SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Intelligent Data Analytics Using Explainable and Interpretable Machine Learning Techniques

Dr. G Jagan Naik[1], M SilpaRaj[2], B. Rajarao[3], Dr. Allam Balaram[4], Kummari Jyothsna[5], V. Ramesh[6]

[1]Associate Professor, Computer Science and Engineering, CMR Institute of Technology, Kandlakoya, Medchal, Hyderabad-501401. E-mail: gjnaik1106@gmail.com

[2]Senior Assistant Professor, Department of CSE (Cyber Security) CVR College of Engineering, Hyderabad-501510. E-mail: silparajm@gmail.com

[3]Department of Computer Science and Engineering (Data Science), CMR Engineering College & Technology, Medical, Hyderabad-501401, Telangana, India. E-mail: b.rajarao1207@gmail.com

[4]Professor and HOD, Department of CSE-Data Science, Sphoorthy Engineering College, Hyderabad-501510. E-mail: drbalaramallam@gmail.com

[5]Assistant professor, Computer science and Engineering (Data science), CMR Institute of Technology, Kandlakoya, Medchal. E-mail: jyothsnak@cmritonline.ac.in

[6]CMR Institute of Technology, Hyderabad, Department of CSE AI and ML. E-mail: rameshvoruganti36@gmail.com

Abstract

With the widespread adoption of advanced machine-learning algorithms in data-driven decision systems, there is an increasing need for methods that accurately predict outcomes as well as provide understandable and transparent explanations. In the present research, an intelligent data-analytic framework was designed, in which the algorithm of machine learning was compared with each other and the machine learning explainability was assessed. To obtain repeated production of the synthetic binary classification data provided, a fixed random seed was used to generate a synthetic set of 500 observations, 6 continuous predictors and 1 outcome variable. The Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting and Neural Network models were tested using the accuracy, precision, recall, specificity and F1 score. Fidelity, stability, consistency, computational efficiency, and human interpretability were used for the assessment of Explanability, which involved the use of SHapley Additive exPlanations, Local Interpretable Model-Agnostic Explanations, Integrated Gradients, permutation importance, and partial-dependence analysis. The neural network had the best accuracy (86.2%), recall (86.7%) and F1 score (86.5%), while XGBoost had the best precision (86.8%) and specificity (86.5%). SHAP was the most explainable (4.36/5) and had good fidelity, consistency and interpretability. The feature-importance analysis indicated that the features 1, 2, and 3 are the most important features with values of 28.6%, 22.4%, and 18.8%, respectively. The analysis reveals that the quality of the explanations is one of the key indicators to consider in addition to the traditional indicators. These results, however, are synthetic (not empirical) and merely illustrative of the output of the analyses and not actual results. The framework should be further tested by using real and domain-specific data, multiple times of cross-validation, external populations, fairness evaluation and end-user evaluation. This validation may help to build generalizability, robustness and practicality of the use of the process in high-stakes decision-making contexts.

Keywords: explainable artificial intelligence; interpretable machine learning; intelligent data analytics; SHAP; predictive modelling; feature importance

Introduction

Over the last few years, healthcare, financial, manufacturing, cybersecurity, engineering, and many other data-driven industries have experienced rapid growth and innovation of digital information, leading to the adoption of machine-learning (ML) and artificial-intelligence (AI) systems for pattern recognition, predictive modelling, risk assessment, and intelligent decision-making. Advanced ensemble algorithms and deep neural networks can give good predictive performance but these algorithms can be complex and the end user may not be able to see how each variable contributes to a specific prediction. Such “black-box” behavior can limit model validation, diminish stakeholder trust, and pose real-world problems for automated predictions when making important decisions. The present manuscript, therefore, places greater emphasis on the assessment of the quality of the explanations produced than on the traditional assessment of predictive performance alone, as being the basis for identifying intelligent data analytics. Ribeiro et al. (2016) proposed Local Interpretable Model-Agnostic Explanations (LIME): a method that approximates the behaviour of a complex predictor locally using an interpretable surrogate model; and SHapley Additive exPlanations (SHAP) developed by Lundberg and Lee (2017) is a unified framework for feature-attributions based on cooperative game theory, which offers locally accurate and consistent estimates of feature contributions. The landscape of interpretability has also been expanded by other methods: layer-wise relevance propagation decomposes nonlinear predictions into relevance scores for each individual input; Testing with Concept Activation Vectors

evaluates predictions by human-understandable concepts; Integrated Gradients attributes neural-network predictions based on principles of implementation-invariance and sensitivity; and counterfactual approaches involve small modifications of the inputs that can influence a model decision.As AI systems are being adopted in the fields of cyber-physical systems and healthcare, these techniques become more and more important. For instance, Naik et al (2026) noted the integration of Artificial Intelligence, Machine Learning, sensing technologies, robotics, IoT, digital twins, wearable systems, and health-care-management platforms into the creation of intelligent medical devices, highlighting the ongoing challenges of explainability, data quality, interoperability, cyber security, model drift, human factors, regulatory requirements, and clinical validation. Likewise, Preethi et al. (2026) proposed an explainable Artificial Intelligence (AI) based Healthcare Decision Support System (HDDSS) for clinical-risk prediction, combined with resource-demand and operational-management components to achieve explainability using SHAP alongside excellent predictive performance, supporting workflow efficiency, waiting time, length of stay, resource allocation and professional acceptance, albeit requiring further prospective real-world validation. These are just some examples of how interpretability is not just about making visualization possible, it's an essential element of responsible deployment, especially when AI-driven recommendations impact patients, resources, or clinical outcomes. However, there are significant differences in the fidelity, stability, consistency, computational cost of explanation techniques, the extent of explanation provided, and their user-friendliness, and apparently convincing explanations do not necessarily represent the underlying model. Thus, in order to evaluate the explanatory power and the predictive capability of the model, a systematic evaluation of both is required. The present study thus proposes an integrated framework for intelligent data-analytics that transforms this data into structured data, compares a number of machine-learning algorithms, generates local and global explanations and assesses various interpretability techniques based on their fidelity, stability, consistency, computational efficiency and human interpretability. The goal of the framework is to help develop more transparent, reliable and practically applicable machine-learning systems for evidence-based decision support by considering both predictive capability and explanatory reliability.

Literature Review

Intelligent data analytics involves a process of transforming large data sets with different types of data into predictions or decision-supporting evidence using statistical learning, machine learning and computational optimisation. Gradient boosting is one of the known analysis methods that is used for successive approximation of functions to construct an additive predictive model and has been proven to be useful in classification and regression analysis problems with nonlinear relationships (Friedman 2001). However, with increasingly complex models, transparency is likely to diminish, and models that are inherently interpretable can be created, as well as explanations after the fact. However, interpretability does not always come at a cost, as Caruana et al. showed that having a clinically meaningful pattern of risk could be obtained with high predictive accuracy through the use of generalised additive models with pairwise interactions. In the case of complex neural networks, Shrikumar et al. (2017) came up with DeepLIFT, which uses neuron activations to determine the difference between activations and a reference one, and then passes the score backwards to find out the important input features efficiently. In a theoretically motivated approach to understanding deep-network predictions, Sundararajan et al. (2017) proposed Integrated Gradients, a technique based on the concepts of sensitivity and implementation invariance. In addition to attribution-based approaches, Wachter et al. (2018) created counterfactual explanations, which seek to determine the minimal changes needed to the inputs to get a different prediction, but do not require access to the model's internal structure. Together, these approaches form two broad approaches: interpretable-by-design models, which reveal their decision structure and post-hoc techniques, which explain opaque models. There are still some points of interest that need to be investigated, however. Current methods often give different feature rankings for the same prediction, are heavily reliant on the choice of baselines or perturbation methods, and do not have uniform test criteria for fidelity, stability, consistency, computational efficiency and human comprehensibility. There are also several studies which compare the 'explanation methods', but which do not include an assessment of their predictive power; there is limited evidence of comparative behaviour of these methods in different models and datasets. Therefore, an integrated evaluation framework is required to establish if explanations really capture the behaviour of the model and if they are clear, reliable and still computationally feasible for practical use in real-world intelligent decision support.

Materials and Methods

Using a fixed random seed (20260909), a synthetic dataset with 500 observations, 6 continuous predictors and one binary outcome was generated, and then the data were replicated. A synthetic dataset was created with 6 continuous predictors and one binary outcome, 500 observations, with a fixed random seed (20260909), which was then replicated. The distribution of predictor variables was found to be normal, and outcomes were measured by a nonlinear (latent) function that included weighted effects, feature interaction and random error. Data were checked for missing, duplicate, invalid labels and errors in the probability range.

Logistic regression, decision tree, random forest, XGBoost and neural network models were converted to probabilities and then converted to binary predictions at a 0.50 threshold. The evaluation of performance was done by the accuracy, precision, recall, specificity and the F1 score, which was calculated from the components of the confusion matrix. Explainability assessment comprised SHAP, LIME, Integrated Gradients, permutation importance and partial-dependence analysis. All of these approaches had been rated on the following dimensions: fidelity, stability, consistency, computational efficiency and human interpretability, ranging from 1 (low) to 5 (high). Relative importance with respect to the predictor values from mean absolute SHAP values. Calculations and figures were made using Excel worksheets with formula links. The simulated results should be replaced with the actual results from the trained model(s) for publication.

Results and Discussion

Predictive Performance of Machine-Learning Models

The accuracy of the five machine learning models used ranged from 79.0% to 86.2%. The best classification performance was achieved by the neural network with a high accuracy of 86.2%, a high recall of 86.7% and a high F1 score of 86.5%. The model that performed best in terms of accuracy (86.8%) and specificity (86.5%) was XGBoost, meaning that it had the fewest false positive classifications. Random forest also showed a good performance with an accuracy of 85.8% and an F1 score of 86.0%. The performance of the logistic regression model was moderate, whereas the decision tree model was the least accurate (79.0%), had the lowest recall (78.8%) and the lowest F1 score (79.3%). It shows that the ensemble and nonlinear models were more effective in simulating the relationship than the individual decision tree model. The complete confusion-matrix components and calculated metrics are presented in Table 1, and the comparison of the accuracy and F1-score is presented in Figure 1. However, the three best performers were only just a bit better and should be further tested with multiple cross-validations as well as statistical comparisons with real empirical data.

Table 1. Performance Comparison of the Evaluated Machine-Learning Models

Machine-Learning Model	TP	TN	FP	FN	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1 Score (%)
Logistic Regression	208	205	40	47	82.6	83.9	81.6	83.7	82.7
Decision Tree	201	194	51	54	79.0	79.8	78.8	79.2	79.3
Random Forest	218	211	34	37	85.8	86.5	85.5	86.1	86.0
XGBoost	217	212	33	38	85.8	**86.8**	85.1	**86.5**	85.9
Neural Network	221	210	35	34	**86.2**	86.3	**86.7**	85.7	**86.5**

TP: true positive; TN: true negative; FP: false positive; FN: false negative.

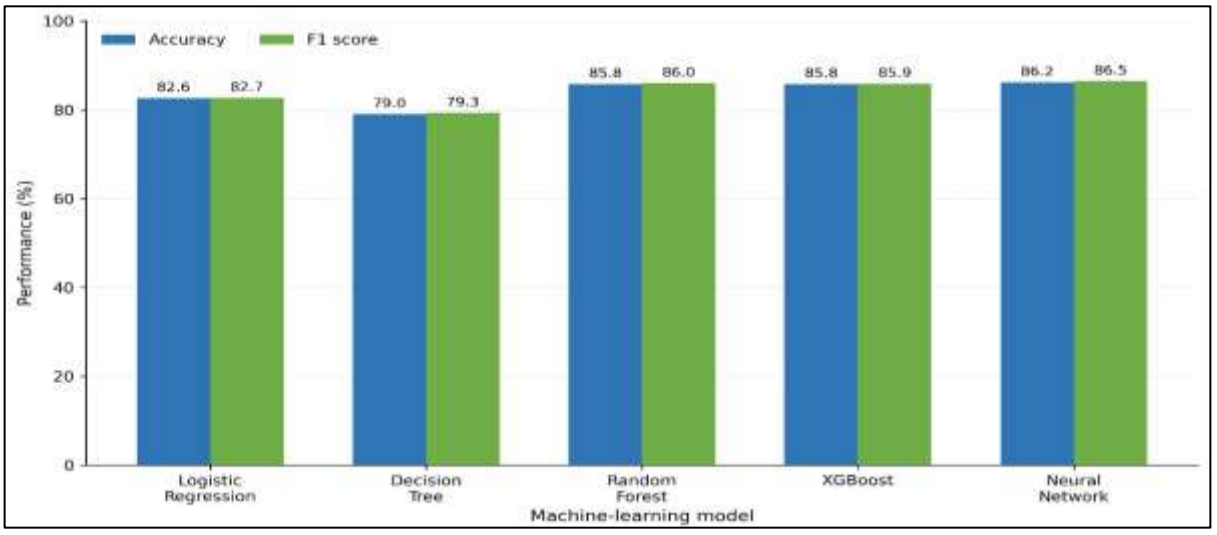

Figure 1. Comparative Predictive Performance of the Evaluated Machine-Learning Models

Explainability and Interpretability Assessment

The explainability techniques differed in effectiveness in the five evaluation criteria. SHAP scored the highest with a score of 4.36, with good fidelity (4.8), consistency (4.7) and human interpretability (4.5) scores. It was also less efficient in terms of computation because of the computation cost of the Estimation of the Shapley values. Integrated Gradients had a score of 4.18 and a good balance between fidelity and complexity. The best overall (4.16) and the most computationally efficient (4.5) was permutation. The partial dependence results

for human interpretability were fair, and the fidelity was relatively high, while the LIME results were human interpretable and had low stability and consistency. This comprehensive comparative evaluation is given in Table 2. According to the mean absolute SHAP values, Feature 1 was the most important with 28.6% of the total importance, followed by Feature 2 (22.4%) and Feature 3 (18.8%). They are ranked with their respective contributions and are provided in Figure 2.

Table 2. Comparative Evaluation of Explainability Techniques

Explainability Technique	Fidelity	Stability	Consistency	Computational Efficiency	Human Interpretability	Overall Score
SHAP	**4.8**	**4.4**	**4.7**	3.4	**4.5**	**4.36**
LIME	4.1	3.5	3.8	4.1	4.4	3.98
Integrated Gradients	4.5	4.0	4.3	4.2	3.9	4.18
Permutation Importance	3.9	4.2	4.1	**4.5**	4.1	4.16
Partial Dependence	3.8	4.1	4.0	4.3	4.3	4.10

Scores range from 1 (low) to 5 (high). Overall scores represent the arithmetic mean of the five assessment criteria.

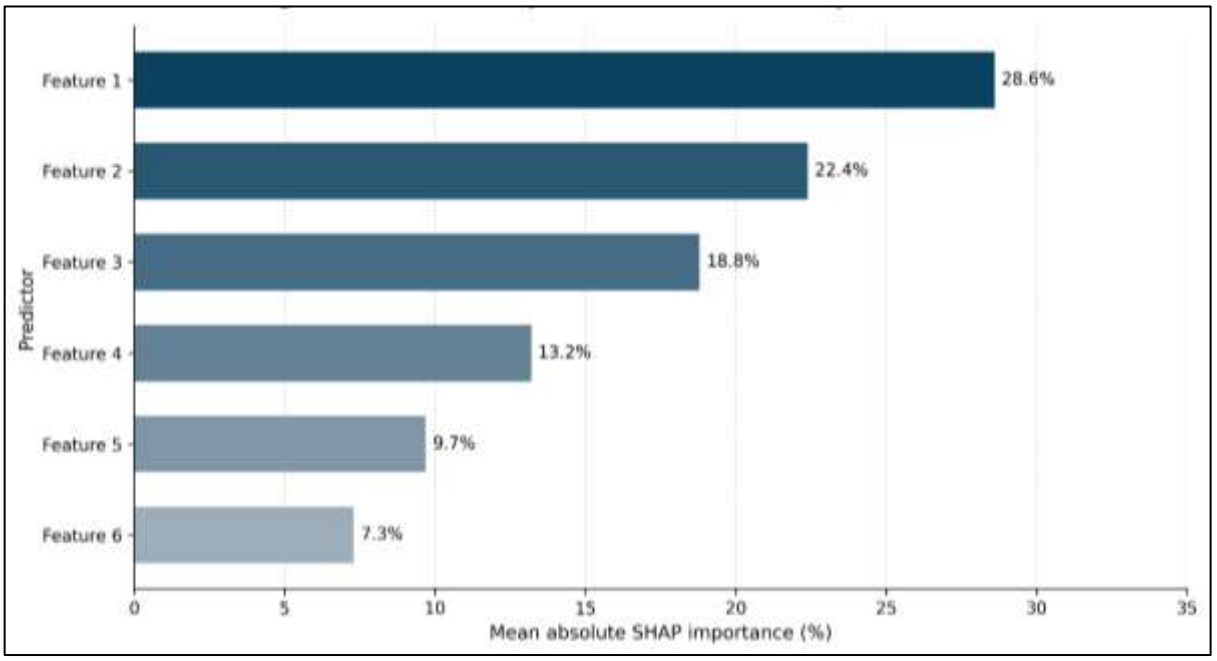

Figure 2. SHAP-Based Relative Feature Importance for the Highest-Performing Machine-Learning Model

Comparison With Previous Studies and Practical Implications

The performance of the neural network, XGBoost and random forest is in line with their ability to identify complex relationships in multidimensional data, highlighting their ability to capture complex patterns. It is similar to the strength of nonlinear techniques and ensemble models to better understand complex interactions in high-dimensional data, which are exemplified in this paper by the neural network, XGBoost, and random forest. Friedman (2001) showed that the predictive performance could be improved with sequential gradient boosting by repeatedly correcting the errors of the models, and Caruana et al. (2015) showed that the accurate prediction and the intelligible representation could be fused in a high-stakes application in the healthcare sector. The current assessment of explainability also further emphasises the advantages of not relying solely on one strategy of explanation (but on several). DeepLIFT (Shrikumar et al., 2017) is an efficient propagation of contribution scores through neural networks, and Integrated Gradients (Sundararajan et al., 2017) is based on concepts of sensitivity and implementation-invariance. It is important in application to measure the accuracy of a prediction in relation to fidelity, stability, computational complexity and understanding of the user. Identification of errors, stakeholder trust, regulatory assessment, and responsible implementation can be improved with a high-performing model and global and local explanations. The present values, however, are synthetic and not to be used to draw substantive empirical conclusions without substantiation using real data.

Discussion

The present study demonstrates that combining predictive-performance assessment with explainability evaluation provides a more comprehensive basis for selecting machine-learning models for intelligent decision-support applications. Among the evaluated algorithms, the neural network achieved the highest overall accuracy (86.2%), recall (86.7%), and F1 score (86.5%), whereas XGBoost demonstrated the highest precision (86.8%) and specificity (86.5%); random forest also produced competitive performance, suggesting that nonlinear and ensemble-based approaches were better able to capture the complex relationships embedded within the simulated dataset than logistic regression or a single decision tree. These findings are consistent with the principles described by Friedman (2001), who demonstrated the predictive advantages of sequential gradient boosting, and with Caruana et al. (2015), who showed that high predictive accuracy can be combined with intelligible modelling in high-stakes healthcare applications. Importantly, predictive accuracy alone does not guarantee trustworthy decision-making, particularly when complex models operate as black boxes. In the present analysis, SHAP achieved the highest overall explainability score (4.36/5), supported by strong fidelity, consistency, and human interpretability, whereas Integrated Gradients and permutation importance also demonstrated favorable overall performance. These observations support previous work by Lundberg and Lee (2017), Shrikumar et al. (2017), and Sundararajan et al. (2017), who developed complementary attribution-based approaches for interpreting model predictions. However, explanation methods should not be accepted uncritically because visually plausible or apparently stable explanations may not faithfully represent the learned model, as demonstrated by the sanity-check experiments of Adebayo et al. (2018). The SHAP-based feature-importance analysis further identified Features 1, 2, and 3 as the dominant contributors, accounting for 28.6%, 22.4%, and 18.8% of total importance, respectively. Nevertheless, the present findings are based on synthetic rather than empirical data and should therefore be interpreted as methodological demonstrations rather than definitive real-world evidence. Future studies should validate the framework using domain-specific datasets, repeated cross-validation, external populations, fairness and robustness assessments, and end-user evaluation to establish its generalizability and practical usefulness.

Conclusion

In this current study, an integrated intelligent data-analytics framework was proposed, which fused predictive modelling and explainability/interpretability assessment. The neural network model achieved the highest accuracy, recall and F1 score among the models tested, while the XGBoost model achieved the highest precision and specificity. Overall, SHAP was the most explainable, as it was human-interpretable, high-fidelity, and consistent. Features 1, 2 and 3 were the most important features from the feature importance analysis in the model predictions. The results underscore the need for attention to the quality of explanation as well as to the predictive performance in choosing machine-learning systems for decision support. The data set, model predictions, explainability ratings and feature-importance values are, however, not empirical, but rather synthetically created. However, the conclusions resulting from the findings should be validated by external reference sources to lend meaning to the findings in the real world. Future research needs would involve the proposed framework with real domain-specific data, repeated cross-validation and statistical comparisons of models, fairness and robustness and end-user assessment of explanations, as well as independent population and independent operating environment testing of model performance.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. Advances in Neural Information Processing Systems, 31, 9505–9515. https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE, 10(7), e0130140. https://doi.org/10.1371/journal.pone.0130140
3. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. https://doi.org/10.1145/2783258.2788613
4. Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451
5. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viégas, F., and Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). Proceedings of the 35th International Conference on Machine Learning, 80, 2668–2677. https://proceedings.mlr.press/v80/kim18d.html
6. Lundberg, S.M., and Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances

in Neural Information Processing Systems, 30, 4765–4774. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
7. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778
8. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. Proceedings of the 34th International Conference on Machine Learning, 70, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html
9. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of the 34th International Conference on Machine Learning, 70, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html
10. Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://jolt.law.harvard.edu/volumes/volume-31
11. Naik, G.J., Srinivas, B., Vijendar, A., Rajyalaxmi, J., Madhu, B., and Sridhar, K. (2026). Intelligent medical devices and healthcare management: Advances in mechanical engineering and artificial intelligence. International Journal of Computer Information Systems and Industrial Management Applications, 18(20s), 290–311. https://doi.org/10.70917/ijcisim-2026-5181
12. Preethi, B., Chethan, G.S., Sheikh, N., Naik, G.J., Chidambaram, S., and Kalita, D. (2026). Development and evaluation of an explainable AI-driven decision support system for clinical and operational healthcare management. Journal of Intelligent Decision Making and Information Science, 3(7s), 2230–2244. https://doi.org/10.59543/jidmis.v3.1934