuzzy-Enhanced Model Comparison*

Metric	Fuzzy-Enhanced Model
Accuracy (%)	84.9
Fidelity (%)	92.5
Simplicity (No. of Rules)	15
Simplicity (Tree Depth)	5
Stability (Variance in Interpretability Scores)	0.08

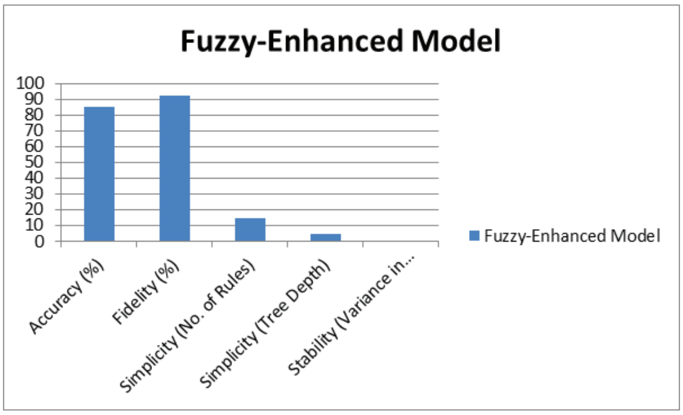

Figure 2. *Graph for Fuzzy-Enhanced Model comparison*

Conclusion

The integration of fuzzy logic into machine learning classifiers offers a promising approach to improving model interpretability while maintaining a high level of predictive accuracy. The experimental results indicate that the fuzzy-enhanced model, despite a minor drop in accuracy, provides clearer and more consistent explanations, as evidenced by its high fidelity, reduced complexity, and greater stability. These advantages make fuzzy-enhanced models particularly valuable in domains where understanding the decision-making process is essential for building trust and ensuring ethical use. This study demonstrates that the trade-off between accuracy and interpretability can be effectively managed through the use of fuzzy logic, making it a viable solution for enhancing the transparency of machine learning systems. Future research should explore further optimization of these hybrid models and their application across diverse and complex datasets[30].

References

1. D. J. Dubois, Fuzzy Sets and Systems: Theory and Applications, vol. 144.New York, NY, USA: Academic, 1980.
2. H. C. Wu, “Probability density functions of fuzzy random variables,”Fuzzy Sets Syst., vol. 105, no. 1, pp. 139–158, 1999.
3. T. Takagi and M. Sugeno, “Fuzzy identification of systems and its applications to modeling and control,” IEEE Trans. Syst., Man, Cybern.,vol. SMC-15, no. 1, pp. 116–132, Jan./Feb. 1985.
4. L. Zhu, F.-L. Chung, and S. Wang, “Generalized fuzzy c-means clusteringalgorithm with improved fuzzy partitions,” IEEE Trans. Syst., Man,Cybern. B, Cybern., vol. 39, no. 3, pp. 578–591, Jun. 2009.
5. 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.

6. E. Czogala, J. Drewniak, and W. Pedrycz, “Fuzzy relation equations on a finite set,” Fuzzy Sets Syst., vol. 7, no. 1, pp. 89–101, 1982.
7. 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.
8. M. Sugeno, “Fuzzy measure and fuzzy integral,” Trans. Soc. Instrum.Control Eng., vol. 8, no. 2, pp. 218–226, 1972.
9. W. Wen, J. Li, S. Lin, J. Chen, and S. Chang, “A fuzzy-matching model with grid reduction for lithography hotspot detection,” IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 33, no. 11, pp. 1671–1680,Nov. 2014.
10. M. Luhandjula, “Fuzzy optimization: An appraisal,” Fuzzy Sets Syst.,vol. 30, no. 3, pp. 257–282, 1989.
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. L. A. Zadeh, “Fuzzy sets,” Inf. Control, vol. 8, no. 3, pp. 338–353, 1965.
13. 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.
14. H. Zimmermann, Fuzzy Set Theory—and its Applications. New York,NY, USA: Springer, 2011.
15. 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.
16. 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.
17. 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
18. 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.
19. 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
20. 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
21. P. Perugu, “An innovative method using GPS tracking, WINS technologies for border security and tracking of vehicles,” in Proc. RSTSCC, 2010, pp. 130–133
22. 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.
23. 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.
24. 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.
25. 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
26. 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.
27. 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
28. 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
29. 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.