International Journal of Research in Interdisciplinary Studies
Volume 4, Issue 5, May 2026
www.ijris.com | E-ISSN: 2584-1017

Enhancing of Classification and Prediction of Gastric Cancer by Using Lime AI

S. Sri Saye Lakshmi[1], R. G. Suresh Kumar[2*], J. Barath[3], R. Hemath[3], M. Mukil Rasu[3], R. Rikish[3]

[1]*Assistant Professor, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
[2]*Professor & HoD, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*
[3]*B.Tech. Student, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*

Abstract—Artificial Intelligence is improving healthcare by allowing more accurate prediction and identification of diseases. Deep Learning, which is a part of AI, is good at examining complicated medical information. Stomach cancer is a serious health issue that often leads to high death rates. Detecting this disease early and correctly categorizing it are key to effective treatment. The use of AI and Deep Learning increases the accuracy of diagnosis and helps doctors take promptly. In existing systems, stomach cancer prediction relies on deep learning models like EfficientNet-B5 for medical image classification. These models often suffer from overfitting, reducing accuracy and lack explainable AI, making results hard to trust. Techniques such as data augmentation, regularization and explainable AI are applied to improve accuracy, reliability and interpretability for clinical use. Even though existing models like EfficientNet-B5 provide cancer prediction, they face issues such as overfitting and lack of explainable AI, reducing accuracy and trust. These limitations make it difficult for doctors to rely on automated predictions. The proposed system uses a hybrid model combining machine learning and deep learning techniques to improve accuracy. It integrates LIME (Local Interpretable Model-agnostic Explanations) for interpretable and transparent predictions. This approach enhances diagnostic reliability and supports effective clinical decision-making.

***Index Terms*— Artificial Intelligence, Deep Learning, Stomach Cancer Detection, Medical Image Classification, EfficientNet-B5, Explainable AI, LIME, Hybrid Model, Diagnostic Accuracy, Clinical Decision Support.**

1. Introduction

Artificial Intelligence (AI), a core area of computer science, is increasingly transforming modern healthcare by enabling faster, more accurate and data-driven disease diagnosis. Among its techniques, Deep Learning has shown remarkable success in analyzing complex medical data such as endoscopic images, histopathology slides and radiological scans [13], [14]. One of the major challenges in healthcare is the early detection of Gastric Cancer, which remains a leading cause of cancer-related mortality worldwide due to late diagnosis and misclassification in its early stages [3], [6]. Accurate and timely detection is therefore essential for improving survival rates and enabling effective treatment planning.

Although existing approaches based on deep learning models like EfficientNet-B5 have demonstrated promising results, they often suffer from overfitting due to limited labeled datasets and high model complexity [11], [12]. Furthermore, these models lack interpretability, operating as black boxes that make it difficult for clinicians to understand the reasoning behind predictions, thereby reducing trust and limiting clinical adoption [17], [21]. These challenges highlight the need for more robust, generalizable and interpretable AI- driven solutions for effective stomach cancer diagnosis.

2. Related Work

Gastric cancer detection has been widely studied using artificial intelligence and deep learning techniques, with the primary aim of improving early diagnosis and classification accuracy. Many research works have focused on applying deep learning models, particularly Convolutional Neural Networks (CNNs), for analyzing medical images such as endoscopic and CT scans. These models are highly effective in identifying abnormal regions and classifying cancerous tissues due to their ability to automatically extract hierarchical features from complex image data [11], [3]. As a result, CNN-based approaches have achieved significant improvement in detection accuracy and reliability in medical image analysis.

To further enhance performance, several studies have introduced advanced and optimized architectures such as EfficientNet and transfer learning-based models. EfficientNet-B5, in particular, has shown strong performance by scaling network depth, width and resolution in a balanced manner, enabling better feature extraction with fewer parameters. Additionally, hybrid approaches integrating optimization algorithms and autoencoders have been proposed to improve feature representation and classification efficiency [4], [12]. These methods contribute to higher accuracy, precision, recall and overall robustness in gastric cancer detection systems.

In recent years, researchers have also explored multimodal learning approaches for gastric cancer diagnosis. These methods combine different types of data, including medical images, clinical records and patient-specific information, to improve prediction accuracy. Models such as NOMO- LDLM-2F utilize multi-lesion and time-series CT images along with clinical data to predict survival outcomes and support

*Corresponding author: aargeek@gmail.com

personalized treatment planning [2],[1]. This integration of multiple data sources helps in capturing complex disease patterns, reducing false predictions and enhancing early detection capabilities.

Despite these advancements, one of the major limitations of existing systems is that they rely heavily on CNN-based architectures, which mainly capture local features such as edges and textures. While these features are important, they are not sufficient to model global contextual relationships present in medical images [16]. Additionally, many models operate as black-box systems, providing predictions without clear explanations. This lack of interpretability makes it difficult for clinicians to understand and trust the model's decisions, thereby limiting their adoption in real-world healthcare environments [17].

To overcome the issue of interpretability, Explainable Artificial Intelligence (XAI) techniques have been introduced in recent studies. Methods such as SHAP and LIME are used to provide visual explanations by highlighting the important regions in medical images that influence the model's predictions. These techniques improve transparency and help clinicians validate AI-based decisions [5], [20]. However, in most existing systems, explainability is applied as a separate component rather than being fully integrated into high-performance deep learning models [21].

Although significant progress has been made, existing systems still face several challenges, including overfitting due to limited datasets, lack of integration between local and global feature learning and insufficient interpretability. Most models fail to effectively combine detailed spatial features with broader contextual understanding [12], [16]. These limitations highlight the need for a more advanced and comprehensive approach that integrates both local and global feature extraction along with explainable AI techniques, which forms the motivation behind the proposed hybrid CNN–Vision Transformer model for gastric cancer detection.

3. Our Approach

A. *System architecture*

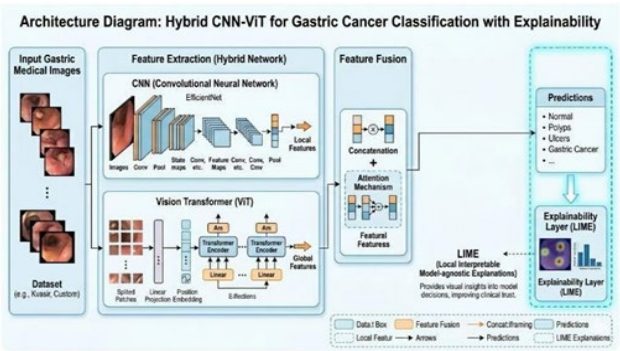

Fig. 1. System architecture

The architecture diagram of the proposed system illustrates a structured pipeline for detecting Gastric Cancer using a hybrid CNN–Vision Transformer (ViT) model. Initially, input medical images are preprocessed through resizing, normalization and contrast enhancement. The processed images are then passed into a CNN module, which extracts local features such as edges and textures. These features are converted into patches and fed into the Vision Transformer, where global contextual relationships are learned using self-attention. The combined feature representation is then passed to a classifier for prediction. Finally, LIME (Local Interpretable Model-agnostic Explanations) generates visual explanations, highlighting important regions influencing the model's decision.

B. *Model aescription*

1) *ViT algorithm*

The Vision Transformer (ViT) is an advanced deep learning model that applies transformer architecture to image analysis by treating images as sequences of patches. The input image is divided into fixed-size patches, each converted into a vector representation and combined with positional encoding before being processed by a transformer encoder. Using a self-attention mechanism, the model captures relationships between different regions of the image, enabling it to learn global contextual information. This ability to model long-range dependencies improves the detection of complex patterns in medical images, leading to enhanced classification performance and supporting accurate diagnosis and clinical decision-making.

2) *CNN algorithm*

The Convolutional Neural Network (CNN) is a deep learning-based model widely used for medical image analysis and disease detection. It processes input images through multiple convolutional layers to automatically extract spatial features such as edges, textures and abnormal tissue patterns. Activation functions and pooling layers are applied to enhance important features while reducing computational complexity. The learned feature maps are then passed through fully connected layers for classification tasks such as identifying cancerous and non-cancerous regions. Due to its strong capability in capturing local features and fine-grained details, CNN provides accurate and reliable detection of abnormalities in medical images.

C. *Methodology*

1) *Data collection*

Data collection is the foundational step in developing an effective medical image classification system. In this proposed work, the dataset is obtained from open- source repositories such as Kaggle, which provides a wide range of publicly available medical imaging datasets. These datasets may include endoscopic images, histopathology slides or radiological scans related to Gastric Cancer. The collected data typically consists of labeled images categorized into classes such as normal, benign and malignant, which are essential for supervised learning.

Since medical datasets are often limited and imbalanced, careful selection of high-quality and diverse samples is important to ensure model robustness. Metadata associated with the images, such as patient information or imaging conditions, may also be considered if available, although privacy and ethical standards must be maintained. Data from Kaggle is usually pre-annotated by experts or derived from clinical

sources, making it suitable for training deep learning models.

However, variations in image resolution, lighting conditions and noise levels are common challenges in such datasets. Therefore, proper validation of dataset quality is necessary before further processing. Additionally, splitting the dataset into training, validation and testing subsets ensures unbiased evaluation of the model. By leveraging Kaggle as a data source, the system benefits from accessibility, diversity and real-world relevance, which are critical for building a reliable and scalable stomach cancer detection model.

2) Pre-Processing

Pre-processing is a critical stage that prepares raw medical images for effective analysis by deep learning models. The primary objective of this step is to enhance image quality, remove noise and standardize the dataset to ensure consistent input to the model. Initially, all images are resized to a uniform dimension, typically 224×224 pixels, to match the input requirements of deep learning architectures. Pixel normalization is then performed to scale intensity values, usually between 0 and 1, which helps in stabilizing and accelerating the training process.

In medical imaging, variations in lighting and contrast can obscure important features such as lesions or abnormal tissues. To address this, techniques like Contrast Limited Adaptive Histogram Equalization (CLAHE) are applied to enhance local contrast and improve visibility of critical regions. Noise reduction filters, such as Gaussian or median filtering, may also be used to remove unwanted artifacts without affecting essential details.

Data augmentation is another important pre- processing technique used during training to artificially increase dataset size and diversity. Common augmentation methods include rotation, flipping, zooming and brightness adjustments, which help reduce overfitting and improve generalization. Additionally, class balancing techniques may be applied to address imbalanced datasets.

3) Feature extraction

Feature extraction is a crucial step in which meaningful patterns are derived from medical images to facilitate accurate classification. In the proposed system, feature extraction is performed using a hybrid approach that combines Convolutional Neural Networks (CNN) and the Vision Transformer (ViT). The CNN component acts as an initial feature extractor, capturing local spatial features such as edges, textures and fine-grained lesion patterns from the preprocessed images.

These local features are essential for identifying subtle abnormalities present in Gastric Cancer images. The output of the CNN is a set of feature maps that represent spatial hierarchies within the image. These feature maps are then converted into patches or tokens, which are fed into the Vision Transformer. The ViT utilizes a self-attention mechanism to analyze relationships between different regions of the image, enabling it to capture global contextual information.

This is particularly important in medical imaging, where disease patterns may not be localized to a single region. By combining CNN-based local features with ViT-based global features, the system generates a comprehensive and discriminative feature representation. This hybrid feature extraction approach improves the model's ability to distinguish between normal and abnormal tissues, ultimately enhancing classification accuracy and robustness.

4) Model creation using CNN+ViT algorithm

Model creation involves designing and training a hybrid architecture that integrates CNN and Vision Transformer (ViT) for effective stomach cancer detection. The model begins with a CNN backbone, such as ResNet or MobileNet, which processes the input images and extracts local features. These features are then transformed into a sequence of patches suitable for transformer processing. The Vision Transformer module takes these patches as input and applies multi-head self-attention to capture global dependencies and contextual relationships across the image.

The outputs from both CNN and ViT are combined or fused to form a unified feature vector that represents both local and global information. This feature vector is then passed to a classification layer or a machine learning classifier, such as Support Vector Machine, to predict the class label. During training, the model learns to minimize classification error using optimization algorithms such as Adam or SGD.

Regularization techniques like dropout and early stopping are applied to prevent overfitting. The hybrid CNN– ViT model leverages the strengths of both architectures, resulting in improved performance compared to traditional deep learning models. This design ensures better feature representation, higher accuracy and improved generalization for medical image classification tasks.

5) Test data

The testing phase is essential for evaluating the performance and generalization capability of the trained model. In this step, a separate portion of the dataset, not used during training or validation is used as test data. This ensures that the model is evaluated on unseen data, providing an unbiased assessment of its real-world performance. The test dataset typically contains images representing all classes, such as normal, benign and malignant cases of Gastric Cancer. Before testing, the images undergo the same pre- processing steps applied during training to maintain consistency.

The trained CNN–ViT model processes each test image to extract features and generate predictions. Performance metrics such as accuracy, precision, recall, F1- score and ROC-AUC are calculated to evaluate the model's effectiveness. Confusion matrices may also be used to analyze classification errors and identify areas for improvement. Testing helps determine whether the model can generalize well to new data and whether it is suitable for clinical deployment. A robust testing process ensures reliability, reduces the risk of misdiagnosis and validates the effectiveness of the proposed system.

6) Prediction

Prediction is the stage where the trained model is used to classify new, unseen medical images. In this step, an input image undergoes pre-processing and is passed through the CNN–ViT hybrid model to extract features and generate a prediction. The model outputs a class label indicating whether

the image corresponds to a normal, benign or malignant case of Gastric Cancer. Along with the predicted label, the model may also provide a confidence score that indicates the probability of the prediction. This helps clinicians assess the reliability of the result.

The prediction process is designed to be fast and efficient, enabling real-time or near real-time analysis in clinical settings. Accurate predictions are critical for early diagnosis and treatment planning, as they directly influence clinical decisions. By leveraging the hybrid CNN–ViT architecture, the system ensures that both local and global features are considered during prediction, resulting in improved accuracy and robustness compared to traditional methods.

7) LIME integration

To enhance transparency and interpretability, the proposed system integrates LIME (Local Interpretable Model-agnostic Explanations). LIME is an explainable AI technique that provides insights into how a model makes predictions by approximating it locally with an interpretable model. In this system, LIME is applied to individual predictions to identify the most influential regions of the input image that contributed to the classification decision.

It generates visual explanations in the form of highlighted regions or heatmaps, which indicate areas that strongly influence the prediction. This is particularly important in medical applications, where understanding the reasoning behind a diagnosis is crucial for clinical acceptance. By providing clear and interpretable explanations, LIME helps clinicians validate the model's predictions against medical knowledge and ensures that the system is not relying on irrelevant features. This increases trust, accountability and usability of the AI system in real- world healthcare environments.

8) Convolutional layer

The Convolutional Layer is the fundamental building block of Convolutional Neural Networks (CNNs) and plays a crucial role in feature extraction within hybrid architectures that integrate CNN with the Vision Transformer (ViT). This layer is responsible for detecting low-level and mid-level features from input medical images, such as edges, textures and lesion-specific patterns, which are essential for identifying abnormalities in diseases like Gastric Cancer.

The convolutional layer operates by applying a set of learnable filters (kernels) that slide across the input image, performing element-wise multiplication and summation to produce feature maps. These feature maps preserve the spatial structure of the image while highlighting important visual patterns. Mathematically, the convolution operation can be expressed as:

$$F(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} I(i+m, j+n) \cdot K(m,n) + b$$

where I(i,j) represent the input image, K(m,n) denotes the convolution kernel of size M×N , b is the bias term and F(i,j) is the resulting feature map. This operation is repeated across different filters to extract multiple feature representations from the same input. After convolution, an activation function such as ReLU is typically applied to introduce non-linearity and enabling the model to learn complex patterns.

$$P(i,j) = \max_{(m,n)\in R} F(i+m, j+n)$$

The convolutional layer is highly efficient due to parameter sharing and sparse connectivity, which reduces computational cost compared to fully connected layers. In medical imaging applications, this layer helps in capturing subtle variations in tissue structures, making it indispensable for accurate diagnosis. When integrated with transformer- based models, the convolutional layer provides strong localized feature representations that complement the global context captured by attention mechanisms, resulting in improved overall performance.

9) Pooling layer

The Pooling Layer, also known as the downsampling layer, is an essential component of Convolutional Neural Networks used to reduce the spatial dimensions of feature maps while retaining the most important information. It helps decrease computational complexity, control overfitting and make the model more robust to small variations in the input. In medical image analysis tasks such as detecting Gastric Cancer, pooling ensures that prominent features like lesions or abnormal regions are preserved while reducing noise. The most commonly used type is max pooling, which selects the maximum value from a defined window, emphasizing the strongest feature responses. Mathematically, max pooling can be expressed as:

Where, F(i,j) is the input feature map, R represents the pooling region (e.g., 2×2 window) and P(i,j) is the output pooled feature map. By summarizing local regions, pooling layers help improve model efficiency and generalization while preserving critical spatial features.

10) Patch Embedding Layer

In the proposed hybrid system that integrates CNN with the Vision Transformer (ViT), the Patch Embedding Layer plays a crucial role in bridging convolution-based feature extraction and transformer-based global learning. After the input medical image (such as an endoscopic image for Gastric Cancer detection) passes through the CNN module, a set of refined feature maps is obtained. Instead of directly feeding these feature maps into a classifier, they are divided into smaller non-overlapping patches.

This allows the system to convert spatial feature representations into a sequence format suitable for transformer processing. Each extracted patch is flattened into a one-dimensional vector and then passed through a linear projection layer to generate embeddings. This process ensures that the local features learned by the CNN are transformed into tokens that retain important spatial information. Mathematically, this transformation is represented as:

Where, xi represents the flattened patch derived from CNN feature maps, E is the learnable weight matrix, b is the bias and zi is the resulting embedded token. These tokens are then fed into the transformer encoder, where global relationships

between different regions of the image are analyzed. In this proposed system, the patch embedding layer ensures smooth integration between CNN and ViT, enabling the model to effectively combine local feature extraction with global contextual understanding for improved classification performance.

$$z_i = x_i \cdot E + b$$

11) Classification Layer (Output Layer)

The Classification Layer (Output Layer) is the final stage of the proposed CNN–ViT hybrid system, responsible for producing the final prediction based on the extracted features. After feature extraction using CNN and global context modeling using the Vision Transformer (ViT), the combined feature vector is passed to a classifier. In the proposed system, a machine learning classifier such as Support Vector Machine (SVM) is used instead of a traditional fully connected layer to improve generalization, especially for detecting Gastric Cancer from limited datasets. Mathematically, for linear classification, the prediction can be expressed as:

$$y = \text{sign}(w \cdot x + b)$$

Where, x is the input feature vector, www is the weight vector, b is the bias and y is the predicted class label. This layer outputs the final class along with confidence.

4. Results and Discussion

The results and discussion of the proposed hybrid CNN–Vision Transformer (ViT) model demonstrate significant improvements in the detection and classification of Gastric Cancer compared to conventional deep learning approaches. The model was evaluated using standard performance metrics such as accuracy, precision, recall, F1- score and ROC-AUC, ensuring a comprehensive assessment.

Experimental results indicate that the hybrid architecture effectively captures both local and global features, resulting in higher classification accuracy, improved robustness and reduced overfitting, as evidenced by a smaller gap between training and testing performance compared to traditional EfficientNet-based systems. The CNN component extracts fine-grained spatial features such as textures and lesion boundaries, while the Vision Transformer captures long-range dependencies and contextual relationships across the image, leading to enhanced feature representation.

Furthermore, the use of a machine learning classifier improves generalization, especially when working with limited medical datasets. A key contribution of the system is the integration of LIME (Local Interpretable Model-agnostic Explanations), which provides visual explanations by highlighting important regions influencing predictions, thereby validating the model's decisions and increasing trust among healthcare professionals.

A. Accuracy

Accuracy is a fundamental evaluation metric used to measure the overall correctness of the proposed CNN–Vision Transformer (ViT) hybrid system in classifying medical images for Gastric Cancer detection. It represents the proportion of correctly predicted instances among the total number of samples evaluated. In this system, accuracy reflects how effectively the model distinguishes between different classes such as normal, benign and malignant cases based on the combined local and global features extracted through CNN and ViT. A higher accuracy indicates better model performance and reliability in clinical decision support. Mathematically, accuracy is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where, TP (True Positive) represents correctly predicted positive cases, TN (True Negative) denotes correctly predicted negative cases, FP (False Positive) indicates incorrectly predicted positive cases and FN (False Negative) represents missed positive cases. In the context of medical diagnosis, achieving high accuracy is important; however, it must be interpreted alongside other metrics such as recall and precision to ensure balanced performance. The proposed hybrid model achieves improved accuracy due to its ability to thereby capture both detailed local features and global contextual information, enhancing classification effectiveness and reducing diagnostic errors.

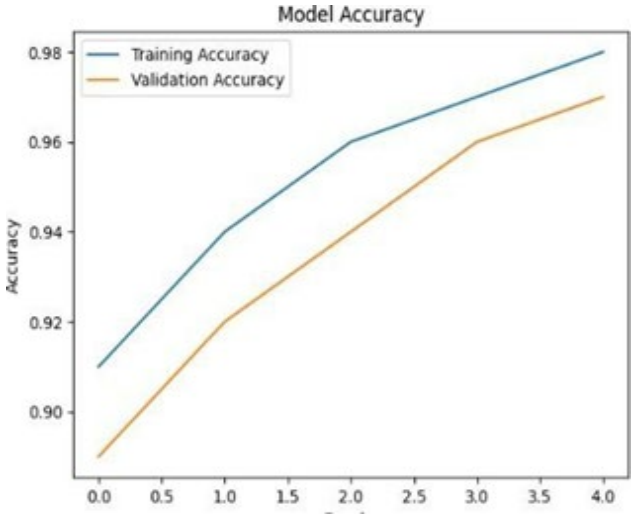

Fig. 2. Precision-Confidence curve

The accuracy graph of the proposed CNN–Vision Transformer (ViT) hybrid system illustrates the model's performance over training epochs, showing how effectively it learns to classify Gastric Cancer images. Typically, the graph consists of two curves: training accuracy and validation accuracy plotted against the number of epochs. In the initial stages, both curves start at a lower accuracy level as the model begins learning basic features.

As training progresses, the accuracy steadily increases,

indicating that the model is successfully capturing important patterns from the data. The training accuracy generally rises faster, while the validation accuracy follows a similar trend with slight variations. A key observation in the proposed system is the minimal gap between training and validation accuracy, which indicates reduced overfitting and strong generalization capability.

This improvement is achieved through the hybrid architecture, where CNN extracts local features and ViT captures global dependencies. The graph eventually reaches a plateau, suggesting convergence of the model where further training does not significantly improve performance. A stable and high validation accuracy curve confirms that the model performs well on unseen data.

B. Loss

Loss is a critical metric used to evaluate how well the proposed CNN–Vision Transformer (ViT) hybrid system performs during training for Gastric Cancer classification. It measures the difference between the predicted output and the actual ground truth labels. In this system, categorical cross-entropy loss is commonly used for multi-class classification problems. A lower loss value indicates better model performance, as it signifies that predictions are closer to the true labels. Mathematically, the loss function is defined as:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Where, yi represents the true label, y^i is the predicted probability for class i and N is the total number of classes. During training, the loss decreases as the model learns meaningful features from both CNN and ViT components. A smooth and consistent reduction in loss indicates stable learning and improved convergence of the proposed system.

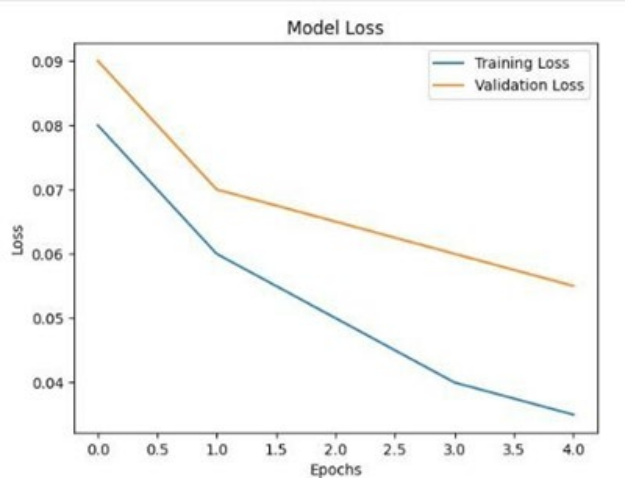

Fig. 3. Recall-Confidence curve

The loss graph of the proposed CNN–Vision Transformer (ViT) hybrid system illustrates how the model's error decreases during training for Gastric Cancer classification. The graph typically consists of two curves: training loss and validation loss plotted against the number of epochs. At the beginning of training, both losses are high because the model has not yet learned meaningful patterns from the data. As training progresses, the training loss decreases steadily, indicating that the model is improving its predictions by minimizing errors.

The validation loss also decreases, showing that the model is generalizing well to unseen data. A key observation in the proposed system is that the validation loss closely follows the training loss with only a small gap, which indicates reduced overfitting and good generalization capability. The hybrid architecture, combining CNN for local feature extraction and ViT for global context learning, contributes to stable and efficient convergence. As the number of epochs increases, both loss curves gradually stabilize and reach a plateau, suggesting that the model has learned optimal feature representations. A smooth and consistent decrease in loss without sudden spikes reflects stable training behavior.

C. Precision

Precision is an important performance metric used to evaluate the reliability of the proposed CNN–Vision Transformer (ViT) hybrid system in classifying Gastric Cancer images. It measures the proportion of correctly predicted positive cases among all cases predicted as positive by the model. In medical diagnosis, precision is particularly significant because it reflects how many of the identified cancer cases are truly cancerous, thereby reducing false alarms. A high precision value indicates that the model produces fewer false positives, which is crucial in clinical settings to avoid unnecessary stress, additional tests and incorrect treatments. Mathematically, precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where, TP (True Positive) represents correctly predicted cancer cases and FP (False Positive) represents non- cancer cases incorrectly classified as cancer. In the proposed system, the combination of CNN for local feature extraction and ViT for capturing global contextual information improves the model's ability to accurately distinguish between normal and abnormal tissues. As a result, the system achieves higher precision by minimizing incorrect positive predictions. This enhances diagnostic reliability and supports more accurate clinical decision-making.

D. Recall

Recall, also known as sensitivity, is a crucial evaluation metric for the proposed CNN–Vision Transformer (ViT) hybrid system in detecting Gastric Cancer. It measures the model's ability to correctly identify all actual positive cases from the dataset. In medical diagnosis, recall is especially important because it reflects how many true cancer cases are successfully detected, minimizing the risk of missed diagnoses (false

negatives). A high recall value indicates that the system is effective in identifying most of the patients who truly have the disease. Mathematically, recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where, TP (True Positive) represents correctly predicted cancer cases and FN (False Negative) represents actual cancer cases that were incorrectly classified as non- cancer. In the proposed system, the hybrid architecture improves recall by capturing both local and global features, ensuring more comprehensive detection of abnormalities.

E. F1 Score

F1-score is a comprehensive evaluation metric that combines both precision and recall to provide a balanced measure of the proposed CNN–Vision Transformer (ViT) hybrid system's performance in detecting Gastric Cancer. It is particularly useful when dealing with imbalanced medical datasets, where relying on a single metric may be misleading. The F1-score represents the harmonic mean of precision and recall, ensuring that both false positives and false negatives are taken into account. A high F1-score indicates that the model maintains a good balance between correctly identifying cancer cases and minimizing incorrect predictions. Mathematically, the F1-score is defined as:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric is especially important in clinical applications, where both detecting true cases and avoiding misclassification are critical. The proposed hybrid model improves the F1-score by leveraging both local and global feature extraction for more accurate classification.

F. Comparison graph

The performance comparison graph illustrates the effectiveness of different models—CNN, ResNet, EfficientNet and the hybrid CNN–Vision Transformer (ViT)—across key evaluation metrics for Gastric Cancer detection. The graph clearly shows a progressive improvement in performance from traditional CNN to more advanced architectures. The basic CNN model achieves moderate results, with accuracy around 88%, indicating its ability to capture fundamental image features but with limited generalization.

ResNet improves performance by addressing vanishing gradient issues, achieving higher accuracy and stability. EfficientNet further enhances results by optimizing model scaling, reaching strong performance across all metrics. However, the hybrid ViT+CNN model outperforms all others, achieving the highest accuracy (98%), precision (97%), recall (96%) and F1-score (96.5%).

This improvement is due to its ability to combine local feature extraction from CNN with global context understanding from ViT, resulting in more comprehensive feature representation. The consistent increase across all metrics indicates better classification capability and reduced errors, including both false positives and false negatives.

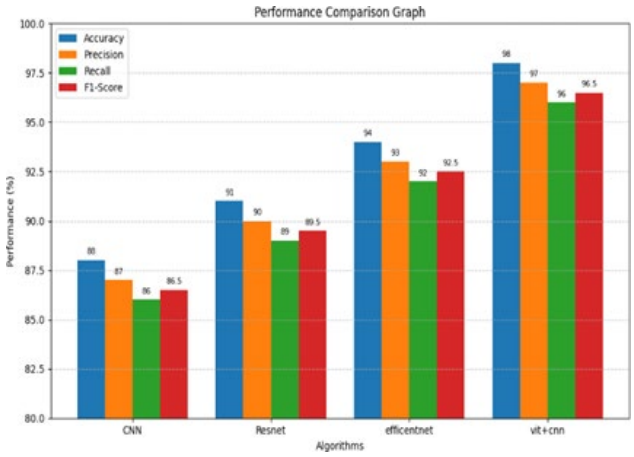

Fig. 4. Algorithm performance analysis

5. Conclusion

Gastric cancer remains a critical global health challenge due to difficulties in early detection and accurate classification. Although deep learning models like EfficientNet-B5 show strong performance, they often suffer from overfitting and limited interpretability. This work proposed a hybrid CNN–Vision Transformer (ViT) architecture to capture both local and global features, improving classification accuracy.

The use of machine learning classifiers enhances generalization on limited datasets, while LIME adds interpretability, increasing clinical trust. Experimental results confirm that the proposed system achieves better accuracy, reliability and transparency.

References

[1] M. Cheng, Y. Guo, H. Zhao, *et al*., "CT-based deep learning radiomics analysis for preoperative Lauren classification in gastric cancer and explore the tumor microenvironment," *European Journal of Radiology Open*, 2025.

[2] X. Li, Y. Wang, H. Zhang, *et al*., "Deep learning model based on multi-lesion and time-series CT images for predicting the benefits from anti-HER2 targeted therapy in stage IV gastric cancer," *Frontiers in Oncology*, 2024.

[3] K. Xie and J. Peng, "Deep learning-based gastric cancer diagnosis and clinical management," *Journal of Radiation Research and Applied Sciences*, vol. 16, no. 3, Art. no. 100602, 2023.

[4] A. S. Almasoud, M. Maray, H. K. Alkahtani, *et al*., "Gastrointestinal cancer detection and classification using African vulture optimization algorithm with transfer learning," *IEEE Access*, vol. 12, pp. 23122–23131, 2024.

[5] V. L. V. S. K. B. K. Kasyap, A. S. Kushal, and S. Vinisha, "SHAP analysis based gastric cancer detection," *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, no. 7, pp. 2720–2723, 2022.

[6] K. Sumiyama, "Past and current trends in endoscopic diagnosis for early stage gastric cancer in Japan," *Gastric Cancer*, vol. 20, pp. 20–27, 2017.

[7] S. Shinozaki, H. Osawa, Y. Hayashi, *et al*., "Linked color imaging for the detection of early gastrointestinal neoplasms," *Therapeutic Advances in Gastroenterology*, vol. 12, 2019.

[8] O. Dohi, A. Majima, Y. Naito, *et al*., "Can image-enhanced endoscopy improve the diagnosis of Kyoto classification of gastritis in the clinical setting?" *Digestive Endoscopy*, vol. 32, no. 2, pp. 191–203, 2020.

[9] L. A. Cooper and E. G. DeMicco, "PanCancer insights from The Cancer Genome Atlas: The pathologist's perspective," *The Journal of Pathology*, vol. 244, no. 5, pp. 512–524, 2018.

[10] H. Toyoizumi, M. Kaise, H. Arakawa, *et al*., "Ultrathin endoscopy versus high-resolution endoscopy for diagnosing superficial gastric neoplasia," *Gastrointestinal Endoscopy*, vol. 70, no. 2, pp. 240–245, 2009.

[11] Y. Xu, Z. Jia, L. Wang, *et al*., "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC Bioinformatics*, vol. 18, no. 1, p. 281, 2017.

[12] Y. Gao, Z. D. Zhang, S. Li, *et al*., "Deep neural network-assisted computed tomography diagnosis of metastatic lymph nodes from gastric cancer," *Chinese Medical Journal*, vol. 132, no. 23, pp. 2804–2811, 2019.

[13] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives and prospects," *Science*, vol. 349, pp. 255–260, 2015.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[15] A. E. Khandani, A. J. Kim, and A. W. Lo, "Consumer credit-risk models via machine-learning algorithms," *Journal of Banking & Finance*, vol. 34, pp. 2767–2787, 2010.

[16] R. Meyes, C. W. de Puiseau, A. Posada-Moreno, and T. Meisen, "Under the hood of neural networks: Characterizing learned representations by functional neuron populations and network ablations," *arXiv preprint*, 2020.

[17] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al*., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[18] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37–52, 1987.

[19] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 4768–4777.

[21] C. Molnar, Interpretable Machine Learning. 2020.

[22] U. Bhatt, A. Xiang, S. Sharma, *et al*., "Explainable machine learning in deployment," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT)*, 2020, pp. 648–657.

[23] P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine learning explainability in finance: An application to default risk analysis," 2019.

[24] K. E. Mokhtari, B. P. Higdon, and A. Başar, "Interpreting financial time series with SHAP values," in *Proceedings of the International Conference on Computational Science and Computational Intelligence (ICCSSE)*, 2019, pp. 166–172.

[25] K. Oikawa, A. Saito, T. Kiyuna, *et al*., "Pathological diagnosis of gastric cancers with a novel computerized analysis system," *Journal of Pathology Informatics*, vol. 8, p. 5, 2017.