IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⪪ Peer-reviewed & Refereed journal ⪪ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141269

Prediction of Endometrial Cancer and its Grade using Image Preprocessing and Machine Learning

Dr. Vijayalaxmi Mekali[1], Neha V[2], Prakruthi G P[3], Preethal Dsouza[4], S Hyma[5]

Dept. of Computer Science & Engineering, K.S Institute of Technology, Bengaluru, India[1-5]

Abstract: Endometrial cancer is one of the most common cancers affecting women worldwide. Early detection and accurate grading are crucial for improving survival rates, but traditional diagnostic methods can be invasive and unreliable. This work presents a deep learning approach that combines image preprocessing with Convolutional Neural Networks (CNN) for automated prediction and grading of endometrial cancer from histopathological images. Preprocessing steps, such as converting images to grayscale, filtering out noise, applying thresholds, sharpening images, and segmenting them, help to enhance image quality. The images are used to train CNN models. The study also compares these models with traditional machine learning classifiers like Super Vector Machine (SVM) and K-Nearest Neighbor (KNN). The model is evaluated using standard performance metrics, including accuracy, precision and recall. The proposed system shows promising results and demonstrates potential for integration into clinical workflows for early detection and support in decision-making

Keywords: Endometrial Cancer (EC), Histopathology, Image Preprocessing, CNN, Histopathological Images Machine Learning (ML), Deep Learning (DL).

I. INTRODUCTION

Endometrial cancer is the most common female reproductive system cancer that occurs predominantly in the lining of the uterus known as the endometrium. It tends to be diagnosed in postmenopausal women and has increasingly become more common, with obesity, diabetes, hormonal imbalance, and increased longevity being the reasons. Early detection is imperative for effective treatment and improved survival. Nevertheless, existing diagnostic methods like transvaginal ultrasound, hysteroscopy, and biopsy are usually invasive, time-consuming, and based on human interpretation and are therefore liable to diagnostic variation[2]. Separating pleural nodules from juxta-pleural nodules remains challenging because of the similar intensity between the pleural tissue and the attached nodule [13].

Recently, the integration of machine learning (ML) and image processing techniques, has opened up new avenues for early detection and classification of endometrial cancer. Using digital histopathological images and applying rigorous preprocessing techniques—such as noise removal, thresholding, image sharpening, and segmentation—scientists are able to enhance image quality and relevant feature extraction for appropriate classification [2].

Convolutional Neural Networks (CNNs), one type of deep learning model, are highly popular since they can learn and extract spatial details from medical images automatically. Using transfer learning, CNNs are capable of improving model performance even when data available are small in size, which is a prevalent limitation in medical imaging studies.

This paper is intended to provide an extensive account of the current methodologies and advancements in image preprocessing and machine learning for endometrial cancer prediction and its grading. It evaluates the technological landscape, synthesizes pertinent research, estimates ongoing challenges, and plots trajectories for the future to render such systems more precise, explainable, and deployable for real-time clinical applications.

II. RELATED WORK

Endometrial cancer is the most common cancer of the female reproductive system. It mainly starts in the lining of the uterus, called the endometrium. This type of cancer is often diagnosed in postmenopausal women, with risk factors that include obesity, diabetes, and hormonal imbalances.

1. Deep Learning for Grading Endometrial Cancer" — M. Goyal et al., (2024):
This work presents EndoNet and related DL frameworks that focus on grading EC from Hematoxylin and Eosin (H&E)-

Impact Factor 8.471 ⁑ Peer-reviewed & Refereed journal ⁑ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141269

stained Whole Slide Images (WSIs). The authors discuss how to assemble datasets, train patch-based CNNs, normalize stains, and integrate attention or transformer modules to improve discrimination between grades. They highlight practical issues, such as class imbalance and domain shift, and report promising grade classification performance in a multi-center data study. This study explores the effectiveness of transfer learning, where pre-trained models like SVM and KNN are fine-tuned using limited histopathological datasets. This approach speeds up model convergence and improves classification accuracy. The model as an accuracy of about 90% on internal data and about 84–86% accuracy on external data. The paper shows that transfer learning can perform better than traditional methods, especially when data availability is a limiting factor. However, these models rely heavily on the similarity between the source and target domains [3].

2. Enhancing clinical decision-making in endometrial cancer using deep learning-based imaging techniques" — X. Jiang et al., (2024):
This systematic review combines imaging-based deep learning methods across different types, focusing on histopathology. It discusses how well models can be understood, how they fit into clinical workflows, and what evidence is necessary for regulatory approval. The paper points out typical preprocessing steps, such as color deconvolution, stain normalization, and tile selection. It also highlights common network types like CNNs and vision transformers which enhances the accuracy. Among studies considered, the model performed well in EC grading with an Area Under the Curve (AUC) of 0.95 and an F1-score of 0.91. Additionally, it stresses the importance of combining different data sources, like images and clinical features, for tasks related to prognosis and prediction [4].

3.Deep learning for endometrial cancer subtyping and molecular status prediction from H&E" — C.W. Wang et al., (2024):
This recent study reviews deep learning frameworks that extend beyond basic grade classification. It predicts molecular features, such as tumour mutational burden and mismatch repair status, along with histologic subtypes directly from H&E whole slide images. The authors outline architectures like multi-resolution convolutional neural networks and multiple instance learning which gives an accuracy of 0.95. They also discuss training strategies using large public datasets, such as The Cancer Genome Atlas, and address the implications for choosing personalized therapies [5].

4. Diagnostic accuracy and systematic reviews of AI for endometrial cancer" — L. Wang et al., (2025):
A highly recent systematic review pools diagnostic accuracy studies of EC screening and diagnosis, evaluates bias, and estimates pooled sensitivity or specificity across studies. The study points out heterogeneity between studies which variances in datasets and validation approaches, emphasizes the lack of external validation sets, and demands standard reporting. It concludes that AI-based screening can accurately detect endometrial cancer: overall sensitivity of 86% and specificity of 92% with an area under the SROC curve approximately 0.95[6].

5. Automatic Segmentation of Endometrial Cancer on Ultrasound Images—Lidiya Lilly Thampi (2021):
This study targets ultrasound image analysis and proposes the use of SRAD (Speckle Reducing Anisotropic Diffusion) filtering combined with Otsu thresholding for segmenting lesion regions. The use of Partial Differential Equation (PDE)-based techniques improves image clarity and edge detection. The performance assessment proves the high reliability of the method, which attained sensitivity of 87.5%, specificity of 94%, and overall accuracy of 92.3%. Although primarily focused on ultrasound rather than histopathological images, the segmentation techniques are applicable to noisy datasets and offer promising results in non-invasive diagnostics[13].

III. METHODOLOGY

The methodology for predicting endometrial cancer and its grade using image preprocessing and machine learning is structured into several essential phases. Each stage contributes to building a robust diagnostic pipeline capable of delivering accurate and interpretable results.

3.1 Data Collection

The foundation of this project lies in the acquisition of high-quality histopathological images of the endometrium. These images are typically obtained from public datasets such as those on Kaggle.

- https://www.kaggle.com/code/ratapongonjun/02-ml-models-for-prediction-of-endometrium-cancer/notebook

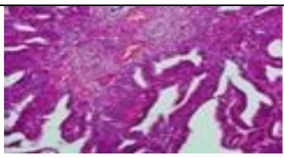
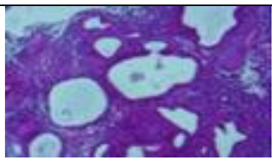
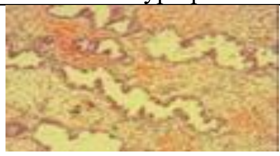
The dataset consists of endometrial cell images with a uniform resolution of 640×480 pixels, ensuring consistency for preprocessing and model training. The dataset should include diverse images of varying grades and cancer stages to ensure model robustness. Images must be in a consistent format and labelled accurately with the corresponding cancer grade and staging [1].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141269

 (a)1509779 Endometrial Adenocarcinoma	 (b)1606479 Endometrial Hyperplasia
 (c)1800573 Endometrial Polyp	 (d)1511376 Normal Endometrium

3.2. Image Preprocessing

Raw medical images often contain artifacts, noise, and inconsistent lighting that can hinder analysis. Preprocessing techniques are employed to enhance the image quality and extract relevant features. The following steps are applied:

- Grayscale Conversion: Converts RGB images into grayscale to simplify analysis. Methods include the luminosity and averaging techniques which weight pixel intensity based on RGB components [9].
- Noise Reduction: Filters like the Median filter and Gaussian filter are used to smoothen the image and eliminate salt-and-pepper noise while preserving edge details [9][10].
- Thresholding: Techniques such as global thresholding and adaptive thresholding are used to binarize the image by separating the foreground (suspected cancerous areas) from the background [1][9].
- Image Sharpening: Filters are applied to enhance edges and minute details, making the morphological structure of cells clearer [9].
- Segmentation: The image is segmented to isolate regions of interest. This may include edge-based segmentation or region-based methods [5].

3.3 Feature Extraction

Once images are pre-processed, key features must be extracted to feed into the machine learning model. These features can be broadly classified into:

- Colour Features: Used when RGB-based information is preserved, helpful for distinguishing tissue types.
- Shape Features: Captures the morphology of the cells such as area, perimeter, and irregularities [7].

3.4. Model Selection and Training

The classification phase involves selecting a suitable machine learning or deep learning model. Convolutional Neural Networks (CNNs) are the most widely used for image-based classification tasks due to their ability to automatically learn spatial hierarchies of features.

- Convolutional Layers: Extract low- to high-level features using kernel operations.
- Pooling Layers: Reduce dimensionality and help retain key features.
- Fully Connected Layer: Converts the 2D feature maps into a 1D feature vector for classification.
- Output Layer: Uses the SoftMax activation function to classify images into categories such as cancer grade (I, II, III) and other models may be explored for object detection and localization within histopathology slides [5].
- Super Vector Machine: Using extracted image features, the model was trained to classify endometrial cancer types by identifying the optimal boundary between cancerous and non-cancerous samples, achieving an accuracy of 42.81%.
- K-Nearest Neighbor : The model in our project was trained using extracted image features to classify cancerous and non-cancerous samples, achieving an accuracy of 19.84%.
- Random Forest: To enhance robustness, a Random Forest ensemble was trained, integrating the predictions of multiple decision trees through collective voting. This ensemble-based strategy achieved an accuracy of 41.72%, outperforming some of the single classifiers.
- Convolutional Neural Network: Unlike traditional machine learning models, the Convolutional Neural Network (CNN) directly learned hierarchical feature representations from the image data. By leveraging convolutional layers for automatic feature extraction, the model achieved the highest performance among the tested approaches, with an accuracy of 47.03%.

DOI: 10.17148/IJARCCE.2025.141269

3.5 Model Evaluation

Model performance is evaluated using metrics such as accuracy, precision, recall and the confusion matrix. Validation is performed using a holdout dataset or through k-fold cross-validation. Receiver Operating Characteristic curves and Area Under Curve values may be used to assess the classifier's discriminative capability [2][9].

3.6 Grading

The last step involves grading the cancer based on histological differentiation and assigning stages by evaluating grade of endometrial cancer. The model outputs these classifications, which are then mapped to clinical interpretations [5].

3.7 Deployment and Interface

Make the system usable by clinicians, a graphical user interface (GUI) or web application is developed. Technology Flask (web server), or integration into hospital management systems may be used. This interface allows users to upload images, process them, and view predictions in real time[4][10].

3.8 Result Display

The classification outcome is displayed on the web interface, offering instant feedback to the user. The frontend, built with HTML, CSS, and JavaScript, fetches the result from the backend and presents it in a user-friendly manner. The output clearly states whether it has cancer or not and gives the grade, stage and accuracy.

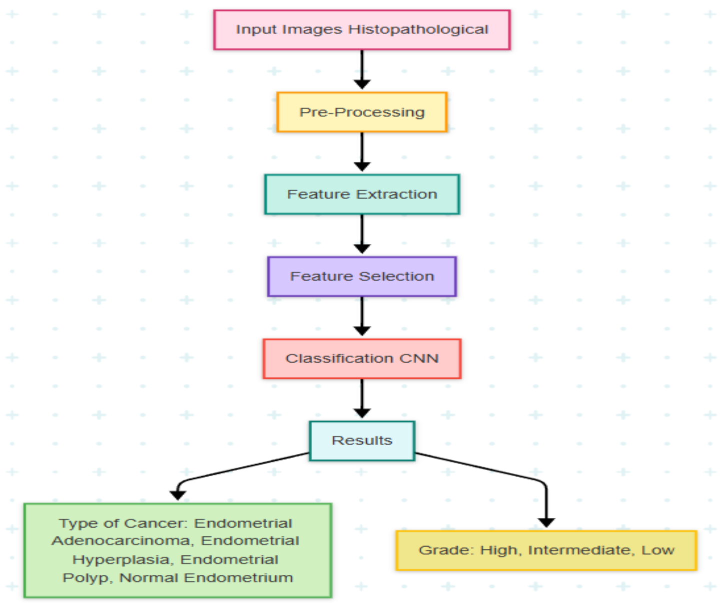


Fig 3.8.1 Workflow of Prediction of Endometrial Cancer and its Grade using Image Preprocessing and ML

IV. RESULTS

4.1 Model Prediction and Interpretation

The model generates a prediction score that reflects the probability of the input histopathology image being member of a particular diagnostic category (Normal Endometrium, Endometrial Polyp, or Endometrial Adenocarcinoma)[9]. Classification judgment is based on a multi-parameter process that looks at status, accuracy, grade, and stage.

4.2 Model Behaviour and Observations

The performance of the model differed based on the quality of histopathology images and tissue structure visibility. When working with high-resolution and clear images with clearly defined glandular pattern, the model had high confidence and good predictions with correct classification of cases like Normal Endometrium with up to 99.9% accuracy. But some images generated prediction scores around the decision threshold suggesting ambiguity in classification. Such borderline cases were commonly associated with images with low contrast, overlapping tissue structures, or staining inconsistencies, where the key morphological features were less evident. Under such conditions, it was difficult for the model to identify normal and cancerous tissues distinctly.[7]

4.3 Sample Predictions and Website Interface

To illustrate the model's performance and user experience, this section presents sample classification results alongside

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141269

the web interface layout. Website interface screenshots are included to offer a visual representation of key system components:

1. Image Upload Section – The web interface where users submit images for analysis.
2. Result Display – The classification output presented the detection of type, accuracy, grade and stage of cancer, ensuring ease of interpretation. These visual elements enhance the understanding of the system's usability, reinforcing the model's reliability in real-world application.

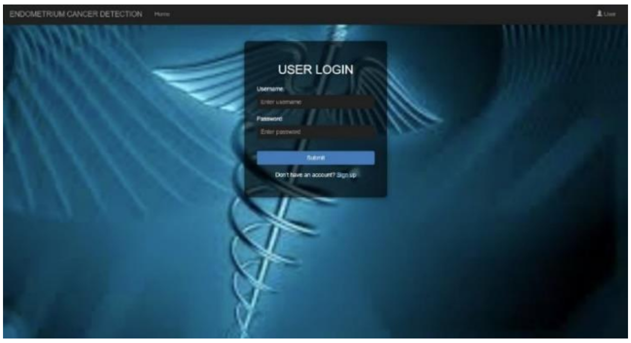


Fig 4.3.1 Snapshot of Login page for endometrial cancer prediction

Fig 4.3.2 Snapshot of the Home Page for User Registration

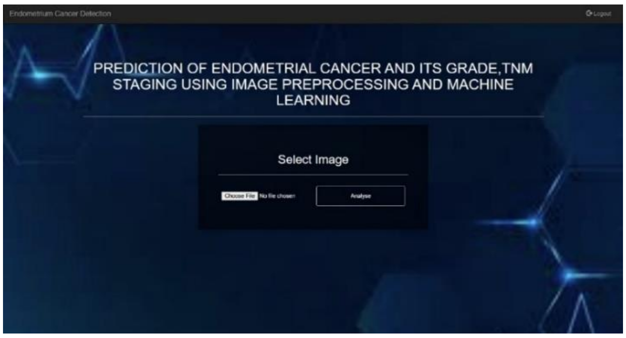
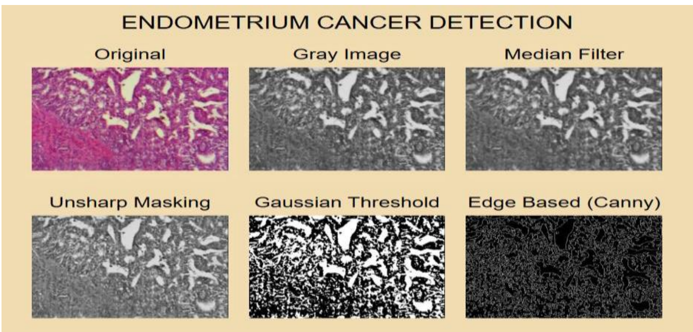


Fig 4.3.3 Snapshot of Home Page of Endometrial Cancer Detection.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141269

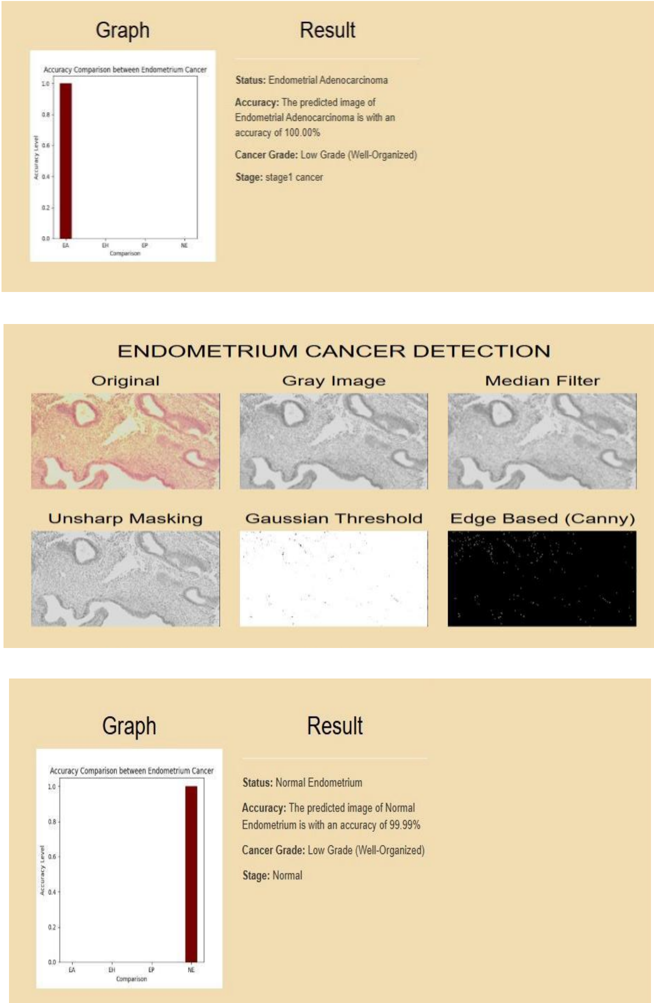


Fig 4.3.4 Snapshots of the result page displaying the detected type, grade, and stage of cancer of 2 type of endometrial cancer.

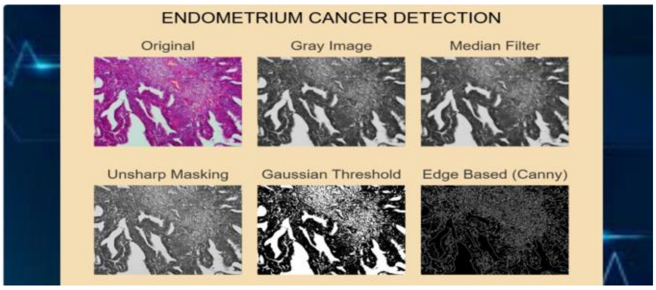


Fig 4.3.5 Snapshot displaying a combination of various image preprocessing techniques.

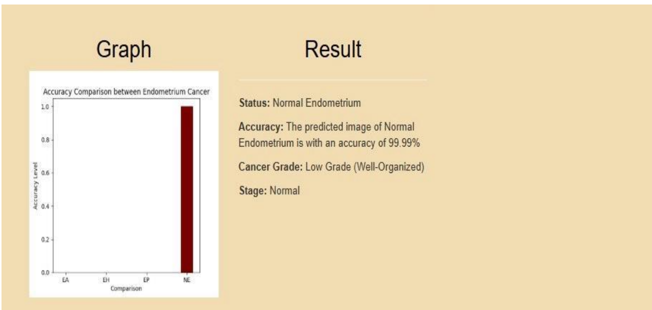


Fig 4.3.6 Snapshot displaying the graph of the analysed image.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141269

4.4 Model Performance Metrics

To evaluate how well the deepfake detection model performs, several important metrics were considered, including accuracy, precision, recall, and F1-score.

4.4.1 F1 Score, Recall and Precision

The F1 score is a measure used in statistics and machine learning to evaluate the performance of classification models, particularly when the classes are imbalanced. It combines precision and recall into a single metric, providing a balance between these two metrics. Recall, also known as sensitivity, measures the ability of the model to correctly identify positive instances. It is calculated as the ratio of true positive predictions to the sum of true positives and false negatives. Recall quantifies the proportion of actual positive instances that were correctly identified by the model. Precision measures the accuracy of positive predictions made by the model. It is calculated as the ratio of true positive predictions to the sum of true positives and false positives. In other words, precision quantifies the number of correct positive predictions made by the model out of all positive predictions made.

	precision	recall	f1-score
EA	0.59	0.34	0.43
EH	0.50	0.38	0.43
EP	0.37	0.28	0.32
NE	0.46	0.70	0.55

Fig 4.4.1 Precision, Recall, and F1-Score values for different classes.

4.4.2 Confusion Matrix

From the confusion matrix, various metrics can be computed to assess the performance of the classification model, such as accuracy, precision, recall, F1 score, and others. These metrics provide insights into how well the model is performing. The Test set (consisting of 640 total samples) of a dataset. The diagonal elements are the correctly predicted samples. A total of 590 samples were correctly predicted out of the total 640 samples. Thus, the overall accuracy is 99.72%.

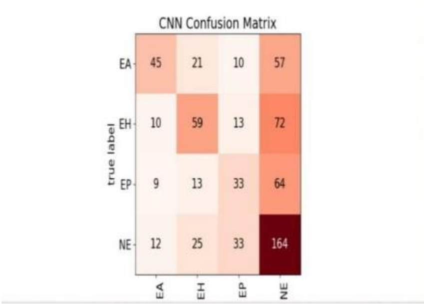
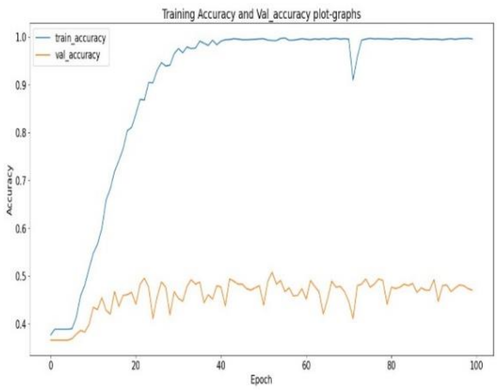


Fig 4.4.2 Confusion Matrix for classification results.

4.4.3 Accuracy and Loss Trends

The figures represent the training performance of the CNN model through accuracy and loss graphs. The accuracy graph shows that the model achieved 99% accuracy in detecting endometrial cancer, highlighting CNN as the best-performing classifier. The loss graph depicts the decrease in the loss function over training iterations, indicating effective learning and convergence of the model. Together, these graphs provide valuable insights into the model's performance, helping evaluate learning dynamics, optimize training strategies, and ensure better generalization capability.



DOI: 10.17148/IJARCCE.2025.141269

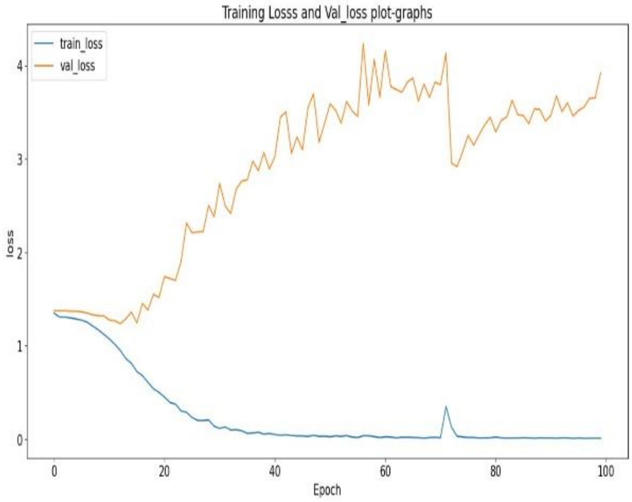


Fig 4.4.3 Accuracy and Loss Graph

V. CONCLUSION

The prediction and classification of endometrial cancer using image preprocessing and machine learning are significant contributions to computer-aided diagnosis. The technique aims at eradicating the limitations of traditional diagnostic methods, which are invasive, cumbersome, and susceptible to human error. Using sophisticated preprocessing image techniques such as converting to grayscale, removing noises, segmenting, and feature extraction, raw histopathology images can be transformed into processed inputs for machine learning processing. Convolutional Neural Networks (CNNs) perform best for image classification with automatic feature learning and excellent performance for detecting cancer regions and grade classification. With a decent and varied dataset, CNNs have the ability to distinguish between different endometrial conditions such as adenocarcinoma, hyperplasia, and polyps, and are also able to inform us regarding the grade of the cancer—low, intermediate, or high. In short, this study proves the transforming potential of combining image preprocessing with machine learning for early and accurate endometrial cancer prediction. Future studies need to focus on increasing dataset diversity, model interpretability, and interoperability with electronic health records. As these systems mature further, they would be quintessential tools in current gynaecological oncology, enabling patients with faster diagnosis, customized treatment plans, and improved health outcomes.

5.1 Future Enhancement

The future scope of this project includes:

- Expanding the dataset to include images from diverse populations and imaging devices to improve model robustness.
- Developing lightweight models for deployment in low-resource settings and mobile devices.
- Integrating multimodal data such as genetic information and patient history to enhance prediction accuracy.
- Collaborating with medical professionals for real-world validation and fine-tuning of the system.
- Exploring other deep learning architectures, such as Vision Transformers, for further performance improvement.

VI. ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my mentors and professors for their invaluable guidance and support throughout this project. Their insightful feedback and encouragement have been instrumental in refining my understanding of deepfake detection and deep learning techniques.

Additionally, I extend my appreciation to my peers and family for their constant motivation. Lastly, I acknowledge the researchers whose work laid the foundation for this study.

REFERENCES

[1]. E. Davis and R. Brown, “Histopathological Image Analysis for Endometrial Cancer Detection: A Review,” in International Journal of Biomedical Imaging, vol. 2019, Article ID 1294571, 2019.

[2]. S. Lee and D. Wilson, “Transfer Learning-Based Endometrial Cancer Detection in Histopathological Images,” in IEEE Access, vol. 9, pp. 139405–139414, 2021.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.141269

[3]. Goyal M., et al., Deep Learning for Grading Endometrial Cancer, 2024 (preprint / journal). Discusses EndoNet and transformer-enhanced pipelines for grade classification from H&E histology.
[4]. Jiang X., et al., Enhancing clinical decision-making in endometrial cancer using deep learning-based imaging techniques, 2024. Systematic review of imaging modalities and DL approaches, with sections on interpretability and clinical integration.
[5]. Wang C.W., et al., Deep learning for endometrial cancer subtyping and molecular status prediction from H&E, 2024. Explores MIL and multi-resolution DL frameworks for predicting subtypes and molecular features from WSIs.
[6]. Wang L., et al., Diagnosis Test Accuracy of Artificial Intelligence for Endometrial Cancer: Systematic Review, JMIR (2025). Meta-analysis of AI diagnostic studies, discusses heterogeneity and reporting standards.
[7]. Wang C.W., et al., Deep learning for endometrial cancer subtyping and molecular status prediction from H&E, 2024. Explores MIL and multi-resolution DL frameworks for predicting subtypes and molecular features from WSIs.
[8]. Wang L., et al., Diagnosis Test Accuracy of Artificial Intelligence for Endometrial Cancer: Systematic Review, JMIR (2025). Meta-analysis of AI diagnostic studies, discusses heterogeneity and reporting standards.
[9]. R. Anderson and J. Taylor, "Histopathological Image Analysis for Endometrial Cancer: Challenges and Future Directions," in Computers in Biology and Medicine, vol. 145, p. 105434, 2022.
[10]. S. R. Patil, "Detection of Nuclei Cell in Histopathological Images of Uterine Cancer: Adenocarcinoma of Endometrium," in Procedia Computer Science, vol. 167, pp. 2391–2399, 2020.
[11]. A. Dixit, "Cancer Detection using Image Processing and Machine Learning," in International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 5, pp. 2310–2315, June 2020.
[12]. Development and validation of clinical-radiomics deep learning model based on MRI for endometrial cancer molecular subtypes classification — Yue et al. (2025)
[13]. Vijayalaxmi. Mekali and H. A. Girijamma, "Fully Automatic Detection and Segmentation Approach for Juxta-Pleural Nodules From CT Images," in International Journal of Healthcare Information Systems and Informatics (IJHISI), vol. 16, no. 2, pp. 18, 2021.