ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov493

Explainable AI (XAI) for Convolutional Neural Networks in Early Cancer Detection

K. Janani[1]

[1]Assistant Professor, PG Department of Computer Applications, Marudhar Kesari Jain College for Women (Autonomous), Vaniyambadi

Publication Date: 2026/11/19

Abstract: Convolutional Neural Networks (CNNs) have become useful tools for analysing medical images because they can learn visual patterns associated with normal and abnormal tissue. In cancer screening and diagnosis, however, a high classification score alone is not sufficient: clinicians also need to understand which image regions influenced a model's output. This paper proposes an explainable deep-learning framework that combines CNN-based image classification with Explainable Artificial Intelligence (XAI) methods to support early cancer detection from histopathological images. The proposed workflow includes image collection, quality assessment, preprocessing, augmentation, CNN training, prediction, and explanation generation. Grad-CAM is used to produce class-specific heatmaps, while LIME and SHAP are considered for local image-region and feature-contribution explanations. The explanations are intended to help users inspect whether the model focuses on tissue regions that are relevant to the predicted class rather than irrelevant background patterns. Model performance can be assessed using accuracy, precision, recall, F1-score, specificity, and area under the ROC curve. Explanation quality should additionally be examined through localization agreement, stability, and review by qualified domain experts. The framework is presented as a research proposal; its diagnostic performance and clinical usefulness must be established through controlled experiments and external validation. It is designed as a decision-support approach and is not intended to replace professional diagnosis.

Keywords: *Explainable Artificial Intelligence, Convolutional Neural Network, Early Cancer Detection, Histopathology, Grad-CAM, SHAP, LIME, Medical Image Analysis.*

How to Cite: K. Janani (2026) Explainable AI (XAI) for Convolutional Neural Networks in Early Cancer Detection. *International Journal of Innovative Science and Research Technology*, 11(11), 661-666. https://doi.org/10.38124/ijisrt/26nov493

I. INTRODUCTION

Cancer outcomes are strongly influenced by the stage at which a disease is identified. Screening programmes, imaging investigations, and pathological examination therefore play important roles in recognising suspicious changes at an early point. Histopathological assessment involves examining tissue samples under a microscope, and the interpretation may require careful attention to cellular structure, tissue arrangement, nuclei, and other visual characteristics. The growing availability of digitised slides has created opportunities to apply computer vision methods to assist image analysis.

CNNs can learn hierarchical visual representations directly from image data. Earlier convolutional layers commonly respond to basic visual patterns, while deeper layers can represent more complex structures. These capabilities have encouraged the use of CNNs for classifying tissue images as normal, benign, or malignant, and for identifying patterns associated with particular cancer types. Nevertheless, a model may produce a confident prediction without providing a clear account of the visual evidence behind it. This lack of transparency can make it difficult for medical professionals to assess whether a prediction is reasonable.

Explainable Artificial Intelligence seeks to make model behaviour more understandable by presenting information about the factors that contribute to an output. For image-based CNNs, visual explanation methods can highlight influential regions, while model-agnostic approaches can estimate the contribution of image segments or input features. Grad-CAM, LIME, and SHAP offer different forms of explanation and should not be treated as interchangeable. A heatmap, for example, indicates regions associated with a model response; it does not by itself prove that the highlighted region is clinically meaningful.

This study proposes an integrated CNN-XAI framework for early cancer detection using histopathological images. The central aim is to combine predictive performance with inspectable explanations, and to evaluate both aspects rather than assuming that a visually appealing explanation guarantees reliability.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov493

➢ *Objectives*

- To design a CNN-based framework for classifying histopathological images into relevant tissue categories.
- To investigate suitable image preprocessing and augmentation procedures for improving model training.
- To integrate Grad-CAM, LIME, and SHAP to explain selected CNN predictions.
- To examine whether the highlighted image regions correspond with annotated or expert-identified tissue areas.
- To evaluate classification performance using accuracy, precision, recall, F1-score, specificity, and ROC-AUC.
- To assess explanation quality through stability, localization agreement, and expert review.
- To identify limitations related to dataset imbalance, image variation, model bias, and clinical generalisation.

II. LITERATURE REVIEW

Deep learning has become an established research direction in computational pathology. CNN-based approaches have been investigated for tissue classification, tumour-region identification, and analysis of whole-slide images. Srinidhi, Ciga, and Martel surveyed deep neural network methods in computational histopathology and discussed learning strategies, datasets, and outstanding methodological challenges [1]. Their review highlights the importance of dataset design and evaluation when developing image-analysis systems.

Public datasets have enabled researchers to compare image-classification methods under more consistent conditions. BreakHis contains breast tumour histopathology images with benign and malignant labels across different magnification levels. BACH and CAMELYON datasets provide additional examples of breast tissue and lymph-node image analysis tasks. These datasets differ in image scale, annotation type, and prediction target; therefore, the selected dataset must match the precise research question.

The need to understand deep-learning decisions has led to the development of visual explanation methods. Selvaraju et al. introduced Gradient-weighted Class Activation Mapping (Grad-CAM), which uses gradients flowing into convolutional feature maps to create class-specific localisation maps [2]. The resulting heatmap can be superimposed on an input image to show areas that contributed strongly to a selected class score. Grad-CAM is useful for visual inspection, although its relatively coarse maps and sensitivity to model behaviour should be considered.

Ribeiro, Singh, and Guestrin proposed Local Interpretable Model-agnostic Explanations (LIME), which approximates a model locally around an individual prediction using an interpretable surrogate [3]. For images, LIME can perturb or mask superpixels and estimate which regions influence a prediction. Lundberg and Lee introduced SHapley Additive exPlanations (SHAP), a unified framework for assigning contribution values to input components [4]. Applying SHAP to images requires an appropriate image masker or segmentation strategy and can be computationally demanding.

Recent work on explainability in medical imaging emphasises that explanation maps should be evaluated rather than judged only by appearance. Explanations may be unstable, may highlight irrelevant image artefacts, or may vary with model architecture and preprocessing. Histopathology images also contain variations caused by staining, scanner settings, magnification, and tissue preparation. These factors make external validation and expert assessment important.

The proposed study brings CNN classification and multiple explanation methods into one evaluation pipeline. Its contribution is not the invention of a new explanation algorithm; rather, it is a structured approach for comparing predictive performance with the usefulness and consistency of explanations in an early cancer detection task.

III. METHODOLOGY

The proposed methodology consists of dataset selection, image preparation, CNN model development, explanation generation, and evaluation. A breast histopathology classification dataset such as BreakHis or BACH may be selected for an initial implementation. The final dataset and class labels must be fixed before model training. Results from different datasets should not be combined unless their labels and acquisition conditions are made compatible.

➢ *Dataset Preparation*

The study may use a publicly accessible, appropriately licensed histopathology image dataset. Candidate resources include:

- BreakHis: breast histopathology images labelled as benign or malignant, available at multiple magnification levels.
- BACH (Breast Cancer Histology): microscopy images organised into normal, benign, in situ carcinoma, and invasive carcinoma categories.
- CAMELYON16: whole-slide lymph-node images designed for detecting breast cancer metastases.

The dataset should be reviewed for class distribution, image resolution, magnification, staining variation, duplicate images, and annotation quality. Splitting must be performed at patient or slide level wherever identifiers are available, so that patches from the same patient or slide do not appear in both training and testing partitions. This reduces the risk of data leakage and overly optimistic performance estimates.

➢ *Image Preprocessing*

Preprocessing prepares the images for consistent model input. The following operations may be considered according to the dataset:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov493

- Remove unreadable or corrupted files and verify the associated labels.
- Resize images to the input dimensions required by the selected CNN architecture while preserving important tissue detail.
- Normalize pixel values and, where justified, investigate stain normalization to reduce colour variation.
- Apply training-only augmentation such as modest rotation, flipping, cropping, and brightness variation.
- Separate training, validation, and test data at patient or slide level before augmentation.
- Use class weighting or a suitable resampling strategy when the class distribution is imbalanced.

Augmentation should be restricted to the training data. Transformations must preserve the diagnostic meaning of the image and should not create unrealistic tissue patterns.

➢ *Feature Learning Using CNN*

A CNN learns image representations through convolutional filters, nonlinear activation functions, and pooling operations. The proposed model may be developed from a baseline CNN or a transfer-learning architecture such as VGG, ResNet, or EfficientNet. The final architecture should be selected through validation rather than assumed to be superior.

The convolutional blocks extract local and progressively more complex patterns. Pooling or strided operations reduce spatial dimensions, and the classification head maps the learned representation to the selected tissue classes. A softmax layer may be used for mutually exclusive classes, while a sigmoid output may be appropriate for a binary task.

Training should use a defined loss function, optimizer, learning rate, batch size, and stopping criterion. Hyperparameters must be selected using validation data. The test set should remain untouched until final evaluation.

➢ *CNN Classification*

The trained network predicts the probability or score associated with each class. For a binary experiment, the output may represent benign and malignant categories. For a multi-class experiment, categories may include normal, benign, in situ carcinoma, and invasive carcinoma, provided that the chosen dataset supports those labels.

The predicted class is obtained using a predefined decision rule. The confidence score should be retained alongside the class output, but a high score must not be interpreted as clinical certainty. Cases with low confidence, conflicting explanations, or poor image quality should be flagged for specialist review.

➢ *Explainable AI Integration*

The XAI component generates complementary explanations for selected CNN predictions. The proposed methods are:

- *Grad-CAM*

Grad-CAM uses the gradient of a target class score with respect to feature maps in a selected convolutional layer. The weighted feature maps are combined to create a coarse heatmap. The heatmap is resized and overlaid on the original image to indicate regions associated with the predicted class. The target layer, class score, and colour scale should be recorded so that explanations can be reproduced.

- *LIME*

LIME divides an image into interpretable regions, commonly superpixels, and creates perturbed versions by masking selected regions. The CNN is queried for each perturbed image, and a simple local surrogate model estimates which regions most influence the selected prediction. The segmentation method, number of perturbations, and random seed should be documented because they can affect the explanation.

- *SHAP*

SHAP estimates the contribution of input components to a model output using Shapley-value concepts. For image inputs, the image can be represented through superpixels or other meaningful segments, and a suitable masker can be used to assess the effect of hiding regions. The resulting attribution map indicates regions that increase or decrease the model score under the selected SHAP procedure. Computational cost and sensitivity to the masking strategy should be considered.

The three methods provide different views of a model prediction. Agreement among them may strengthen the case for further inspection, but disagreement should not be concealed. Explanations should be reviewed together with the original image and the model output.

➢ *Explanation Validation*

Explanation quality should be assessed separately from classification accuracy. Where lesion or tissue-region annotations are available, the highlighted regions can be compared with reference annotations using suitable localization measures such as intersection over union or pointing-game accuracy. Explanation stability can be examined by applying small, clinically reasonable image transformations and measuring whether the explanation changes substantially. Domain experts may also review whether highlighted regions are relevant to the target pathology. Such review should use a predefined assessment protocol.

➢ *Performance Evaluation*

The CNN classifier can be evaluated using the following measures:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov493

Table 1 Performance Evaluation

Metric	Purpose
Accuracy	Proportion of all test cases classified correctly.
Precision	Proportion of predicted positive cases that are truly positive.
Recall / Sensitivity	Proportion of actual positive cases correctly identified.
Specificity	Proportion of actual negative cases correctly identified.
F1-score	Harmonic mean of precision and recall.
ROC-AUC	Measures discrimination across classification thresholds.
Localization agreement	Measures overlap or positional agreement between explanation maps and reference regions, when annotations are available.
Explanation stability	Assesses consistency of explanations under controlled, small input changes.

For early cancer detection, recall/sensitivity is particularly important because missed positive cases can be consequential. Precision and specificity must also be reported to understand false alarms. Confidence intervals and class-wise results should be included where sample size permits. The model should be evaluated on an independent test set and, ideally, an external dataset from a different source.

➢ *Ethical, Privacy and Clinical Considerations*

Only datasets with appropriate access permissions should be used. Patient identifiers must be removed or protected, and data handling should follow applicable institutional and legal requirements. The study should report dataset provenance, inclusion criteria, class balance, and known demographic or acquisition limitations where such information is available. The proposed system is a research decision-support tool; it must not be presented as an autonomous diagnostic system. Any future clinical use would require prospective evaluation, regulatory review where applicable, and oversight by qualified healthcare professionals.

➢ *Overall Workflow*

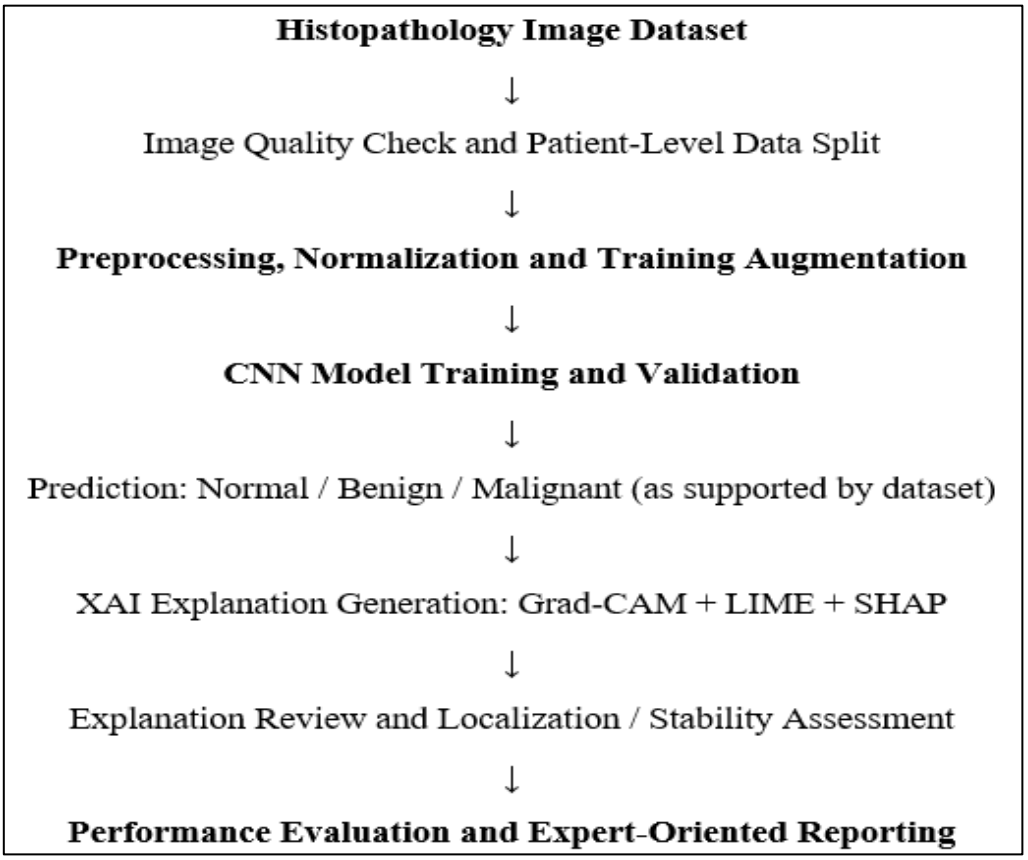


Fig 1 Proposed Workflow of the CNN-XAI Framework for Histopathological Image Analysis.

IV. RESULTS AND DISCUSSION

➢ *Classification Performance*

The proposed model should be evaluated on a held-out test set after the training and validation stages are completed. The following table provides a reporting format. Values are intentionally left blank because no model has been trained and no experimental measurements have been supplied.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov493

Table 2 Classification Performance

Performance metric	Experimental result
Accuracy	To be measured
Precision	To be measured
Recall / Sensitivity	To be measured
Specificity	To be measured
F1-score	To be measured
ROC-AUC	To be measured
External test performance	To be measured

The results should include the confusion matrix and class-wise scores. If several CNN architectures are tested, the comparison should use the same data partitions and evaluation procedure. A model should not be selected solely on accuracy when the classes are imbalanced.

➢ *Explanation Analysis*

Grad-CAM, LIME, and SHAP outputs should be generated for correctly classified cases, misclassified cases, and uncertain predictions. Representative examples should show the original image, predicted class, confidence score, and corresponding explanation map. The analysis should identify whether the explanation focuses on tissue structures of interest or on irrelevant regions such as image borders, labels, staining marks, or scanning artefacts.

Where expert annotations are available, localization measures can be calculated. Where annotations are unavailable, explanation maps should be described as qualitative visualisations and should not be claimed to demonstrate lesion localization accuracy. Expert assessment can provide additional evidence, but the review protocol and number of reviewers must be reported.

➢ *Comparison of XAI Methods*

Table 3 Comparison of XAI Methods

Method	Explanation output	Considerations
Grad-CAM	Class-specific heatmap over image regions	CNN-oriented; heatmaps may be coarse.
LIME	Local importance of perturbed image segments	Model-agnostic; results can vary with segmentation and perturbations.
SHAP	Contribution values for image segments or features	Flexible attribution framework; may require substantial computation.

The comparison should consider visual clarity, repeatability, computational cost, and agreement with reference regions. No single explanation method should be assumed to provide a complete account of the model's reasoning.

➢ *Discussion*

The proposed framework addresses two connected requirements in medical image analysis: identifying suspicious tissue patterns and making the model's output more inspectable. CNNs provide a mechanism for learning image features, while XAI methods can expose regions or components associated with individual predictions. This combination may help researchers and clinicians identify cases where a model appears to rely on inappropriate visual cues.

Several limitations must be considered. A model trained on one dataset may not generalise to images collected using different scanners, staining protocols, or patient populations. Class imbalance and data leakage can inflate reported performance. Explanation methods are approximations and may be sensitive to implementation choices. In addition, highlighting a region does not establish a causal explanation or prove that the region is diagnostically decisive. These issues require careful experimental design, independent validation, and expert review.

V. CONCLUSION

This paper presents a proposed Explainable AI framework that integrates CNN-based histopathology image classification with Grad-CAM, LIME, and SHAP to support research into early cancer detection. The framework covers dataset preparation, preprocessing, model development, prediction, explanation generation, and evaluation. It places emphasis on assessing both predictive performance and explanation quality rather than treating model accuracy as the only measure of success.

The study does not claim clinical effectiveness or report experimental performance because implementation results have not yet been obtained. Future work should train and compare suitable CNN architectures, validate the model using independent datasets, measure explanation localization and stability, and conduct structured review with qualified pathology experts. With rigorous validation and appropriate clinical oversight, explainable image-analysis systems may contribute to transparent and responsible decision-support research.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov493

REFERENCES

[1]. C. L. Srinidhi, O. Ciga, and A. L. Martel, “Deep neural network models for computational histopathology: A survey,” Medical Image Analysis, vol. 67, Article 101813, 2021. https://doi.org/10.1016/j.media.2020.101813

[2]. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. https://doi.org/10.1109/ICCV.2017.74

[3]. M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should I trust you?’: Explaining the predictions of any classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

[4]. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, vol. 30, 2017.

[5]. CAMELYON16 Challenge, “Grand Challenge: CAMELYON16 dataset.” https://camelyon16.grand-challenge.org/

[6]. F. A. Spanhol et al., “A dataset for breast cancer histopathological image classification,” IEEE Transactions on Biomedical Engineering, vol. 63, no. 7, pp. 1455–1462, 2016. https://doi.org/10.1109/TBME.2015.2496264