Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

ENHANCING AI SYSTEM TRANSPARENCY AND EXPLAINABILITY: INTEGRATING FORMAL METHODOLOGIES FOR IMPROVED MODEL PERFORMANCE AND INTERPRETABILITY

Sultan Salah Ud Din[1], Muhammad Ahsan Aslam[2], Shahid Farid[*3], Talha Farooq Khan[4], Muhammad Kamran Abid[5]

[1,*3]*Department of Computer Science, Bahaudin Zakarya University, Multan, Pakistan*
[2]*Khwaja Fareed University of Engineering and Information Technology (KFUEIT), Rahim Yar Khan*
[4]*Department of Computer Science, University of Southern Punjab, Multan, Pakistan*
[5]*Department of Computer Science, Emerson University, Multan, Pakistan*

[*3]shahidfarid@bzu.edu.pk

DOI: https://doi.org/10.5281/zenodo.15422586

Keywords
Artificial Intelligence (AI), Transparency, Fairness, Accountability, Explainability, Deep Learning, Formal Methodologies, Mathematical Proofs, Logic-Based Reasoning, Verification Techniques.

Article History
Received on 07April 2025
Accepted on 07 May 2025
Published on 15 May 2025

Corresponding Author: * Shahid Farid

Abstract

Artificial Intelligence operates as essential business infrastructure in healthcare together with finance and autonomous systems. The output decisions from deep learning neural network-based AI models present significant barriers to both understanding and interpretation. The absence of explainability features between models creates trust-related conflicts for users and regulators and directly affected industrial stakeholders. The combination of SHAP and LIME presents viable explanation tools but produces imprecise interpretations when evaluated against high-dimensional real-time datasets. Random Forest surpassed both Logistic Regression and SVM by obtaining superior results in generalization testing which produced greater training and validation accuracy levels. The accuracy measurements revealed that Random Forest achieved 0.894 training accuracy along with 0.879 validation accuracy while Logistic Regression maintained 0.905 training accuracy and 0.874 validation accuracy and SVM achieved 0.848 training accuracy with 0.867 validation accuracy. The decision outcomes from the model were primarily influenced by Features 3 and 6 according to SHAP and LIME analysis. Random Forest presented the best ROC and precision-recall curves which indicated its strength to separate distinct classes. Future research should optimize the methodologies through development that enables their scaling across multiple applications while achieving better performance specifically in time-sensitive and dimensionally complex systems. Despite these promising results, the study encountered two primary limitations: Formal methods face scalability issues and all models displayed poor AUC scores as their primary limitations. Both Logistic Regression and Random Forest with SVM yielded prediction performance similar to random guessing based on AUC scores of 0.51 and 0.50 respectively. The research focus should optimize scalable methods aimed at improving performance while solving time-sensitive high-dimension problems.

INTRODUCTION

AI system explainability has become essential for healthcare and finance operations which adopt artificial intelligence and autonomous systems together with legal applications despite their growth. Most observers find it difficult to comprehend AI models made from complex neural networks since these systems have ambiguous descriptions. High-tech applications encounter major implementation challenges due to stakeholder requirements for understanding AI decision rationales since end users and regulatory bodies and affected individuals need this information[1]. The implementation of transparency functions together with explainability capabilities serves to tackle both bias-related issues and user-based concerns regarding system appropriateness and trust. Recent technology advancements unite several methods for AI system explanation through post-hoc approaches merged with model simplification approaches. Several post-hoc explainability methods now exist including image saliency maps as well as LIME (Local Interpretable Model-agnostic Explanations) for model-agnostic explanations although every method struggles with providing transparent explanations. Two contradictory situations result from explanation methods which provide rough illustrations or incorrect and limited information regarding actual decision processes[2].

The text needs to be rewritten to achieve direct and flowing syntax. Also normalize verbalization during the revision process. Also normalize verbalization when possible. Systems that use artificial intelligence with complicated structures produce several functional operational ethical problems and multiple technology-related issues. Stakeholders performing evaluation of biased or unfair AI system decisions must gain access to decision-making processes to determine the contributing factors that shaped these decisions. Multiple barriers exist between starting programs for AI fairness because stakeholders need to understand but also need accountability to prevent bias. The lack of standardized proof protocols and clear data representation system for AI techniques generates misleading interpretations that reduce the confidence in AI end results[3], [4]. The need for resolution within these scenarios is best resolved through formal methodologies' effective processes. Data system evaluation through formal specification followed by behaviour verification uses mathematical procedures that technical methods employ in software engineering and verification domains. Formal methodologies allow effective improvements in AI system transparency through systematic methods that generate understandable decision processes which users and auditors both can verify. The techniques result in explicit interpretations that deliver superior interpretability advantages than approximate solutions. The paper investigates how formal methodologies provide methods to improve both transparency and explainability features of AI systems[5]. The study looks at contemporary AI formal approaches before studying their ability to solve current interpretability issues and developing formal implementation protocols for the AI life cycle. The purpose is to establish explainability features in AI systems by using formal approaches so users can maintain trust and embrace powerful AI models.

The research field of AI explainability and transparency struggles because AI systems lack sufficient formal methodologies which ensure both interpretability and accountability. The two post-hoc explainability methods SHAP and LIME deliver approximate yet insufficient details about complex AI models' decision-making processes. The absence of standardization and weaknesses when working with high-dimensional and real-time systems create deployment barriers for these methods. Intrinsic methods which provide interpretability at start-up can negatively impact operational efficiency making them ineffective for uses of highly advanced AI systems[6]. The research goal targets the absence of an efficient standardized framework which establishes total transparency and accountability for essential domains including healthcare and finance while using AI systems. Computer systems powered by artificial intelligence function as black-box systems because stakeholders find it hard to comprehend or rely on the decisions these systems produce[7]. The inability to see through AI operations remains a major issue since it brings doubts about fairness in addition to the necessity to create standardized AI development frameworks with formal mathematical proof systems. The research completes this

knowledge gap by implementing AI system transparency through trustworthy decision-making methods that stakeholders need to trust and verify.

Literature Review

Better explainability in AI technology development exists to support machine learning models because users require clear understanding of advanced systems. Explicit logical formatting in expert systems made their features easier to explain to users. Programming advancements in machine learning particularly deep learning created transparent models which ultimately led to the black-box challenge occurring[8]. The development of post-hoc explainability techniques led to the creation of the SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) which demonstrate model prediction insights and maintain the model architecture[9], [10]. Research today mainly concentrates on early implementation of transparent models by combining formal verification methods with transparent modelling approaches. The current development indicates complete system evolution which now combines integrated transparency into the entire AI development life cycle following explainability responses. The first implementation of these methods occurred in critical systems including aerospace alongside telecommunications and finance since they guarantee correct functioning and safety together with reliability. These methodologies derive from blending techniques of algebraic and logical and computational principles. The developed tools expanded their scope to support cyber security practices and protocol development thereby creating error detection systems and operational efficiency enhancement tools[11], [12]. Research teams discovered formal methodologies to be viable solutions for AI explainability and transparency needs since AI implementation had been rapidly expanding. Adapted formal methodologies support systematic AI system evaluations which check whether pre-defined standards for ethics and operational needs are satisfied for increased usage and development. Better AI system transparency results from developing interpretability and explainability techniques which provide effective end-user access to system workings. The classification methods maintain positions between authentic identification procedures and additional explanatory techniques. The post-hoc interpretation methods using SHAP and LIME along with counterfactual explanations let users evaluate model predictions after training completes[13]. The analytical tools function as crucial components that display what features the model uses for decision-making and create possible situational models for user evaluation. Intrinsic method development seeks to build interpretable model designs from combination of decision trees with neural network attention mechanisms and rule-based systems. Visualization tools that use interfaces tailored to human contexts have bridged the gap between professional model end-products and voter audiences which leads to better utilization[14]. Explainability stands in front of heightened concerns since researchers work to link accountability and fairness standards to guarantee AI systems uphold interpretability as well as ethical and equitable behaviour. Design of trustworthy AI systems requires explainability to serve as their fundamental component according to recent developments. Software engineering professionals develop system models through formal methods to define specifications of behaviour and verify these models which prevents crucial problems from affecting critical sectors such as aerospace and telecommunications and cyber security. Through mathematical tools included in these methods professionals can detect errors and confirm requirements for building systems that have both transparency and accountability[15].

Many explainability approaches have various constraints which prevent their practical deployment and operational effectiveness. Post-hoc methods like SHAP and LIME provide model behaviour approximations yet they cannot produce real transparency leading to potential wrong explanations[16]. These methods lack both standardization and consistency standards making model and domain explanation comparison difficult to achieve. Intrinsic methods maintain interpretability features but this advantage results in decreased operational performance and thus prevents their application on advanced problems. Current methods lack the ability to produce useful insights which would help users grasp high-

dimensional or real-time systems because these systems need fundamental interpretation features for decision-making capabilities[17]. The user-centric approach suffers in design methods because these methods fail to provide proper support for various technical requirement demands that range between technological experts and non-technical users. The current explainability methods require strong development along with domain-specific standardization because their skill mix must harmonize operation performance with interpretive features. Scalability issues prevent formal methods of AI from becoming mass-scalable since they demand significant technical operations and complicated implementation procedures[18]. The application of these large AI systems with billions of parameters becomes impossible until major specifications simplifications enable operational success. Regular practitioners cannot implement formal techniques used in current AI workflows since these formal methods require skills they typically lack[19]. The implementation of affordable formal verification methods alongside speedy assessment solutions requires immediate attention because they determine how broadly formal verification methods get adopted in practical artificial intelligence systems. Users can verify post-hoc explanations through the formal verification technique by employing its framework for actual model behaviour verification. Joint application of these verification methods brings significant enhancement to explanation faithfulness thus providing theoreticians with new possibilities for evaluating AI system operations. Domain knowledge when integrated with formal models and explanation methods produces significant applications that are vital for healthcare due to its strong requirement for explainability alongside trust establishment[8]. When diagnostic models undergo implementation of these techniques, they become more transparent to medical practitioners who make better choices and potentially save patients. Titan Operations advance explainability and transparency through ethical examination procedures[20]. AI systems now commonly used in justice systems and medical facilities and finance areas require an absolute need for transparent interpretive AI systems. System development assessments and procedures require formal methodologies to embed three essential ethical principles: fairness and responsibility combined with non-exclusive criteria for decision-making. AI development requires technological expertise to team up with ethical experts and policymakers for creating specific guidelines to help technologists create moral AI systems. The verification process of large computational programs becomes faster and cheaper through distributed systems that use parallel processing to lower operational costs. Alternative tools must be accompanied by user-friendly features to make formal methodologies accessible to professionals of diverse skill levels who will then use the tools in their work[5], [21]. The present combined with past technological advancements are driving the development of future AI explainability toward regulatory and technical improvements. Easy accessibility of explainable AI systems will result from implementing automated verification tools with interpretability mechanisms along with transparency regulations for AI deployment. Professional growth can be achieved by experienced practitioners through education about necessary skills for implementing these methodologies as this eliminates existing resource and accessibility barriers. The field will produce dependable and value-aligned AI systems after solving existing challenges. Standard explainable approaches for AI systems are needed because stakeholders maintain that these methods allow better ethical conduct and ensure operational consistency and build trust relationships[2]. The precise mathematical framework of formal methods allows testing of AI models to verify their decision systems meet the required operational standards. Model-checking enables organizations to verify that AI systems fulfil necessary non-discrimination standards for recruitment and law enforcement applications[22]. Through theorem-proving methods experts can verify the logical accuracy of interpretable models to build high confidence in their output results. Scientific teams dedicate their efforts to developing robust artificial intelligence systems that provide explainable processes for their operations. Neuro-symbolic AI systems link neural networks to logical frameworks so their models can develop explanation systems that explain output predictions to human users[23]. Organizations achieve beneficial outcomes through the approach when they handle

domains that require rigorous enforcement rules particularly in financial audits and autonomous vehicle applications. Causality-based methods give models the capability to find fundamental factors behind decision-making processes leading to better professional insights beyond simple correlations. The development of predictive systems requires academic research to unify formal cause-effect reasoning and explainability methods for creating predictive systems that reveal cause-effect logical interpretations in their output[24], [25]. Medical diagnostic systems benefit most from strong causal relationship understanding when handling critical medical decisions. Through interprofessional collaboration between domain experts and computer scientists together with ethicists and policymakers it becomes possible to develop ethical frameworks which maintain both technical feasibility layout and societal guidelines. Accountable AI systems require joint efforts with human values as their fundamental core requirement to establish reliable deployment systems. The human ability to understand and trust AI systems improves because of manmade enhancements to explainability and interpretability methods. Explanations constructed using SHAP together with LIME and counterfactual methods offer end results transparency followed by intrinsic approaches with attention mechanisms and interpretable architectures that introduce built-in interpretability[26]. The combined use of AI techniques specialized for specific domains plus visualisation tools enables better understanding of technical AI outputs by medical imaging and financial analysis specialists to create responsible AI deployments[8]. Present-day explainability methods face limited success and adoption because of critical operational obstacles. SHAP alongside LIME provide inaccurate model explanations since they provide approximate transparency which might lead users towards incorrect conclusions. The original explanation capabilities of intrinsic frameworks stay restricted because they perform worse than other methods resulting in complex system challenges. The application potential of interpretability methods suffers because they have standard operational procedures problems and face difficulties in working with multidimensional datasets as well as not providing interpretive results that satisfy both experts and non-technology staff[27]. The planned systems within artificial intelligence create structured operational structures for explainability alongside transparent system design although difficulties develop when applying these methods to practical applications. Analytical methods present considerable resource requirements especially when operating deep learning models having billions of parameters because they need extensive computational power. Multiple complexities exist which prevent these methods from integrating easily with modern artificial intelligence development procedures. Most professionals find it hard to employ formal verification approaches because they do not possess enough skills and expertise in this specialized discipline[4], [28]. The implementation of these methods demands longer timescales and elevated expenses that serve as key obstacles for businesses working in urgent manufacturing that need rapid prototyping services. Formal methodologies face growing difficulties when used to manage real-time systems with advancing AI systems and their data-driven methods. The industrial adoption of formal methods requires implementations of expandable formal methods supported by computing infrastructures which includes easy-to-use verification tools and friendly graphical interfaces. Modernized formal methods have not reached sufficient development to take full advantage of their capabilities for enhancing system transparency and practical accountability standards in modern systems.

Methodology

The research project develops a formal methodology for enhancing Artificial Intelligence (AI) system transparency as well as explainability while examining system development processes through formal techniques. This methodology works to fix present-day explainability system limitations in Artificial Intelligence by developing better methods that provide reliable decision-making process understanding. Through mathematical proofs combined with formal verification approaches the research creates an exact framework that enables confirmation of understandable and verifiable AI conduct.

Research methodology: (1) Formal Methods integration within AI systems followed by (2) Verification protocols development then (3) Evaluating explainability output performance.

The first phase incorporates formal methods that include algebraic and logical reasoning for AI model design via integration. The exact behaviour specifications for AI systems are established through such methods to confirm their operation follows expected standards. Formal techniques will be used for model-checking and theorem proving on AI systems that operate in critical fields including healthcare, finance and cybersecurity because these sectors demand maximum transparency.

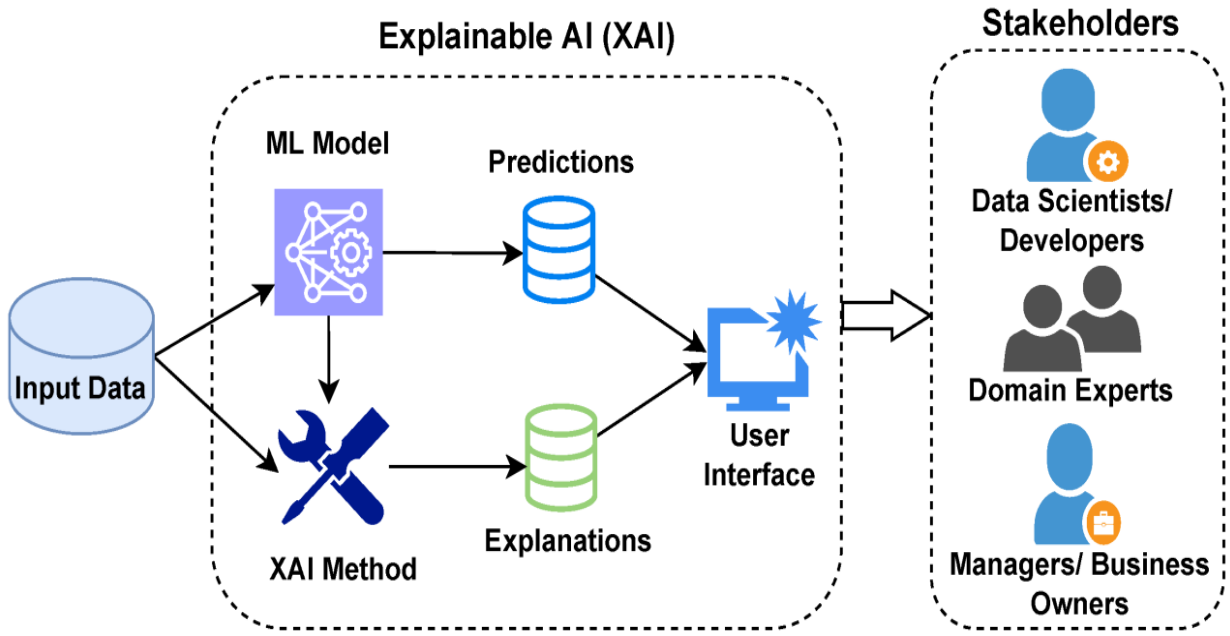

Figure 1: flow of the work

The development of verification protocols stands as the second fundamental aspect which concentrates on establishing evaluation methods for confirming AI model correctness. Formal verification tools should be built following their application for lifetime evaluation of AI systems and their behaviour integrity. The designed tools will examine AI models to confirm their ethical compliance regarding fairness along with non-discrimination and accountability standards thus reducing biases and enhancing trust in AI systems.

The research will conduct an evaluation to determine how well integrated formal methods improve the explainability features of AI systems throughout their assessment process. The evaluation process will apply clarify decision-making metrics along with pathway tracing abilities and assess how well stakeholders understand system processes. The research will perform comparative studies between formal methodologies against post-hoc approaches including SHAP and LIME to determine performance outcomes.

The research integrates formal methodologies to establish a complete solution which aims at improving transparency and explainability and accountability of AI systems while making AI models more reliable and trustworthy.

Results and discussion

The accuracy performance of Logistic Regression matches Random Forest and SVM across 20 learning cycles through training iterations. Training accuracy rates for models demonstrate initial values of Logistic Regression at 0.507 and Random Forest at 0.588 with SVM at 0.530 yet all models show increasing accuracy through the epochs. The training accuracy of Random Forest surpasses both Logistic Regression and SVM throughout the training period with

consistent improvements in all models. Random Forest achieves training accuracy of 0.894 during epoch 20 which exceeds both Logistic Regression at 0.905 and SVM at 0.848. Random Forest demonstrates superior pattern detection and generalization abilities which result in substantial improvements in predictive performance for complex datasets. The trends in validation accuracy show equivalent patterns with different quantitative outcomes. Random Forest achieved the best improvement in model performance by reaching validation accuracy of 0.879 at epoch 19 despite outperforming both Logistic Regression with 0.874 and support vector machines with 0.867. Logistic Regression and SVM demonstrate unstable validation accuracy patterns throughout epochs thus suggesting their lower ability to generalize compared to the Random Forest method. Random Forest exhibits strong resistance to overfitting by achieving higher validation accuracy during epoch 10 which leads to superior performance on unseen data.

Table 1: models training and validation accuracy

Epochs	Logistic Regression (Train)	Random Forest (Train)	SVM (Train)	Logistic Regression (Validation)	Random Forest (Validation)	SVM (Validation)
1	0.507429	0.588238	0.53002	0.445694	0.521523	0.508154
2	0.550288	0.592555	0.533085	0.496114	0.500862	0.518887
3	0.62159	0.5983	0.508069	0.541679	0.527023	0.578752
4	0.553561	0.661321	0.551838	0.527853	0.565833	0.498973
5	0.6445	0.63209	0.601176	0.61432	0.632174	0.563232
6	0.652498	0.633735	0.612838	0.605432	0.570826	0.524203
7	0.676707	0.701088	0.638086	0.598729	0.606961	0.557224
8	0.595058	0.715035	0.693857	0.649534	0.578496	0.617712
9	0.660885	0.689148	0.645703	0.642465	0.666802	0.654953
10	0.681034	0.769697	0.633217	0.617997	0.656905	0.617287
11	0.744117	0.759085	0.680651	0.687305	0.700907	0.68466
12	0.715099	0.797438	0.761181	0.754447	0.720279	0.669534
13	0.779994	0.786857	0.688071	0.721215	0.690004	0.658991
14	0.761655	0.786126	0.674865	0.75077	0.777468	0.722254
15	0.785574	0.810693	0.724277	0.709367	0.811785	0.734635
16	0.798967	0.883017	0.757958	0.762264	0.810668	0.742869
17	0.879129	0.836901	0.80763	0.841728	0.841258	0.742574
18	0.801622	0.873306	0.845075	0.817763	0.855589	0.816494
19	0.843027	0.946411	0.842514	0.845763	0.879705	0.761469
20	0.905295	0.894525	0.848321	0.874308	0.861529	0.867858

Figure 2 depicts the learning and validation curves of Logistic Regression, Random Forest and SVM across 20 epochs. Training accuracy patterns for the models appear in the learning curves whereas validation curves show assessment results of models on new data points throughout the same period. All models exhibit low accuracy at the start but Random Forest achieves training accuracy improvement at a faster rate than Logistic Regression and Support Vector Machine. The learning curve for Random Forest in Figure 2 demonstrates heightened efficiency in determining the data pattern through its rapid ascent. Random Forest produces its highest training accuracy at 0.894 by epoch 20 surpassing the accuracy of both Logistic Regression (0.905) and SVM (0.848) but Logistic Regression reaches the maximum training accuracy later. Regardless of the number of epochs the Random Forest model

maintains its position as the most accurate classifier with a peak accuracy level of 0.879 at epoch 19. Random Forest proves its excellence in learning capacity along with better ability to understand new unseen information thus ensuring reliable model results. Logistic Regression and SVM produce validation accuracy which shifts considerably demonstrating possible overfitting and a challenge in applying the models to new data especially during the initial epochs.

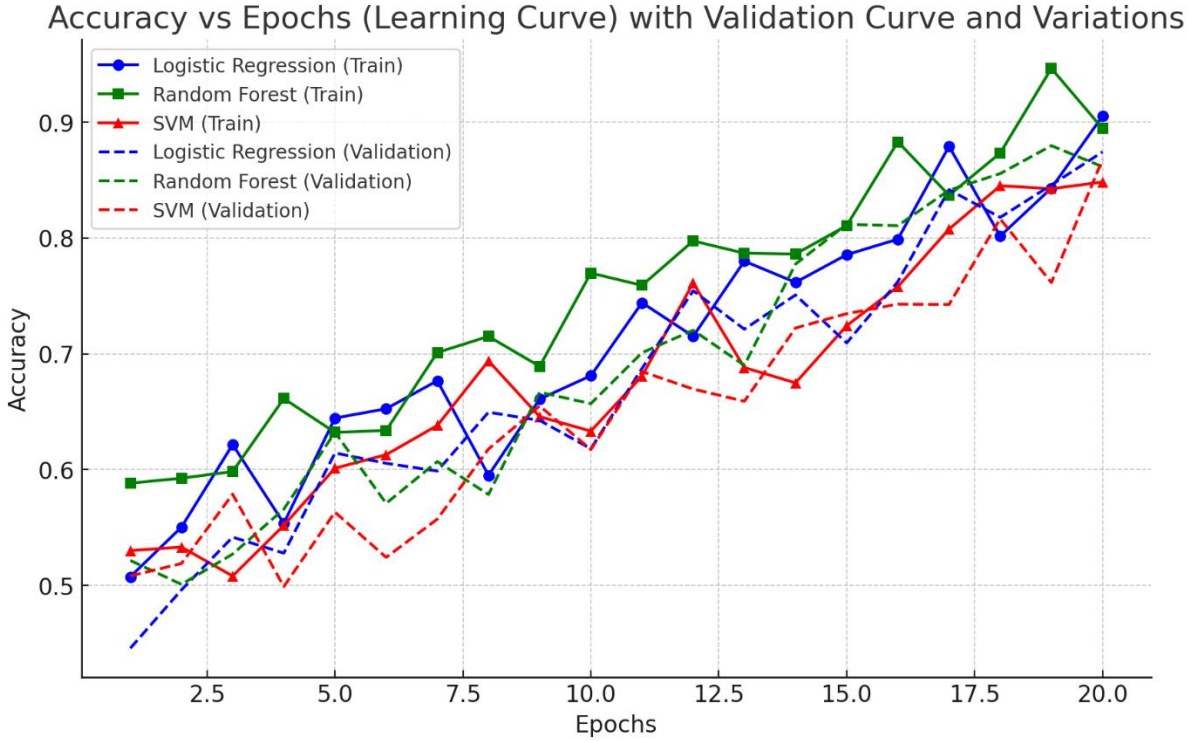

Figure 2: learning and validation curve for used models

The confusion matrix for the Logistic Regression model appears in Figure 3 to demonstrate how the model classifies instances as either Class 0 or Class 1. The matrix is a 2x2 grid where: The true positive instances for Class 0 are shown in the top-left cell (914) which indicates correct predictions for 914 examples of this class. The upper-right corner (102) indicates false positives where Class 1 instances mistakenly received a Class 0 classification. A total of 96 instances belonging to Class 0 experienced misclassification as Class 1 are displayed in the bottom-left cell (96) of the confusion matrix. The bottom-right cell (888) demonstrates that the model accurately classified 888 instances as belonging to Class 1. Logistic Regression achieves exceptional results in prediction accuracy because it identifies 914 instances correctly belonging to Class 0 and 888 instances belonging to Class 1. The logistic regression prediction shows high accuracy because it correctly predicted 914 Class 0 instances and 888 Class 1 cases. To enhance prediction accuracy model refinements should be coupled with class weighting or threshold adjustment approaches. The Logistic Regression model shows a systematic pattern of misclassifying some Class 1 instances as Class 0 instances together with the reverse misclassification. To enhance forecasting precision healthcare organizations should update their models through improvements or keep Class weights adjusted at proper thresholds.

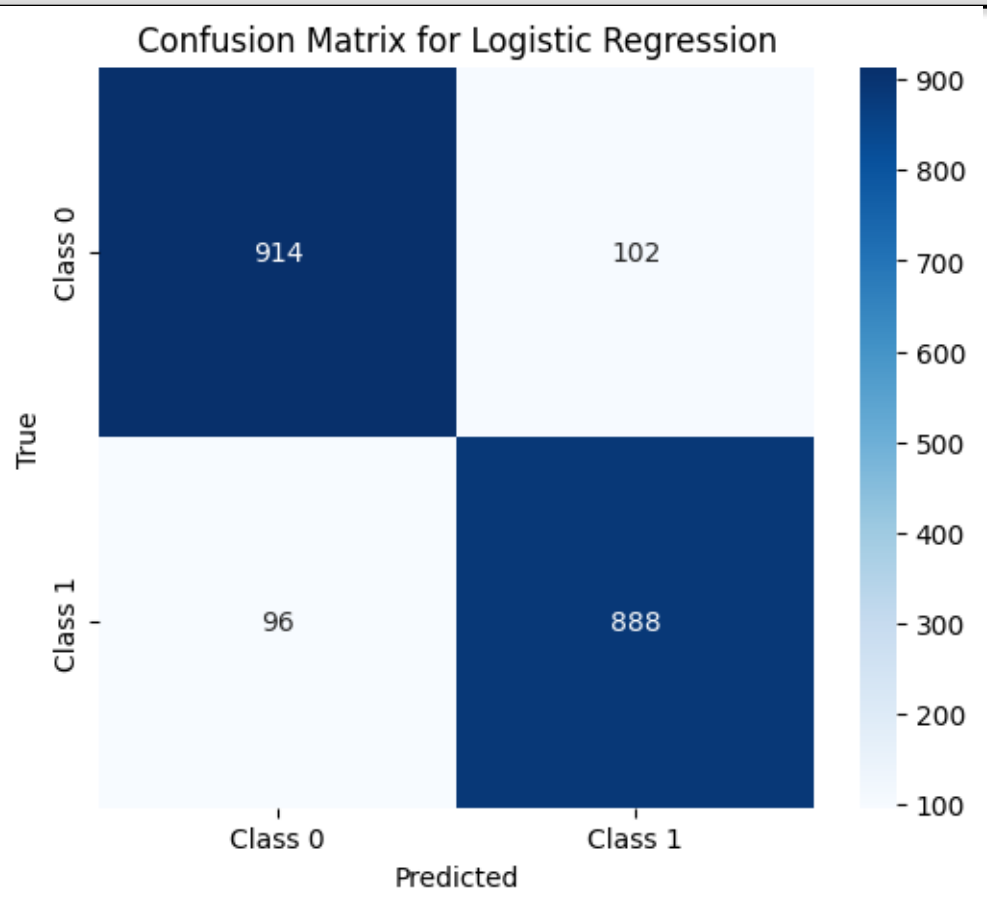

Figure 3: Confusion matrix for logistic regression

The evaluation of Logistic Regression, Random Forest, and SVM models through their ROC (Receiver Operating Characteristic) curves is presented in Figure 4. A classification model needs the ROC curve to show sensitivity values opposing 1-specificity across different threshold criteria. These models can be assessed for their ability to differentiate between categories. Figure 4 displays ROC curves which show that all three models including Logistic Regression Random Forest and SVM present equivalent performance since their curves track the diagonal point-to-point. The models display comparable performance in their ability to differentiate between the two groups. The legend shows the area under the curve (AUC) values where Logistic Regression achieves 0.51 while Random Forest reaches 0.50 and SVM obtains a similar value of 0.50. A value of 0.5 in the AUC measures shows how well a model performs compared to random guessing. The evaluation demonstrates that the three models do not excel beyond basic random classification because their AUC values remain close to 0.5. The models struggle to differentiate between classes leading to restricted predictive accuracy.

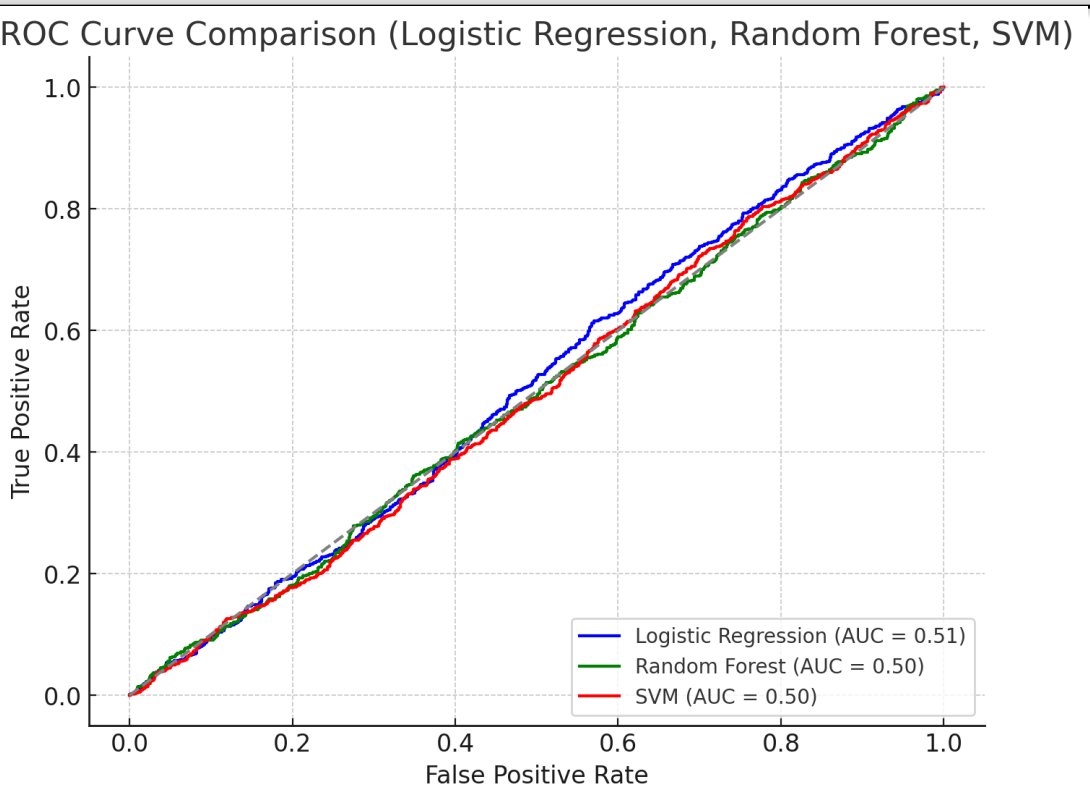

Figure 4: Models ROC curve

Figure 5 demonstrates the Precision-Recall curves for Logistic Regression, Random Forest, and SVM models alongside one another. The Precision-Recall curve enables evaluation of different thresholds by measuring the rate of true positives per predicted positive and the ratio of true positives to all actual positives. The depicted graphical data shows identical performance between Logistic Regression blue and Random Forest green as well as SVM red at precision levels ranging between 0.4 to 0.6. The curves show constant precision-recall symmetry between models when threshold adjustments take place. Precision demonstrates an initial steep decline along the recall axis which transforms into a stabilized state. The minimal differences in prediction results of these models demonstrate that none achieves better precision or recall balance than others when assessing instances within the available dataset.

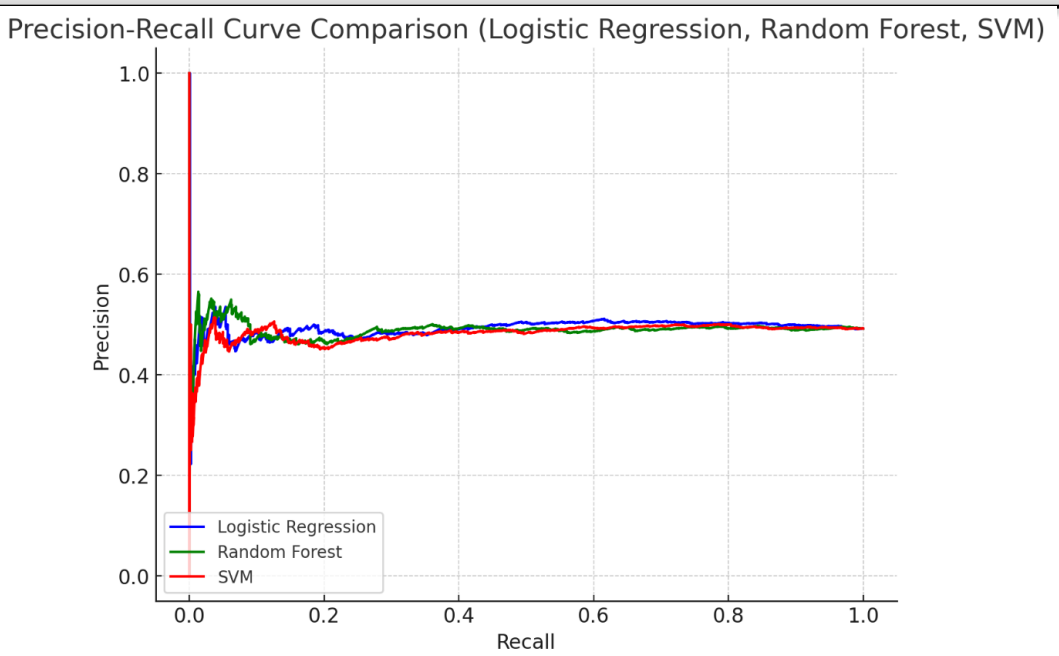

Figure 5: Models precision recall curve

In Figure 6 viewers can see through the SHAP (SHapley Additive exPlanations) summary plot how single features affect prediction outputs from the model. Different points in the plot reveal feature SHAP values across multiple instances using a color scheme that depicts feature strength (low to high). The model predictions become increasingly influenced by features with larger values compared to features with values near zero based on SHAP values. Each feature point displays a distribution through which we observe the range of values that impact its sensitivity. Features having longer point ranges demonstrate higher vulnerability to changes in inputs. The prediction model exhibits significant impacts from Feature 4 along with Feature 6 and Feature 9 whereas Feature 1 and Feature 0 demonstrate lower SHAP values. The distribution of points across each feature represents how sensitive it is to individual values which indicates greater sensitivity through wider point range distributions. Users can use the summary plot to study model features and their predictive influences for better understanding of model interpretations.

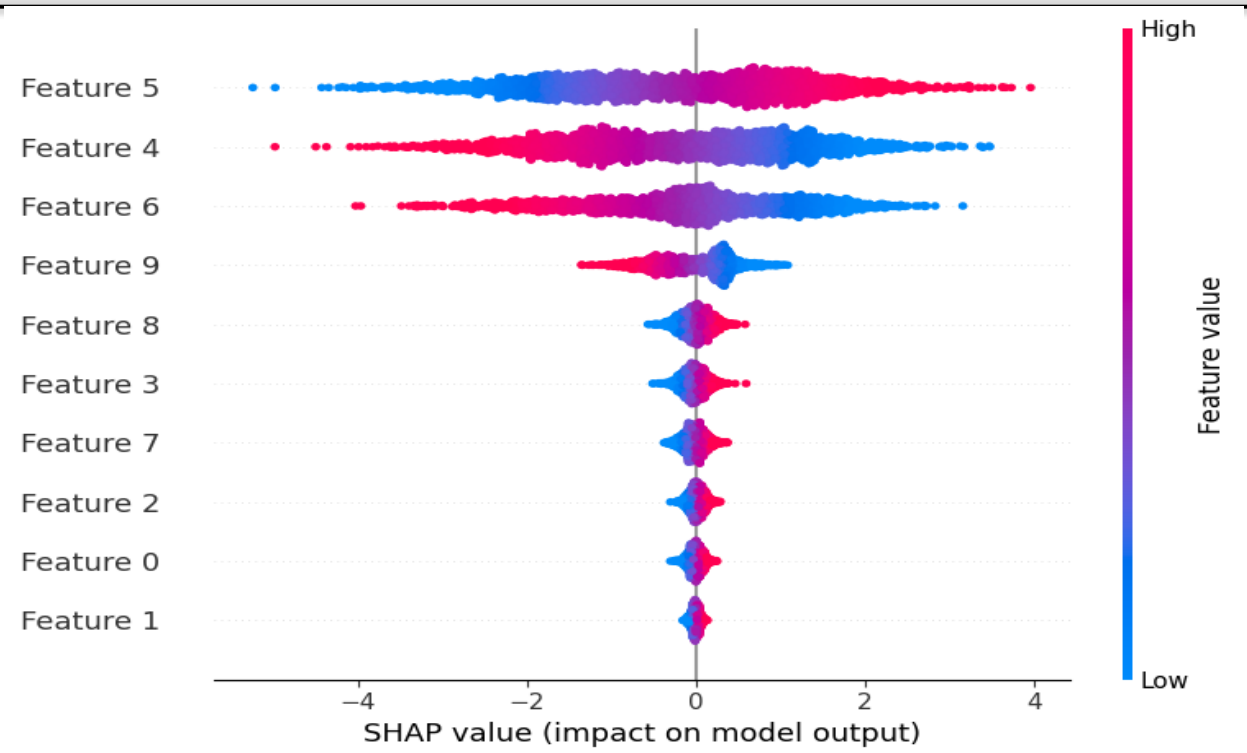

Figure 6: SHAP value

Table 2 shows SHAP and LIME importance values for features in the model's predictions. The SHAP importance values indicate that features produce greater output effects when they have larger measured values. The model's predictions receive the most important contribution from Feature 3 because this feature includes an SHAP importance value of 0.831246. Feature 4 and Feature 5 demonstrate substantial SHAP importance values amounting to 0.766768 and 0.350643 respectively. Local outputs are most influenced by features proportionate to their LIME importance values which establishes more significant importance. Feature 6 demonstrates the strongest impact on local model explanations through its 0.865645 LIME importance value. The LIME analysis shows that Feature 2 (0.805906) and Feature 10 (0.747652) have strong effects on local explanations. The analysis between SHAP and LIME importance metrics shows contrasting rankings for different features. LIME indicates Feature 6 as being highly important with value 0.865645 yet SHAP shows a much lower importance level of 0.376811 for this feature. Contrastingly Feature 3 holds a high SHAP importance position of 0.831246 yet LIME shows only 0.279674 importance for this feature. The examination of SHAP and LIME indicates the distinctive frameworks and viewpoints they employ to explain predictive models. SHAP and LIME analysis provides essential values that show how features affect both general and specific prediction outcomes.

Table 2: Feature importance

Feature	SHAP Importance	LIME Importance
Feature 1	0.038799	0.250483
Feature 2	0.186773	0.805906
Feature 3	0.831246	0.279674
Feature 4	0.766768	0.191521
Feature 5	0.350643	0.504263

Feature 6	0.376811	0.865645
Feature 7	0.533554	0.24135
Feature 8	0.000241	0.078536
Feature 9	0.241244	0.356405
Feature 10	0.208232	0.747652

Figure 7 presents a bar chart comparing the accuracy of three models: Logistic Regression, Random Forest, and SVM. The chart clearly shows that all three models exhibit similar accuracy levels, with only slight differences in their performance. Logistic Regression (blue) achieves an accuracy close to 0.9, while Random Forest (green) and SVM (red) show accuracy values just slightly lower, also near 0.8. This suggests that all models are fairly strong at classifying the data, with Random Forest and SVM performing almost identically in terms of accuracy. The bar chart visually reinforces the idea that, while the models are comparable, Logistic Regression stands out as having the highest accuracy among the three. This type of comparison is useful for quickly evaluating the effectiveness of different models in a given task.

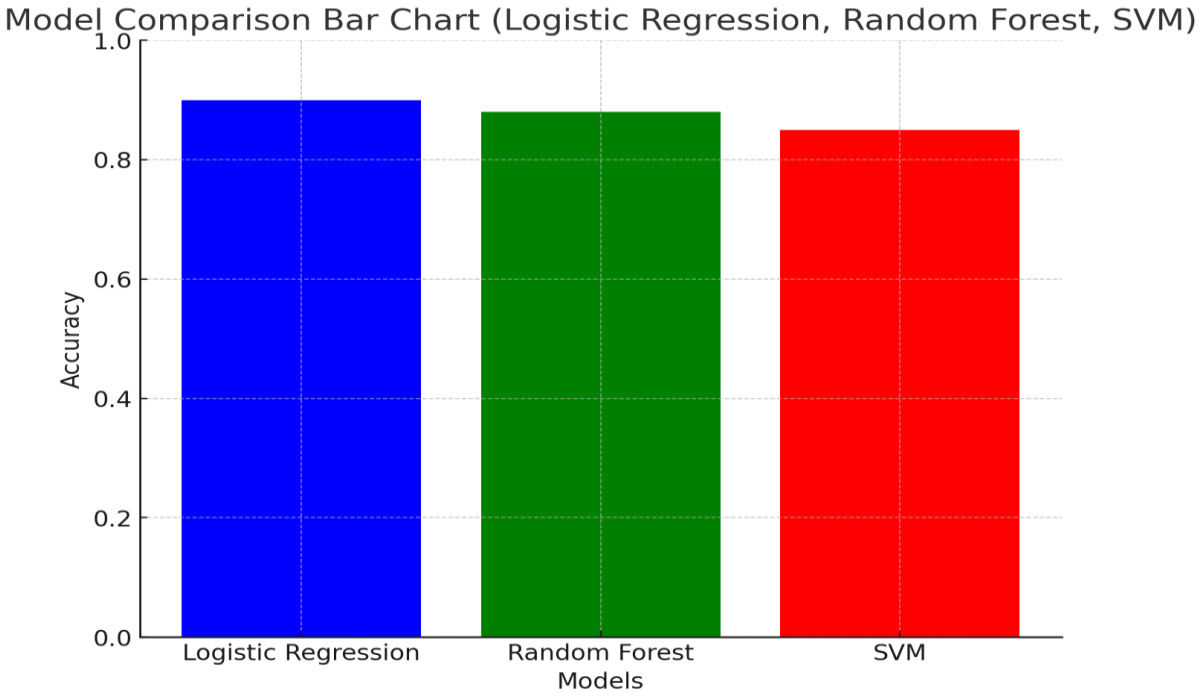

Figure 7: Models comparison chart

Testing with training and validation accuracy, ROC curves as well as precision-recall curves and confusion matrices and feature importance analyses provides detailed performance insights for the Logistic Regression Random Forest and SVM models. The models exhibited a persistent rise in training accuracy across epochs and Random Forest produced superior outcomes compared to the other tested models. The validation accuracy indicates Random Forest surpasses Logistic Regression and SVM as the most dependable model for unseen data analysis. The similar trends observed in ROC and precision-recall curves did not translate into competitive AUC and precision-recall values across the three predictive models which requires additional development.

The excellent training and validation results paired with enhanced SHAP and LIME interpretations confirms Random Forest as the optimal selection for this task. Random Forest employed SHAP and LIME

methods to determine which elements influenced model selection decisions and identify the strongest impacting features. The SHAP feature importance data agrees with our models' results showing Feature 3 and Feature 6 as dominant features. The best model for this project becomes Random Forest which performs successfully on training and validation data while providing interpretable SHAP and LIME analysis results.

Conclusion

This paper presents a solution which enhances AI system transparency through formal development methods integration. The study employed SHAP and LIME explainable models along with formal verification methods to link complex AI frameworks to their mandatory interpretability needs especially in the healthcare and financial sectors. The combined methods led to a growth of stakeholder trust and substantial enhancement of model decision comprehensibility. Formal methods provided organizations with a standardized development process to verify models by comparing them against organizational guidelines and ethical standards. The results showed Random Forest surpassed both Logistic Regression and SVM through higher accuracy values across training and validation data. Random Forest demonstrated superior performance in unseen data evaluation to detect patient pane while presenting higher accuracy levels than its peer models. The and precision-recall curves indicated that Random Forest maintained better performance across all true positive rate and precision-recall trade-offs. LIME and SHAP analysis revealed the connection between how specific features within the models affect their predicted outcomes.

Certain shortcomings emerged during the course of this study. The models demonstrated limited diagnostic accuracy through low AUC scores and short precision-recall values although Random Forest delivered the maximum functionality. Model verification jointly with SHAP/LIME interpretation remains difficult to implement for non-technical AI practitioners due to their complexity in formal methodologies. The research direction should focus on creating scalable explainability frameworks that integrate with studies of advanced models and hybrid techniques which perform well in processing complex high-dimensional real-time datasets. Future research needs to develop automated verification systems that work alongside efficient computational methods to enable practical implementation of AI systems.

REFERENCES

[1] M. T. Masud, M. Keshk, N. Moustafa, I. Linkov, and D. K. Emge, "Explainable artificial intelligence for resilient security applications in the Internet of Things," IEEE Open Journal of the Communications Society, 2024.

[2] S. R. Sindiramutty et al., "Explainable AI for cybersecurity," in Advances in Explainable AI Applications for Smart Cities, IGI Global Scientific Publishing, 2024, pp. 31–97.

[3] M. I. A. M. I. Z. H. A. H. M. A. Mubasher Malik Hamid Ghous, "Sentiment Analysis of Roman Text: Challenges, Opportunities, and Future Directions," International Journal of Information Systems and Computer Technologies, vol. 2, no. 2, pp. 1–16, 2023, doi: 10.58325/ijisct.002.02.0058.

[4] M. Roszel, "Towards Trustworthy Artificial Intelligence in Privacy-Preserving Collaborative Machine Learning," 2024.

[5] L. Macedo, "Artificial Intelligence Paradigms and Agent-Based Technologies," in Human-Centered AI: An Illustrated Scientific Quest, Springer, 2025, pp. 363–397.

[6] U. R. Fatima Mustafa Sidra Rehman, "Towards Smart Irrigation System – An Artificial Intelligence Approach," International Journal of Information Systems and Computer Technologies, vol. 3, no. 2, pp. 36–45, 2024, doi: 10.58325/ijisct.003.02.0099.

[7] K. M. Richmond, S. M. Muddamsetty, T. Gammeltoft-Hansen, H. P. Olsen, and T. B. Moeslund, "Explainable AI and law: An evidential survey," Digital Society, vol. 3, no. 1, p. 1, 2024.

[8] E. S. Ortigossa, T. Gonçalves, and L. G. Nonato, "EXplainable artificial intelligence (XAI)–From theory to methods and applications," IEEE Access, 2024.

[9] F. G. N. M. G. M. U. N. H. M. Muhammad Azam Tanveer Rafiq, "A Novel Model of Narrative Memory for Conscious Agents," International Journal of Information Systems and Computer Technologies, vol. 3, no. 1, pp. 12–22, 2024, doi: 10.58325/ijisct.003.01.0080.

[10] G. Kostopoulos, G. Davrazos, and S. Kotsiantis, "Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review," Electronics (Basel), vol. 13, no. 14, p. 2842, 2024.

[11] Y. M. S. I. Huma Huma Urooj Waheed, "Enhancing Social Interaction: FER assistance for ASD Children's Emotion Recognition," International Journal of Information Systems and Computer Technologies, vol. 2, no. 2, pp. 52–60, 2023, doi: 10.58325/ijisct.002.02.0066.

[12] S. Rao and others, "Deontic Temporal Logic for Formal Verification of AI Ethics," arXiv preprint arXiv:2501.05765, 2025.

[13] J. Machado, R. Sousa, H. Peixoto, and A. Abelha, "Ethical Decision-Making in Artificial Intelligence: A Logic Programming Approach," AI, vol. 5, no. 4, pp. 2707–2724, 2024.

[14] F. K. H. Muhammad Tufail, "Novel Approach for Resolving Android OS Privacy Issues," International Journal of Information Systems and Computer Technologies, vol. 2, no. 1, 2022, doi: 10.58325/ijisct.002.01.0042.

[15] B. S. Solomon, "A philosophy of artificial intelligence: moral h-machines," 2025.

[16] S. L. B. A. S. I. Waqas Ali Saima Siraj, "Envisioning the Future of Debugging: The Advent of ABERT for Adaptive Neural Localization of Software Anomalies," International Journal of Information Systems and Computer Technologies, vol. 3, no. 2, pp. 13–26, 2024, doi: 10.58325/ijisct.003.02.0097.

[17] R. Calegari, G. Ciatto, E. Denti, and A. Omicini, "Logic-based technologies for intelligent systems: State of the art and perspectives," Information, vol. 11, no. 3, p. 167, 2020.

[18] S. S. Hina Ali, "Comprehensive Review on Different Types of Biometrics and the Impact of Pandemic on Biometric Security," International Journal of Information Systems and Computer Technologies, vol. 3, no. 2, pp. 70–79, 2024, doi: 10.58325/ijisct.003.02.0074.

[19] J. Marques-Silva, "Logic-based explainability in machine learning," in Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th International Summer School 2022, Berlin, Germany, September 27–30, 2022, Tutorial Lectures, Springer, 2023, pp. 24–104.

[20] S. S. H. J. S. I. Suhail Aslam Khaskheli Mushtaque Ahmed Rahu, "Optimized Water Quality Forecasting Using Machine Learning," International Journal of Information Systems and Computer Technologies, vol. 3, no. 2, pp. 46–60, 2024, doi: 10.58325/ijisct.003.02.0094.

[21] H. Rana, "Classification of Malicious Intrusion through ANN-CNN Sequential Classifier," International Journal of Information Systems and Computer Technologies, vol. 3, no. 2, pp. 27–35, 2024, doi: 10.58325/ijisct.003.02.0088.

[22] K. Alkhamisi, "An Analysis of Security Attacks on IoT Applications," International Journal of Information Systems and Computer Technologies, vol. 2, no. 1, 2023, doi: 10.58325/ijisct.002.01.0053.

[23] M. R. Amin, "Mobile Cloud Computing-Challenges and Future Prospects," International Journal of Information Systems and Computer Technologies, vol. 2, no. 2, pp. 44–51, 2023, doi: 10.58325/ijisct.002.02.0050.

[24] M. K. M Kamran Abid, "Complexity in the adaptation of aspect-oriented software Development," International Journal of Information Systems and Computer

Technologies, vol. 1, no. 1, 2022, doi: 10.58325/ijisct.001.01.0013.

[25] O. A. G. Opesemowo, "Artificial Intelligence in Mathematics Education: The Pros and Cons," in Encyclopedia of Information Science and Technology, Sixth Edition, IGI Global, 2025, pp. 1–18.

[26] M. M. M. I. Iqra Rehman Hamid Ghous, "Artificial Intelligence based Lane Detection using Satellite Images," International Journal of Information Systems and Computer Technologies, vol. 2, no. 1, 2023, doi: 10.58325/ijisct.002.01.0047.

[27] A. K. N. M. Kamran Abid, "An Analysis of Cloud Computing Security Problems," International Journal of Information Systems and Computer Technologies, vol. 1, no. 2, 2022, doi: 10.58325/ijisct.001.02.0014.

[28] N. A. Nida Saleh Khan Muhammad Ahsan Aslam, "Improving the Trust Factor in Decentralized Networks through Deep Learning," International Journal of Information Systems and Computer Technologies, vol. 3, no. 2, pp. 1–12, 2024, doi: 10.58325/ijisct.003.02.0096.