



# Explainable Artificial Intelligence in CyberSecurity

Shravani Shivaji Shitole

Department of Computer Engineering

Nutan Maharashtra Institute of Engineering & Technology, Pune,  
India'

[shravaneeshitole26@gmail.com](mailto:shravaneeshitole26@gmail.com)

Mr. Pritam Ramesh Ahire

Department of Computer Engineering,

Nutan Maharashtra Institute of Engineering & Technology, Pune,  
India.

[pritamahire33@gmail.com](mailto:pritamahire33@gmail.com)

**Abstract**— As artificial intelligence (AI) increasingly integrates into cybersecurity systems, the need for transparency and trust in AI-driven decisions becomes critical. Explainable AI (XAI) seeks to address this by making the inner workings of AI models more interpretable and their outcomes more understandable to human users, including security analysts, IT administrators, and decision-makers. In cybersecurity, where the stakes involve detecting and mitigating threats in real time, opaque “black-box” AI systems may hinder timely and informed responses. XAI offers the potential to clarify how models identify malware, detect anomalies, or flag vulnerabilities, ensuring that AI tools complement human expertise rather than replace it. By providing insights into the decision-making process, XAI enhances trust, enables better compliance with security regulations, and improves accountability, making AI a more reliable partner in fortifying digital defenses. This paper explores the applications of XAI in cybersecurity, its advantages in fostering human-AI collaboration, and the challenges associated with balancing explainability, performance, and complexity in high-stakes environments

**Keywords**— *Explainable AI, Cybersecurity, XAI, Trust in AI, AI Transparency, Anomaly Detection, AI Interpretability, AI-Driven Threat Detection, Human-AI Collaboration, Black-box Models, AI Accountability, Cyber Threat Mitigation.*

## I. INTRODUCTION

As artificial intelligence (AI) becomes more integral to cybersecurity, it is significantly enhancing the ability to detect and respond to a wide range of cyber threats. AI systems are adept at analyzing large volumes of data, identifying patterns, and predicting potential vulnerabilities, allowing for rapid and often automated responses to emerging attacks. These capabilities are invaluable in a landscape where cyber threats are becoming increasingly sophisticated and frequent. However, despite AI's transformative potential, there remains a critical limitation: the “black-box” nature of many AI models.

AI models, particularly those based on deep learning or complex algorithms, often provide highly accurate results, but with little to no explanation of how they arrive at their conclusions. This lack of transparency can be a major concern in cybersecurity, where understanding the rationale behind a decision—whether it be identifying malware, flagging phishing attempts, or detecting abnormal network activity—is essential for professionals to trust, verify, and act on AI-driven insights. Without this understanding, security teams may struggle to interpret the severity or nature of a threat, leading to slower response times or incorrect decisions, and ultimately undermining the effectiveness of AI solutions.

Explainable AI (XAI) addresses this challenge by providing mechanisms that make AI models more transparent and interpretable. XAI techniques enable AI systems to offer human-understandable explanations for the decisions they make, shedding light on how specific conclusions are reached. In the context of cybersecurity, this capability is crucial for several reasons.

First, XAI enhances trust in AI-driven security systems. When security teams understand how AI models reach their decisions, they can more confidently rely on those systems to detect and respond to threats. For instance, if an AI model flags a network intrusion, XAI can help explain which data points or patterns were involved in that determination, allowing security professionals to verify the legitimacy of the alert and respond accordingly. This fosters greater trust in AI-driven processes and reduces the likelihood of overlooking or misinterpreting critical alerts.

Second, XAI is essential for regulatory compliance. Many industries, particularly finance, healthcare, and critical infrastructure, are subject to stringent cybersecurity regulations that require transparency in decision-making processes. For example, the General Data Protection Regulation (GDPR) in Europe mandates that individuals have the right to understand decisions made by automated systems, including those that involve AI. In such contexts, XAI plays a pivotal role by ensuring that AI models meet these transparency requirements, allowing organizations to comply with legal standards and avoid penalties.

Moreover, XAI improves the overall effectiveness of AI in cybersecurity by facilitating better decision-making. Cybersecurity professionals can use the insights provided by XAI to refine their responses to detected threats. For example, XAI can help explain why an AI model prioritizes certain incidents over others, allowing teams to allocate resources more efficiently and

respond to the most critical issues first. Additionally, by providing visibility into the inner workings of AI models, XAI makes it easier to identify and correct potential biases or inaccuracies, leading to more accurate and reliable threat detection over time.

XAI also plays a key role in improving AI models. By explaining how AI models arrive at their conclusions, XAI allows security teams to provide feedback that can be used to refine the models. For instance, if an AI system consistently flags benign activities as threats, XAI can help identify the features or patterns that are causing the false positives, enabling teams to fine-tune the model and reduce errors. This continuous feedback loop not only improves the accuracy of AI models but also enhances their interpretability and fairness.

In addition to these operational benefits, XAI serves as a powerful educational tool for cybersecurity professionals. As AI becomes more embedded in security workflows, it's crucial for teams to understand how these systems function. XAI helps bridge the knowledge gap by making AI systems more accessible and understandable, allowing security teams to learn from the model's decisions and gain deeper insights into the types of threats they are facing. This increased understanding leads to more informed decision-making and better collaboration between human analysts and AI systems.

Lastly, XAI promotes greater accountability in AI-driven cybersecurity solutions. In traditional AI models, the lack of transparency makes it difficult to audit and evaluate how decisions are made, which can be problematic in high-stakes environments like cybersecurity. With XAI, organizations can audit AI models to ensure that they are making fair and accurate decisions. This not only improves the accountability of AI systems but also helps organizations identify vulnerabilities in their AI models that could be exploited by attackers, enhancing overall security.

#### A. Background

The background of Explainable AI (XAI) in cybersecurity stems from the increasing reliance on AI to detect and manage sophisticated cyber threats. As AI models, especially deep learning, became widely used, their "black-box" nature made it difficult to understand how decisions were made, leading to trust issues. Regulatory and compliance requirements, such as GDPR, also demand transparency in AI systems. Additionally, adversarial attacks targeting AI models, the need for human-AI collaboration, and the high rate of false positives in cybersecurity tools highlighted the need for XAI to provide clear, interpretable explanations for AI-driven decisions. This ensures AI systems are both effective and trustworthy.

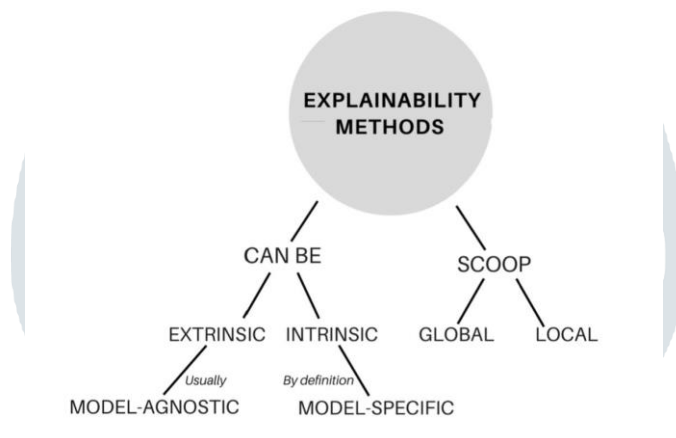


Fig 1: A visual representation of xai taxonomy

## II. ARCHITECTURE OF EXPLAINABLE AI IN CYBERSECURITY

The architecture for Explainable AI (XAI) in cybersecurity integrates traditional AI models with techniques designed to make decisions transparent and understandable. This system is structured to facilitate real-time threat detection and explanation. Below is a detailed description of the key components in the architecture:

### 1. Data Collection Layer

**Data Sources:** This layer is responsible for gathering data from various cybersecurity-related sources, including:

**Network Traffic:** Monitoring packets for potential anomalies.

**System Logs:** Collecting logs from servers, firewalls, and endpoint devices.

**User Behavior Analytics (UBA):** Tracking user interactions to detect abnormal activities.

**Threat Intelligence Feeds:** Integrating known threat information from external sources.

**Preprocessing:** Before feeding into AI models, the data is cleaned, normalized, and preprocessed for better accuracy. This may involve filtering noise, aggregating time-series data, and removing redundant features.

### 2. AI Model Layer

**Core AI Model:** This component contains the machine learning (ML) or deep learning (DL) models trained to detect threats and cyber anomalies. Examples of models used are:

**Random Forests:** Useful for classification tasks, such as identifying malicious activity.

**Support Vector Machines (SVMs):** For anomaly detection by separating normal from abnormal traffic.

**Neural Networks:** Deep learning models used for pattern recognition, such as malware classification or intrusion detection.

These models continuously learn from the incoming data and update themselves to detect new and evolving cyber threats.

### 3. XAI Explanation Layer

The heart of the architecture is the explanation layer, which provides insights into the AI model's decision-making process. The most common XAI techniques include:

- **LIME (Local Interpretable Model-Agnostic Explanations):** Provides local approximations of the model's behavior by explaining individual predictions. For example, LIME can explain why a specific network activity is flagged as malicious.
- **SHAP (Shapley Additive Explanations):** SHAP computes the contribution of each feature to the final prediction. For example, if an AI model detects a potential phishing attack, SHAP will show how different features (email sender, domain, content) contribute to this classification.
- **Feature Importance:** This method ranks input features based on their importance in making predictions. For instance, it highlights which aspects of network traffic were most significant in flagging an anomaly.
- **Saliency Maps:** Used in deep learning, these visualizations show which input features (such as regions of a network) most influence the model's decision. In cybersecurity, saliency maps can indicate the most critical parts of a network that need immediate attention.

### 4. User Interface (UI) Layer

**Dashboard:** The XAI-based cybersecurity system is equipped with a user interface that visualizes predictions and explanations in an intuitive manner.

**Real-time Alerts:** Notifications about potential threats are accompanied by explanations of why they were flagged.

**Visual Explanations:** Graphical representations such as feature importance charts, SHAP plots, or saliency maps provide insights into the reasoning behind decisions.

**Threat Classification:** Alerts are classified based on the severity of threats, and explanations are provided for each detected event.

**User Feedback Mechanism:** Security experts can provide feedback to the AI model based on the explanations provided, which in turn helps improve the system's performance. For example, experts can mark alerts as false positives, and the AI will adjust its future predictions accordingly.

### 5. Feedback Loop

The feedback loop allows for continuous improvement of the AI model:

**Model Refinement:** Based on user feedback, the AI model fine-tunes its predictions. If a false positive or missed threat is identified, the system can learn from these mistakes to avoid them in the future.

**Explanation Refinement:** The system evolves not only in terms of predictive accuracy but also in providing clearer, more accurate explanations to cybersecurity professionals.

### 6. Security and Compliance Layer

**Auditing:** The system logs all decisions and explanations to ensure compliance with regulatory requirements. This is crucial for industries like healthcare or finance, where understanding the basis of a decision is essential for accountability.

**Bias Mitigation:** Regular audits of the AI system help ensure that there are no biases in the decision-making process. This is particularly important in cybersecurity to prevent the system from unfairly targeting specific types of network traffic or user behavior.

End-to-End Process Flow:

1. **Data Ingestion:** Network traffic, system logs, and UBA data are collected and preprocessed.
2. **Prediction:** The core AI model analyzes the data and flags any suspicious activity or anomalies.
3. **Explanation:** The explanation layer (using LIME, SHAP, etc.) generates human-readable explanations for why a particular activity was flagged as a threat.
4. **Action:** Security professionals use the provided explanations to assess the threat and take action (e.g., block traffic, investigate further).
5. **Feedback:** Professionals give feedback on the decisions, helping the system refine both its threat detection and explanation quality.
6. **Audit & Compliance:** The system logs all decisions and explanations for future reference, compliance, and audits.

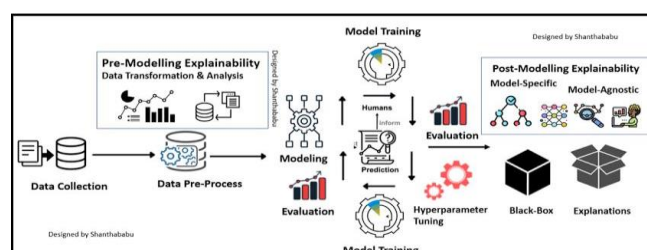


Fig 2 : Architecture of explainable AI in cybersecurity

### III. METHODOLOGY

The methodology for implementing Explainable AI (XAI) in cybersecurity involves a series of steps designed to build, train, explain, and evaluate AI models that can detect and respond to cyber threats in a transparent and interpretable manner. The following describes the overall approach, from data gathering to model deployment.

#### 1. Data Collection and Preprocessing

The first step in any AI-driven cybersecurity system is gathering relevant data. The types of data used can vary but often include:

**Network Traffic Data:** Capturing packets sent over the network to detect anomalies such as abnormal traffic patterns, port scans, or denial-of-service attacks.

**System Logs:** Collecting logs from firewalls, servers, routers, and end devices to track activities that might indicate a security breach.

**User Behavior Analytics (UBA):** Monitoring user actions to detect insider threats or compromised accounts through deviations from normal behavior.

**Threat Intelligence Feeds:** External feeds that provide known attack vectors, malware signatures, or other threat intelligence to enhance detection.

After gathering data, the next step is preprocessing, which involves:

**Data Cleaning:** Removing noise, handling missing values, and dealing with inconsistent entries.

**Feature Engineering:** Identifying key features that are relevant to cybersecurity, such as traffic volume, IP address patterns, login attempts, or malware signatures.

**Normalization:** Scaling and transforming data to ensure uniformity and improve model performance.

#### 2. Model Development

Once the data is prepared, the next step is building and training machine learning (ML) or deep learning (DL) models to detect cybersecurity threats. The model development process includes:

**Model Selection:** Choosing the appropriate algorithms for specific cybersecurity tasks:

**Anomaly Detection:** Algorithms like Support Vector Machines (SVM), Autoencoders, or Isolation Forests are used to detect unusual patterns in network behavior.

**Threat Classification:** Supervised learning models such as Random Forest, Decision Trees, or Neural Networks are trained to classify different types of attacks like malware, phishing, or ransomware.

**Behavioral Analysis:** Reinforcement learning models or Markov chains may be used to predict abnormal user behaviors.

**Training:** The selected models are trained using the preprocessed data, with a focus on achieving high accuracy in threat detection.

**Testing and Validation:** The models are tested using holdout datasets or through cross-validation to evaluate their performance. Metrics such as precision, recall, accuracy, and F1-score are used to gauge the effectiveness of the model.

#### 3. Integration of Explainability

Explainability is introduced by incorporating XAI techniques into the models to provide clear, interpretable explanations of the AI system's decisions. The primary methods used for this include:

**LIME (Local Interpretable Model-Agnostic Explanations):** LIME provides localized explanations for individual predictions by approximating the AI model with a simpler, interpretable model. In cybersecurity, this might mean explaining why a specific network activity is flagged as suspicious by highlighting the most important features (e.g., unusual IP address or packet size).

**SHAP (Shapley Additive Explanations):** SHAP values are used to show the contribution of each feature to a model's decision. In a cyberattack classification scenario, SHAP can explain how specific features (e.g., time of login, number of failed attempts) contributed to the decision to flag an activity as an attack.

**Saliency Maps:** For deep learning models, especially in visual and temporal data analysis, saliency maps are used to highlight which parts of the input (e.g., sections of network traffic) contributed most to the prediction.

**Decision Trees and Rule-based Models:** In cases where simpler models like decision trees can be employed, these inherently interpretable models provide clear pathways that can explain why a certain outcome was predicted.

#### 4. Model Evaluation and Explainability Testing

The explainability of the model is evaluated along with its performance:

**Explainability Metrics:** These include how easily humans can understand the model's explanations, the completeness and clarity of the explanation, and the trustworthiness of the information provided.

**Performance Metrics:** Metrics such as accuracy, precision, recall, and F1-score are calculated to ensure the model's effectiveness in detecting threats. The balance between performance and explainability is crucial. For example, a very complex neural network may be accurate but difficult to interpret, whereas simpler models might offer better transparency but lower detection accuracy.

Human-in-the-Loop Testing: Cybersecurity experts interact with the system, reviewing predictions and the accompanying explanations. Their feedback helps fine-tune the model for both accuracy and clarity of explanations.

## 5. Real-Time Deployment

Once the model has been tested and validated, it is deployed in real-time environments:

Integration with Security Operations: The model is connected to live data streams (e.g., network traffic or user activity) to monitor for real-time threats.

Continuous Learning and Feedback: As new data is collected, the model continues to learn and adapt, refining its predictions and explanations based on updated threat patterns.

User Interface and Explanation Delivery: An easy-to-understand interface is crucial for communicating the AI's decisions to cybersecurity analysts. The UI typically includes:

Alerts: When a potential threat is detected, the system generates alerts along with explanations of why the activity was flagged.

Visualizations: Techniques like SHAP plots or saliency maps are displayed to help analysts understand which features contributed most to the decision.

## 6. Continuous Improvement and Feedback Loop

Once the XAI system is live, continuous monitoring, feedback, and updates are essential for improvement:

Human-AI Collaboration: Security experts provide feedback on the system's predictions and explanations. If an alert is marked as a false positive or false negative, the AI system uses this feedback to improve its future performance.

Model Retraining: Periodically, the AI model is retrained using the latest data and feedback to stay up-to-date with evolving threats.

Explanation Refinement: The explanations generated by XAI techniques are also refined based on user feedback, ensuring that they remain relevant and easy to understand.

## 7. Security and Compliance Audits

Compliance Monitoring: Many industries (such as finance and healthcare) require adherence to cybersecurity regulations. XAI models facilitate this by providing clear, auditable explanations of their decisions.

Bias Detection and Mitigation: Regular audits are conducted to ensure that the AI models are not introducing biases, such as favoring certain network traffic or discriminating against certain types of users. Explanations help in identifying any areas where the model may be behaving unfairly or inconsistently.

# IV. DATA COLLECTION AND MANAGEMENT

Data Collection and Management in Explainable AI for Cybersecurity involves gathering diverse data sources like network traffic, system logs, user behavior, and threat intelligence. The data is then cleaned, transformed, and stored securely using scalable solutions like SIEM systems or cloud storage. Preprocessing includes normalizing, labeling for supervised learning, and handling missing or noisy data. Real-time processing is enabled for immediate threat detection. Ethical considerations such as anonymization and compliance with privacy laws are crucial. Continuous feedback loops improve the model's accuracy and transparency over time, allowing security experts to understand and refine AI predictions.

# V. PREDICTIVE ANALYTICS AND DECISION SUPPORT

In Explainable AI (XAI) for cybersecurity, predictive analytics is used to identify patterns and trends in vast amounts of security data, such as network logs, user activity, and system events. It helps forecast potential cyber threats, detect anomalies, and assess risks by learning from historical data and real-time inputs. Traditional AI models often act as "black boxes," offering predictions without clear reasoning, which can be problematic in critical fields like cybersecurity. XAI addresses this by making the decision-making process transparent, enabling security professionals to understand why certain predictions or threat alerts are made.

For decision support, XAI enhances the ability of cybersecurity teams to make informed responses to threats. It provides clear explanations of AI-driven predictions, offering insights into why a certain action, like flagging suspicious activity or prioritizing a security alert, is recommended. This interpretability is crucial for ensuring that security decisions are trusted and aligned with the organization's policies. Additionally, XAI allows cybersecurity teams to collaborate with AI systems, refining decisions based on clear, understandable rationales, leading to better incident management and response strategies. Through this, XAI enhances both predictive accuracy and the practical usability of AI in real-world cybersecurity scenarios.

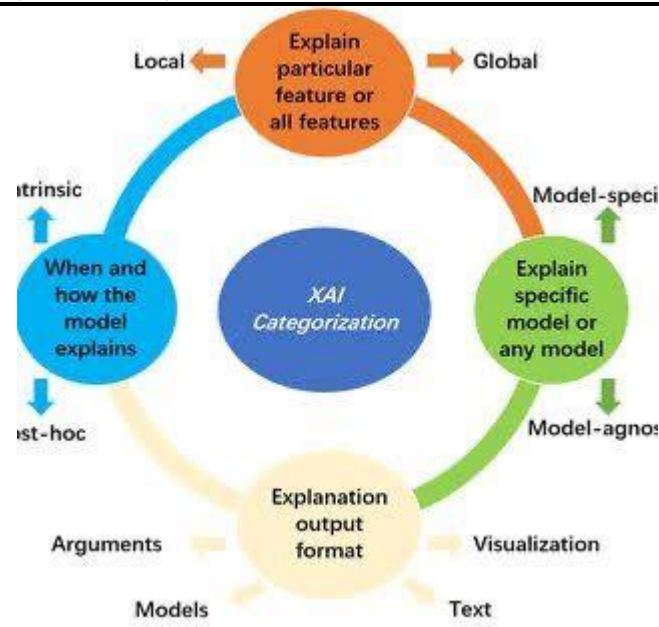


Fig 3 : Applications

## VI. Challenges and Limitations

Explainable AI (XAI) in cybersecurity encounters several challenges and limitations. One major issue is the complexity of AI models, particularly deep learning algorithms, which can be difficult to interpret. This complexity can lead to oversimplifications in the explanations provided, potentially misleading users. Additionally, the lack of standardized frameworks for XAI results in inconsistent practices across organizations, making it challenging to adopt uniform approaches.

There is often a trade-off between model accuracy and interpretability; more interpretable models may not achieve the same predictive performance as complex models. The dynamic nature of cyber threats further complicates the situation, as XAI systems may struggle to keep pace with rapidly changing attack patterns. Data quality is also a concern, as effective XAI relies on high-quality data, which can be difficult to obtain due to issues like bias or incompleteness.

Furthermore, while XAI aims to improve user understanding, the explanations provided may still be too complex for non-expert users, hindering trust in the system. Ethical concerns around privacy also arise, as explanations must not expose sensitive information. Integrating XAI with existing cybersecurity frameworks can pose significant challenges, particularly for organizations with legacy systems that may not be compatible.

Scalability is another issue; as organizations expand and their networks grow more complex, maintaining performance while providing real-time explainability can be technically demanding. Finally, the reliance on human expertise remains critical; while XAI can support decision-making, skilled cybersecurity professionals are essential for interpreting and applying the insights provided. Together, these challenges underscore the need for ongoing research and innovation to enhance the effectiveness of XAI in the cybersecurity domain.

## VII. Future Research Directions

Future research directions for Explainable AI (XAI) in cybersecurity can focus on several key areas. First, there is a need for advanced XAI techniques that enhance interpretability while maintaining model accuracy, including user-centric explanation models that adapt to different levels of technical expertise. Integrating XAI with emerging technologies like blockchain and zero-trust architectures could improve security and transparency in AI-driven decisions.

Addressing ethical considerations and mitigating bias in AI models is crucial, with research aimed at creating frameworks for fair and accountable decision-making. Enhancing real-time explanation capabilities is essential as the landscape of cyber threats evolves, allowing for immediate, context-aware insights.

Additionally, facilitating better collaboration between AI systems and human operators is an important area, focusing on intuitive interfaces that present AI predictions and explanations clearly. Finally, ensuring the scalability of XAI solutions in large, complex environments will be vital for effective deployment in enterprises. Together, these research directions can significantly advance the effectiveness and usability of XAI in cybersecurity.

## VIII. Conclusion

Explainable AI (XAI) is vital for enhancing cybersecurity by providing transparency and interpretability in AI decision-making. As cyber threats grow more sophisticated, the ability to understand AI predictions becomes essential for fostering trust among cybersecurity professionals. Integrating XAI into cybersecurity frameworks can improve threat detection, incident response, and regulatory compliance.

Despite challenges such as model complexity and ethical considerations, future research should focus on developing advanced explainability techniques, user-friendly interfaces, and scalable solutions. Ultimately, the successful adoption of XAI in

cybersecurity will empower security teams to make informed, data-driven decisions, strengthening defenses against evolving threats and protecting sensitive information and critical infrastructure.

## IX. References

1. Ahire, Pritam Ramesh, and K. Ulaga Priya. "Monitoring Body Mass Index (BMI) Pre & Post Covid-19 Outbreak: A Comprehensive study in Healthcare." *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)*. IEEE, 2024.
2. Ahire, Pritam. "Predictive and Descriptive Analysis for Healthcare Data, A Hand book on Intelligent Health Care Analytics-Knowledge Engineering with Big Data" <https://www.wiley.com/enus/Handbook+on+Intelligent+Healthcare+Analytics%3A+Knowledge+Engineering+with+Big+Data-p-9781119792536> Published by Scrivener Publishing." (2021).
3. Ahire, Pritam, et al. "LSTM based stock price prediction." *International Journal of Creative Research Thoughts* 9.2 (2021): 5118-5122.
4. Ahire, Pritam R., and Preeti Mulay. "Discover compatibility: Machine learning way." *Journal of Theoretical & Applied Information Technology* 86.3 (2016).
5. Ahire, Pritam R., Rohini Hanchate, and Vijayakumar Varadarajan. "Indigenous Knowledge in Smart Agriculture." *Advanced Technologies for Smart Agriculture*. River Publishers, 2024. 241-258.
6. Hanchate, R., & Anandan, R. (2023). Medical Image Encryption Using Hybrid Adaptive Elliptic Curve Cryptography and Logistic Map-based DNA Sequence in IoT Environment. *IETE Journal of Research*, 1–16. <https://doi.org/10.1080/03772063.2023.2268578>
7. Ahire, Pritam Ramesh, Rohini Hanchate, and K. Kalaiselvi. "Optimized Data Retrieval and Data Storage for Healthcare Applications." *Predictive Data Modelling for Biomedical Data and Imaging*. River Publishers 107-126.
8. M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword," *Nature Mach. Intell.*, 1638 vol. 1, no. 12, pp. 557–560, Dec. 2019. 1639
9. D. Gunning and D. Aha, "Darpa's explainable artificial intelligence 1640 (XAI) program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019. 1641
10. P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and 1642 M. A. Przybocki, "Four principles of explainable artificial intelligence," 1643 NIST Interagency, Gaithersburg, MD, USA, Internal Rep. NISTIR-8312, 1644 Aug. 2020, doi: 10.6028/NIST.IR.8312. 1645
11. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust 1646 you?" Explaining the predictions of any classifier," in *Proc. 22nd 1647 ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, 1648 pp. 1135–1144. 1649
12. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model 1650 predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, 1651 pp. 1–10. 1652
13. M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision 1653 model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 1654 no. 1, Apr. 2018, pp. 1–9
14. R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and 1655 F. Giannotti, "Local rule-based explanations of black box decision sys- 1656 tems," 2018, arXiv:1805.10820. 1657

15. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and 1658 D. Batra, “Grad-CAM: Visual explanations from deep networks via 1659 gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis. 1660 (ICCV), Oct. 2017, pp. 618–626. 1661
16. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, 1662 and P. Das, “Explanations based on the missing: Towards contrastive 1663 explanations with pertinent negatives,” in Proc. Adv. Neural Inf. Process. 1664 Syst., vol. 31, 2018, pp. 1–12. 1665
17. S. Morgan. (2020). Special report: Cyberwarfare in the C-suite, 1666 online. Cybercrime Magazine. [Online]. Available: <https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/> 1668
18. Enisa Threat Landscape 2020—List of Top 15 Threats, ENISA, Athens, 1669 Greece, 2020. 1670 [12] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: 1671 A review of machine learning interpretability methods,” Entropy, vol. 23, 1672 no. 1, p. 18, Dec. 2020. 1673
- a. Rawal, J. McCoy, D. B. Rawat, B. Sadler, and R. Amant, “Recent 1674 advances in trustworthy explainable artificial intelligence: Status, chal- 1675 lenges and perspectives,” IEEE Trans. Artif. Intell., no. 4, Aug. 2021, doi: 1676 10.1109/TAI.2021.3133846. 1677
- b. Rai, “Explainable AI: From black box to glass box,” J. Acad. Market- 1678 ing Sci., vol. 48, no. 1, pp. 137–141, Jan. 2020. 1679
19. Kale, T. Nguyen, F. C. Harris, Jr., C. Li, J. Zhang, and X. Ma, 1680 “Provenance documentation to enable explainable and trustworthy 1681 AI: A literature review,” Data Intell., pp. 1–41, Feb. 2022, doi: 1682 10.1162/dint\_a\_00119. 1683
20. Adadi and M. Berrada, “Peeking inside the black-box: A sur- 1684 vey on explainable artificial intelligence (XAI),” IEEE Access, vol. 6, 1685 pp. 52138–52160, 2018. 1686
21. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, 1687 “Trends and trajectories for explainable, accountable and intelligible 1688 systems: An HCI research agenda,” in Proc. CHI Conf. Hum. Factors 1689 Comput. Syst., 2018, pp. 1–18. 1690
22. Q.-S. Zhang and S.-C. Zhu, “Visual interpretability for deep learn- 1691 ing: A survey,” Frontiers Inf. Technol. Electron. Eng., vol. 19, no. 1, 1692 pp. 27–39, 2018. 1693
23. Q. Zhang, Y. N. Wu, and S.-C. Zhu, “Interpretable convolutional neural 1694 networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 1695 Jun. 2018, pp. 8827–8836. 1696
24. P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, 1697 “Explainable artificial intelligence: An analytical review,” Wiley Inter- 1698 discipl. Rev., Data Mining Knowl. Discovery, vol. 11, no. 5, p. e1424, 1699 2021. 1700
- a. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, 1701 A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, 1702 and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, 1703 taxonomies, opportunities and challenges toward responsible AI,” Inf. 1704 Fusion, vol. 58, pp. 82–115, Jun. 2020. 1705
25. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, 1706 “Explaining explanations: An overview of interpretability of machine 1707 learning,” in Proc. IEEE 5th Int. Conf. Data Sci. Adv. Analytics (DSAA), 1708 Oct. 2018, pp. 80–89. 1709
26. G. Riccardo, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and 1710 D. Pedreschi, “A survey of methods for explaining black box models,” 1711 ACM Comput. Surv., vol. 51, no. 5, pp. 1–42, 2018. 1712

27. M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, “A systematic review 1713 of explainable artificial intelligence in terms of different application 1714 domains and tasks,” Appl. Sci., vol. 12, no. 3, p. 1353, Jan. 2022. 1715
28. E. Tjoa and C. Guan, “A survey on explainable artificial intelligence 1716 (XAI): Toward medical XAI,” IEEE Trans. Neural Netw. Learn. Syst., 1717 vol. 32, no. 11, pp. 4793–4813, Oct. 2021. 1718
29. Mittelstadt, C. Russell, and S. Wachter, “Explaining explanations in 1719 AI,” in Proc. Conf. Fairness, Accountability, Transparency, Jan. 2019, 1720 pp. 279–288. 1721
30. T. Miller, “Explanation in artificial intelligence: Insights from the social 1722 sciences,” Artif. Intell., vol. 267, pp. 1–38, Feb. 2018. 1723
31. S. Mohseni, N. Zarei, and E. D. Ragan, “A multidisciplinary survey and 1724 framework for design and evaluation of explainable AI systems,” ACM 1725 Trans. Interact. Intell. Syst., vol. 11, nos. 3–4, pp. 1–45, Dec. 2021.

