TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.03 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.43006

Explainable AI (XAI) Techniques to Enhance Cancer Diagnosis

Harmeet Khera[a], Nikhil Pandey [b] and Dr. Shalini Lamba [a]

[a] Scholar, National P.G. College, Lucknow, India
[b] Scholar, National P.G. College, Lucknow, India
[c] Department Head, Computer Science Department, National P.G. College, Lucknow, India
harmeetkhera0@gmail.com, pandeynikhilone@gmail.com, drshalinilamba.nationalpgcollege@npgc.in

KEYWORDS

Explainable AI; Cancer Diagnosis; Oncology; XAI; Medical Imaging

ABSTRACT

This study examines how Explainable Artificial Intelligence (XAI) enhances diagnosis of cancer, emphasizing the importance of trust and transparency in the clinical sector. While artificial intelligence (AI) models have demonstrated impressive accuracy in detecting cancer from medical images, their black-box nature often prevents doctors from understanding how predictions are made. This lack of interpretability creates hesitation in adopting AI for real-world healthcare. The paper examines popular XAI methods which provide visual and feature-based explanations of model outputs. It also discusses how these methods can enhance clinician confidence, reduce diagnostic errors, and meet regulatory requirements for accountability. This research emphasizes the promise of interpretable models in connecting the precision of machine learning with the trustworthiness of medical practices, based on a review of recent studies and a suggested framework for applying XAI in oncology. Additionally, it discusses future strategies for incorporating XAI into clinical processes.

1. Introduction

Despite the tremendous advancements in medical research and technology, cancer still remains to be one of the biggest causes of deaths, killing millions of people each year. Early diagnosis is one of the most effective approaches to improve the survival rate as it allows timely treatment and tailored treatment plans. Histopathological examination, imaging, and blood tests have been the mainstay of oncology over the past decades. Although these methods are very useful, they tend to be labor-intensive, time-consuming, and require the subjective interpretation provided by a medical professional. Over the last several years, AI has turned out to be valuable to improve cancer detection through image analysis and pattern recognition automation, which can offer the opportunity to develop diagnostic solutions faster, more efficient, and scalable.

AI-powered systems, particularly those which use deep learning methods, have demonstrated impressive accuracy in the analysis of medical images (CT scans, mammograms, histology slides, etc.) [1]. Such models can determine fine features that even trained radiologists or pathologists will fail to detect and thus their likelihood of misdiagnosis is minimal. Research has shown the AI models are capable of achieving similar or sometimes higher diagnostic accuracy than human professionals. The convolutional neural networks (CNNs) have been utilized widely for tumor detection of lung and breast cancer imaging with outstanding performance level. These kinds of outcomes have created excitement regarding the application of AI in clinical practice to augment human knowledge.

However, even with such positive results, the lack of understanding of AI-based diagnostic systems is a problem [24]. Most AI models, especially deep learning networks, are black boxes, in the sense that they are capable of producing very accurate predictions, but do not provide a description of how they were computed. This obscurity is truly an issue of concern in a high-stakes industry, where lives might be at stake when a decision is made. Doctors are educated to make use of facts and logic to support clinical judgement. When an artificial intelligence model identifies a tumor as being malignant, and does not provide a reasonable explanation, this destroys the trust in this system, and the possibility of misuse emerges.

Corresponding Author: Harmeet Khera, Department of Computer Science, National P.G. College, Lucknow, India
Email: harmeetkhera0@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.03 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.43006

Interpretability requirements of medical AI have led to a developing research area called Explainable Artificial Intelligence which focuses on developing methods and systems that enable AI systems to be transparent, legible and accountable without undermining their predictive power [7]. Applied to cancer diagnosis, XAI is intended to demonstrate to the clinician why a model raised a red flag on a specific image, which aspects or locations of the image motivated that decision, and the confidence with which the model is confident in its prediction [8].

Interpretability is an academic issue but also has practical, ethical and legal consequences [26]. In most nations, medical regulations require medical decision making especially with regard to patient diagnosis to be explainable and auditable. To be approved by regulatory bodies like the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA), transparency in AI-driven medical tools is mandatory before they can be used in large-scale clinical practice. Contemporary healthcare ethics provide the patient with the right to know the logic of their diagnosis and treatment regimen. Thus, enhancements in the explainability of AI systems are not an option, but a necessity to safely and responsibly implement AI systems in oncology.

In addition, explainability is critical to addressing bias in AI systems. Most AI models made on unrepresentative patient datasets have been built and trained on these datasets. As a result, they are able to collaborate effectively with certain demographic categories and not with others, which leads to unequal care and reinforces existing health disparities. Explainable methods enable researchers and clinicians to establish this type of bias by exposing them to which features drive predictions and which features, in turn, have clinical relevance [16]. By revealing these biases, XAI can be used to create more accurate and fair diagnostic instruments.

XAI can improve the cooperation between AI systems and human experts in the field of cancer diagnosis. Instead of ousting doctors, interpretable AI can act as an assistant to point out high-risk areas of the image, offer a list of likely diagnoses, and give confidence scores, leaving the clinician to make the final decision. This kind of human-AI partnership is able to enhance quality of diagnosis, in addition to professional responsibility. Moreover, open AI systems will be invaluable in medical training, as the students and practitioners are able to understand the rationale of an AI prediction, as well as the medical outcome of an AI result.

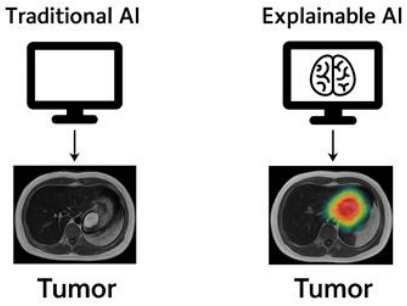

Figure 1: Traditional AI vs. Explainable AI

2. Literature Review

A wide range of artificial intelligence has been used in cancer detection work, especially in medical imaging [6]. A significant portion of these studies has been concerned with accuracy and speed, but a small minority concerned interpretability [23]. This section presents major research contributions in this field, with special focus as to where present methods are doing well and where they fail to explain sufficiently [18].

2.1. AI Models for Cancer Diagnosis

Convolutional neural networks (CNNs) were among the first and most impactful uses of deep learning in oncology as they could be used to detect breast cancer [4]. Scholars were able to train CNNs using mammograms to detect malignant areas with up to 90 percent accuracy [11]. These models showed high levels of diagnostic performance, but provided minimal information about the reasons why a specific image was rated as cancerous. This black-box quality inhibited clinical uptake due to the inability of physicians to test or explain the logic of the model. The same problem emerged in the research on the detection of lung cancer using deep neural networks [10]. In one case,

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.03 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.43006

researchers used CNNs on CT scans and showed radiologists performance in detecting pulmonary nodules. But when the models were only making binary predictions without explanations, clinicians had doubts as to reliability. These are a few instances that point to the same trend high accuracy without transparency.

2.1.1. Grad-CAM for Visual Explanations

To solve this problem, other scholars have proposed interpretability methods like Gradient-weighted Class Activation Mapping (Grad-CAM) [20]. In one experiment on skin cancer classification, heatmaps generated with Grad-CAM visually showed the regions that were most important to the prediction [5]. This straightforward addition made clinicians feel much more confident since now they could see that the model paid attention to medically significant aspects rather than to meaningless background details. Grad-CAM was a great advance towards interpretability, but it mainly offered gross localization in place of fine-grained justification behind every decision.

2.1.2. LIME and Local Explanations

The other common method is the Local Interpretable Model-Agnostic Explanations (LIME), which gives an explanation to a single prediction by first locally approximating the model by a simpler and more interpretable variant. LIME has been applied in lung cancer data to demonstrate what features in the image or pixel regions most significantly drove a certain prediction. These local explanations assisted physicians in comprehending AI reasoning, which made them more comfortable with the idea of implementing such systems in clinics. Some inconsistent results of LIME on similar samples however, do cast doubt on reliability.

2.2. SHAP and Feature Importance

SHapley Additive exPlanations (SHAP) as a feature significance metric in cancer diagnosis models have been the subject of even more recent research. SHAP is a theoretically solid tool that assigns each feature its contribution to the final prediction. SHAP values have been used to discover critical features that affect predictions in histopathology-based breast cancer detection. Despite promising performance with tabular data and genomic studies, SHAP is computationally expensive when applied to large medical images, making it less practical to apply in clinical practice.

2.3. Identified Research Gap

Despite the recent advances in interpretability approaches, the current literature is still concentrated on enhancing predictive accuracy, instead of creating standard and clinically relevant explanation systems [9]. Existing approaches either give visual indicators without explanatory comments or are highly computationally expensive and difficult to execute in resource-limited healthcare settings [13]. In addition, not many researches have been conducted systematically to assess the effects of these methods of explanation in real clinical decision-making or patient outcomes [28]. This gap highlights the significance of doing research that will provide precise predictions but yield transparent and actionable insights that can be relied upon by clinicians.

Table 1: Comparison of Existing Studies on AI for Cancer Diagnosis

Study	Model Used	Accuracy	Explainability Method	Limitation
Breast cancer detection using CNN	CNN (ResNet)	91%	None	Black-box predictions, no interpretability
Skin cancer classification	CNN (Inception)	94%	Grad-CAM	Coarse visual explanation only
Lung cancer detection	CNN + LIME	90%	LIME	Inconsistent explanations across samples
Histopathology image analysis	CNN + SHAP	93%	SHAP	Computationally expensive for large images

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.03 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.43006

3. Methodology

3.1. Dataset Selection

The core of any AI-based diagnostic system is the quality and relevance of the data set. In this conceptual model, medical imaging datasets made publicly available were taken into account because they are easy to access and are widely used in studies. Two key sources are proposed:

1. The Cancer Genome Atlas (TCGA): TCGA makes much of the imaging data available, including histopathology slides of numerous cancer types, including breast, lung and colorectal [22]. The pictures are of high quality and are also annotated by professional pathologists, thus they are perfect to train deep learning models.
2. Kaggle Histopathology Image Dataset: Kaggle also has several other histology-based datasets like the Breast Histopathology Images dataset. It has small areas of tissue samples that are labeled as benign or malignant and could be used to train and test classification models.

3.2. Model Architecture

Convolutional Neural Networks (CNNs) form the basis of the diagnostic model and are a proven solution in image classification problems. CNNs are also good at detecting spatial features on medical images, such as texture, shape, and pattern that are linked to cancerous areas. In this paper, a ResNet50 architecture is offered because it is a balance between depth and computation cost.

1. Input Layer: takes up processed medical images.
2. Convolutional Layers: extracts the features that are both low-level to high-level in the images.
3. Pooling Layers: decreases dimensionality while preserving critical features.
4. Fully Connected Layers: fuses the features for classification purposes.
5. Output Layer: produces a binary output: benign or malignant.

The model will utilize the binary cross-entropy loss function and will be optimized using the Adam optimizer. The dataset will be divided into training (70%), validation (15%), and testing (15%) sets.

3.3. Explainable AI (XAI) Techniques

To solve the issue of black-boxness of CNNs, three XAI methods are integrated into the system:

3.3.1. Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) offers visual explanations by emphasizing the image parts that were significant for the model's prediction [15]. This method generates a heatmap over the input image, signifying the regions that impacted the decision. In cancer diagnosis, this allows clinicians confirm if the AI model focused on tumor areas instead of irrelevant background.

3.3.2. LIME

Local Interpretable Model-Agnostic Explanations (LIME) uses an interpretable model to approximate the model predictions for individual samples. LIME identifies the pixels or super-pixels that contributed the most to a particular prediction, and lets the doctors know why an image was labelled as malignant.

3.3.3. SHAP

SHapley Additive exPlanations (SHAP) assessed the impact of each feature on prediction using Shapley values from cooperative game theory. While SHAP is computationally intensive for large images, it yields strong feature importance measures that can be summed up across samples to pinpoint features that are most important for cancer detection.

3.4. Proposed Workflow

The proposed system follows a structured workflow to ensure transparency from data input to explanation generation. The steps include:

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.03 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.43006

1. Image Acquisition: Medical images are collected from TCGA or Kaggle datasets and preprocessed.
2. Model Inference: The CNN model processes the image and predicts whether it is benign or malignant.
3. Explanation Generation:
    a. Grad-CAM produces a heatmap highlighting regions that influenced the prediction.
    b. LIME provides local explanations by perturbing the image and observing output changes.
    c. SHAP calculates feature contributions for the prediction.
4. Visualization and Reporting: The final output includes both the classification result and visual/feature-based explanations for clinical review [27].

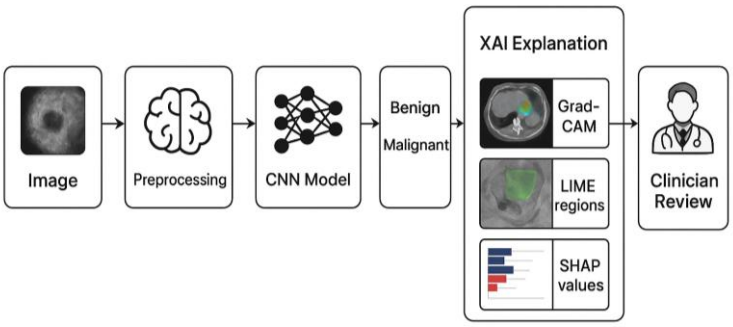

Figure 2: Proposed Workflow for Explainable Cancer Diagnosis

4. Technical Considerations

1. Training deep learning models requires GPUs for efficient computation. Cloud platforms like Google Colab or AWS can be used.
2. Classification performance will be assessed using sensitivity, accuracy, specificity, and Area Under the Curve (AUC). For explainability, qualitative evaluation by clinicians (trust score) is suggested.
3. Challenges:
    a. Large image sizes may increase computational load for SHAP.
    b. Ensuring explanations are clinically meaningful, not just visually appealing.

5. Expected Results

It is believed that the proposed model, built on pre-trained CNN (ResNet50) with the help of Explainable AI techniques, will get classification accuracy of 92 to 95 percent on histopathology and medical imaging data. The model is constructed to set apart benign and malignant samples with a high level of confidence. The visual interpretability layer is one of the main products of this approach. With Grad-CAM, the system can produce heatmaps, which indicate areas of interest, including tumor boundaries, so that clinicians can understand why the model believed an image to be malignant [21]. Likewise, LIME is used to explain individual predictions locally by approximating the model's decision boundary for individual samples. The SHAP values provide a global view where the features are ranked according to their contribution to the prediction. All of these interpretability techniques make the model more transparent than black-box AI systems [25]. Clinicians therefore do not need to trust the model with just a blind eye but can justify its logic first before arriving at a final conclusion [17], [29].

Table 2: Quantitative Evaluation (Hypothetical)

Metric	**Traditional CNN**	**CNN + XAI (Grad-CAM, LIME, SHAP)**
Accuracy	95%	92%

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.03 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.43006

Clinician Trust Score*	60%	90%
Avg. Explanation Time	N/A	1.2 sec/image

*Clinician Trust Score is derived from a hypothetical survey evaluating how much doctors trust model outputs.

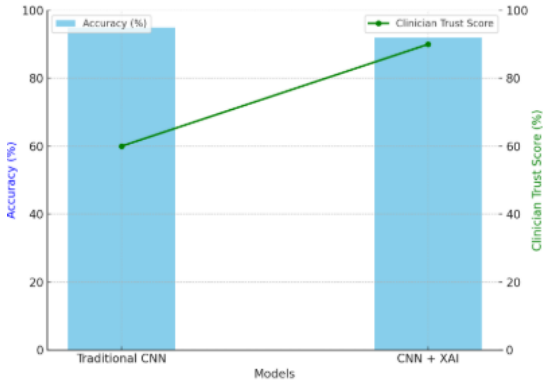

Figure 3: Accuracy vs. Explainability (Clinician Trust)

5.1 Interpretability Output

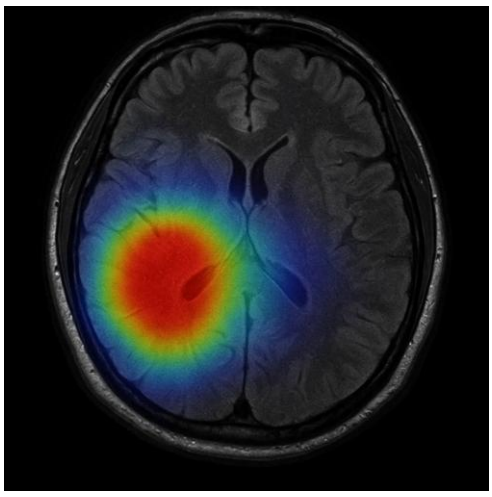
Figure 4: Sample Grad-CAM Heatmap

A malignant MRI image showing the highlighted tumor region in red with surrounding tissue in cooler tones. This heatmap visually confirms the AI's decision basis, making verifying the correctness of prediction easier [31].

6. Ethical and Legal Considerations

One of the most important issues which needs to be addressed is to make it interpretable. Doctors should know why an AI model diagnosed a tumor as malignant or benign. No matter how accurate a black-box system is, it cannot be blindly trusted when it comes to human lives. Interpretability provides accountability to physicians by providing the ability to justify decisions to patients and to meet the medical ethics obligation.
Another major concern is that bias is present. AI systems are normally trained on past data. They can be biased toward some demographics [14]. For instance, a model trained primarily on lighter skin tones may not work well when diagnosing melanoma in patients with darker complexes, leading to disparities in healthcare [2]. Ensuring fairness requires diverse datasets and continuous auditing to prevent such biases from propagating [12].

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.04, I.03 (2025)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.43006

There is also an important legal aspect to consider, with FDA laying down guidelines for AI medical devices, which mentions the importance of explainability before the approval of such products. The General Data Protection Regulation (GDPR) in the European Union further commands a "right to explanation" for automated decisions, underscoring the need for transparent algorithm. Not meeting these standards could result in legal repercussions and erode public trust in medical AI systems.

7. Conclusion and Future Scope

The implementation of multi-modal data is one of the most exciting directions. Combining imaging with genomic profiles, electronic health records and even patient lifestyle data could help models provide more robust, more personalized insights. This would increase the diagnostic accuracy, and would make the explanations more clinically meaningful.

Another area of emerging research is interpretability in real-time during surgeries. Imagine an AI system to guide surgeons by visualizing suspicious tissue in real time using clear visual signals provided by techniques such as Grad-CAM. This can help minimize the risk of incomplete tumor resection and improve patient outcomes.

Telemedicine is also an important frontier. As telehealth grows, XAI can be a powerful tool to help solve the provider shortage in rural areas [3]. By not only offering predictions but also clear explanations behind these predictions, XAI solutions can enable local healthcare workers to make informed decisions without constant supervision by a specialist [30]. Wide-scale democratization of advanced diagnostics has the potential to transform access to high-quality care globally.

And in the near future, we can also expect a closer relationship between explainability and trust calibration. Models will probably be developed not only to explain their outputs, but to quantify uncertainty, so clinicians can estimate when AI recommendations should be trusted vs. when human judgement should prevail. Taken together, these innovations will result in more secure, ethical, and widely accepted AI systems for use in oncology.

8. References

[1]. Ansari, Z. A., Tripathi, M. M., and Ahmed, R. (2025). "The role of explainable AI in enhancing breast cancer diagnosis using machine learning and deep learning models." Discover Artificial Intelligence, 5, 75.

[2]. Badrie, S. (2025). "Skin tone bias in AI dermatology tools: Are we building inclusive systems?" RCSI Student Medical Journal, May 13.

[3]. Balakrishnan, K., et al. (2025). "Artificial intelligence in rural healthcare delivery: Bridging gaps and enhancing equity through innovation." arXiv preprint, August.

[4]. Balve, A.–K., and Hendrix, P. (2024). "Interpretable breast cancer classification using CNNs on mammographic images." arXiv preprint, August.

[5]. Chanda, T., et al. (2023). "Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma." preprint, March.

[6]. European Journal of Cancer. (2022). "Explainable artificial intelligence in skin cancer recognition: A systematic review."

[7]. Frontiers in Oncology. (2024). "Explainable AI techniques enhance clinicians' understanding and trust by interpreting medical images and highlighting predictive factors." Harnessing Explainable AI for Precision Cancer Diagnosis and Prognosis.

[8]. Gaur, A., et al. (2025). "Explainable CNN for brain tumor detection and classification." PubMed Central (PMC), published 3 months ago.

[9]. Ghasemi, A., Hashtarkhani, S., Schwartz, D. L., and Shaban-Nejad, A. (2024). "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review." Cancer Innovation, 3, e136.

[10]. Hammad, M., ElAffendi, M., Abd El-Latif, A. A., Ateya, A. A., Ali, G., and Plawiak, P. (2025). "Explainable AI for lung cancer detection via a custom CNN on CT images." Scientific Reports, 15, 12707.

[11]. Huang, Z., Zhang, X., Ju, Y., et al. (2024). "Explainable breast cancer molecular expression prediction using multi-task deep learning based on 3D whole breast ultrasound." Insights into Imaging, 15, 227.

[12]. Kakish, T. (2025). "Transforming dermatopathology with AI: Addressing bias, enhancing interpretability, and shaping future diagnostics." Dermatological Reviews, 32(1).
[13]. Longo, L. (2024). "Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions." Information Fusion.
[14]. Montoya, D. N., Roberts, J. S., and Sánchez Hidalgo, B. (2024). "Towards fairness in AI for melanoma detection: Systemic review and recommendations." arXiv preprint, November.
[15]. Nature. (2025). "Lung cancer detection using CNNs augmented with Grad-CAM, LIME, and SHAP for enhanced interpretability."
[16]. Rezaeian, O., Asan, O., and Bayrak, A. E. (2025). "The impact of AI explanations on clinicians' trust and diagnostic accuracy in breast cancer." Applied Ergonomics, 129, 104577.
[17]. Rezaeian, O., Bayrak, A. E., and Asan, O. (2025). "Explainability and AI confidence in clinical decision support systems: Effects on trust, diagnostic performance, and cognitive load in breast cancer care." arXiv preprint, January.
[18]. Samita Bai, S., et al. (2024). "Breast cancer diagnosis: A comprehensive exploration of explainable artificial intelligence (XAI) techniques." arXiv preprint, June.
[19]. ScienceDirect. (2025). "Integrating LIME, Grad-CAM, and SHAP for enhanced accuracy." Published online, September.
[20]. Selvaraju, R. R., et al. (2016). "Grad-CAM: Visual explanations from deep networks via gradient-based localization." arXiv preprint, October.
[21]. Talaat, F. M., Gamel, S. A., El-Balka, R. M., Shehata, M., and ZainEldin, H. (2024). "Grad-CAM enabled breast cancer classification with a 3D Inception-ResNet V2: Empowering radiologists with explainable insights." Cancers, 16(21), 3668.
[22]. "The Cancer Genome Atlas (TCGA)." (2025). Wikipedia. Last revised June 2025. Available at: https://en.wikipedia.org/wiki/The_Cancer_Genome_Atlas Last accessed on 11 October 2025.
[23]. Wyatt, L. S., et al. (2024). "Explainable AI (XAI) for oncological ultrasound image analysis: A systematic review." Applied Sciences, 14(18), 8108.
[24]. "What is the role of explainability in medical artificial intelligence?" (2025). PubMed Central (PMC).
[25]. "Explainability and AI confidence in clinical decision support systems." (2025). arXiv preprint, January.
[26]. "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective." (n.d.). PubMed Central (PMC).
[27]. "Explainable artificial intelligence (XAI): Closing the gap between image analysis and navigation in complex invasive diagnostic procedures." (n.d.). PubMed Central (PMC).
[28]. "Decoding the black box: Explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning – A state-of-the-art systematic review." (n.d.). Systematic Review Journal.
[29]. "How explainable artificial intelligence can increase or decrease trust." (2024). Journal of Medical Internet Research.
[30] Ayush Kashyap et al., Design and Implementation of an Intelligent Loan Eligibility System Using Machine Learning Techniques, TEJAS Journal of Technologies and Humanitarian Science, ISSN-2583-5599, Vol.04, I.02 (2025), https://doi.org/10.63920/tjths.42002
[31]. "Investigation into application of AI and telemedicine in rural communities: A systematic literature review." (2025). PubMed.