Received: 8 November 2025 | Revised: 5 February 2026 | Accepted: 25 May 2026 | Published online: 22 June 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–16
DOI: 10.47852/bonviewJCCE62028130

BON VIEW PUBLISHING

HistoPath AI: An Edge-Powered Web Platform for Real-Time Lung Cancer Diagnosis

Rajkumar Maharaju[1,*], Sai Prasad Ellaboina[1] and Rama Valupadasu[1]

[1] *Department of Electronics and Communication Engineering, National Institute of Technology Warangal, India*

Abstract: Lung cancer remains a major cause of mortality worldwide, underscoring the importance of accurate and timely diagnosis. Manual evaluation of histopathology slides is labor-intensive and error-prone, particularly in resource-constrained settings. This work introduces HistoPath AI, an edge-based web application for real-time lung cancer detection. The system employs a fine-tuned EfficientNetB7 model to classify histopathology images into benign, adenocarcinoma, and squamous cell carcinoma, achieving 99.8% validation accuracy with strong precision and recall. A novel Computational Severity Index approach stratifies disease progression into four levels, supporting clinical decision-making. To ensure practical deployment, the model is optimized into FP32, FP16, and INT8 formats using TensorFlow Lite and integrated on a Raspberry Pi 4. The framework includes benchmarking of latency, resource utilization, and thermal stability, alongside a Flask-based role-driven interface for secure access, patient registration, and automated reporting. HistoPath AI demonstrates a scalable, privacy-preserving, and deployable solution for point-of-care cancer diagnostics. The proposed framework emphasizes offline inference and local data processing, enabling privacy-preserving operation without reliance on cloud infrastructure. The results highlight the feasibility of deploying deep learning-based histopathology analysis on low-cost embedded hardware for real-time use. This study establishes a strong engineering foundation for future extensions involving clinically informed validation and real-world deployment in resource-limited healthcare settings.

Keywords: histopathology, Computational Severity Index (CSI), adenocarcinoma, squamous cell carcinoma, Edge AI

1. Introduction

Cancer develops when normal cells multiply without control. In the United States, it is the second leading cause of death, and lung cancer is among the most commonly diagnosed types [1]. Major risk factors for lung cancer include smoking, exposure to secondhand smoke, and long-term contact with polluted air [2]. Survival rates for lung cancer vary widely. About 40% of patients live at least one year after diagnosis, but only around 10% survive 10 years, and the five-year survival rate is about 15%. The type of tumor and its stage at diagnosis have a significant impact on these outcomes [3].

To identify lung disorders, medical professionals employ diagnostic techniques like MRIs, CT scans, X-rays, and microscopic analysis of tissue samples. To confirm a diagnosis, these images are usually examined by hand, which can take several hours [4–6]. This procedure adds to the workload and may result in mistakes or conflicts in medical professionals' interpretations. According to recent research, deep learning can sometimes outperform skilled pathologists, particularly when it comes to determining whether cancer has spread [7]. Because they expedite image analysis and minimize mistakes or treatment delays, computer-aided diagnosis systems are growing in popularity. Artificial intelligence (AI) enables models to understand patterns from ground truth information instead of depending on existing data. Machine learning models learn specific tasks gradually to accurately classify the new data without constant human input. By applying signal and image processing to datasets, these models can be further enhanced, assisting in the identification of growth patterns, form variations, and peculiar pathological features. This approach leads to clearer prognoses, augments feature detection, and advances model training. Convolutional neural networks (CNNs), random forests, batch normalization, and support vector machines (SVMs) are some AI methods that have recently gained popularity, particularly for examining medical images like MRI, CT, and X-rays [8]. Although these methods produce good results, it is not possible to correctly distinguish between multiple kinds of lung cancer using only CT or MRI scans [9, 10]. These scans don't show what's going on inside the cells; instead, they primarily show larger tissue features like spots or tumors. Through the direct examination of tissues and cellular elements made possible by a biopsy, disease-related alterations can be better understood. Because whole tissues are examined and classified, frequently by hand, in research and clinical settings, histopathology is becoming more and more important [11]. This procedure supports identifying illnesses. Although it can be difficult to

***Corresponding author:** Rajkumar Maharaju, Department of Electronics and Communication Engineering, National Institute of Technology Warangal, India. Email: rm712110@student.nitw.ac.in

integrate all of this information, this strategy assists in detecting disease earlier and improves treatment.

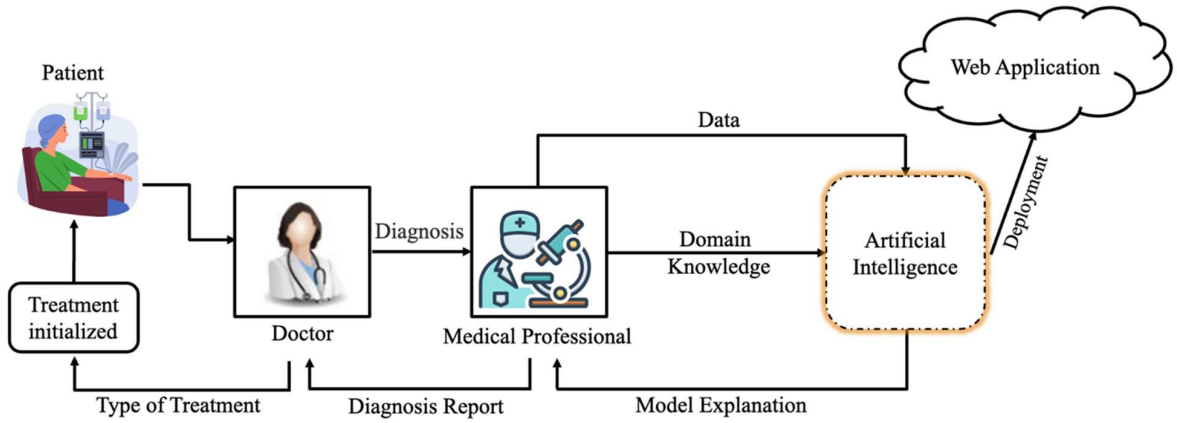
Nowadays, AI technology is revolutionizing healthcare by assisting doctors in making an accurate diagnosis without manual intervention with the help of deep learning methodologies. Figure 1 shows the designed smart diagnostic ecosystem, where the patients are subject to diagnostic investigations such as CT scan, MRI, biopsy, and laboratory testing by doctors who are examined by medical experts like radiologists or pathologists. The AI system can learn from patient data and domain knowledge to generate accurate, transparent diagnostic predictions for expert validation. Moreover, the deployment of the AI system is possible via a cloud-based web application.

In this context, Hatuwal and Thapa [12] applied the CNN on the LC25000 dataset, containing lung and colon histopathology images. Upon testing their model on lung cancer data, their experimental outcomes proved that the model had excellent capability with high levels of training and validation accuracy. At an accuracy level of 65.10%, Roy et al. [13] introduced the Reg-STN + SORD model for video prediction using lung ultrasound images obtained from Italian hospitals. Utilizing lung histopathology images, Šarić et al. [14] demonstrated an accuracy of 75% and 72% when using VGG and ResNet, respectively, with less accurate results due to the complexity of datasets and network parameters. In their case, Mehmood et al. [15] enhanced the accuracy of AlexNet to 98.8%, compared to its baseline accuracy of 89%, but the use of the old architecture made the model less robust. Handcrafted feature-based methodologies are available as well. Numerous research projects have been dependent on histopathology slide images, especially the LC25000 dataset. In particular, Mangal et al. [16] conducted their research on one type of cancer, such as lung or colon cancer classification. Masud et al. [17] proposed a lightweight CNN with an accuracy of 96.33%.

Gabralla et al. [18] used four CNN models (VGG16, ResNet50, InceptionV3, DenseNet121) with SVM, and their approach yielded 100% accuracy, although they did a single train-test split 70/30 with post-processing based on their test set, causing potential data leakage. Additionally, they developed a complicated system that was not end-to-end trainable. Kumar et al. [19] compared handcrafted features with Deep Neural Network (DNN) features by employing several types of classifiers, and they concluded that DNN outperforms the handcrafted ones in terms of accuracy, achieving 98.60% accuracy, 98.63% precision, an F1-score equal to 0.985, and an Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) value of 1. On the other hand, Chhillar and Singh [20] attained 100% accuracy using a LightGBM classifier with handcrafted features, but there may be some issues with their generalizability. Provath et al. [21] introduced a Global Context Attention block into their neural network, decreasing the number of parameters, reducing computations, and increasing model accuracy. They achieved 99.76% accuracy, 99.4% precision, and 99.6% sensitivity on LC25000, along with strong results on GLaS and CRAG datasets. The current study uses histopathological images from the LC25000 dataset to create an automatic model for classifying lung cancer. Dataset splitting, neural network selection, on-the-fly augmentation, training, validation, and final deployment on an edge device are all part of the suggested pipeline. Recent work has also extended deep learning to edge devices for medical diagnostics. In order to identify COVID-19 and lung-related conditions from chest X-ray images, Bhosale and Patnaik [22] introduced LDCNet, a compact CNN, and showed its efficacy with high accuracy on a Raspberry Pi platform. Alghareb and Hasan [23] investigated parallel execution on Raspberry Pi clusters and discovered that using architectures such as EfficientNet greatly accelerated inference speed while preserving high diagnostic accuracy for tasks pertaining to colon and brain cancer. Moreover, the work by Xu et al. [24] presented a detailed review related to edge-based AI for healthcare imaging applications and related hardware-embedded decisions on model optimization and practical applications. Collectively, the above-mentioned works clearly indicate the rising concern for edge-based deep learning solutions within the modern healthcare system framework.

The major accomplishments of the presented work are itemized below. First, a severity indexing approach for histopathology images is presented through the assessment of the processed images via signal processing analysis, extensive annotation of the entire dataset of images with severity levels varying from low to high, and so on. Second, a deep learning approach is designed for the classification task that incorporates the severity information, resulting in a lower computation cost and increased accuracy than the existing approaches. Third, the designed approach is optimized for effective execution on edge devices, with a particular emphasis placed on Raspberry Pi devices. Finally, a secure, user-friendly, and effective web application is designed, catering to the needs of various users, including clinicians, administration, patients, and other medical experts.

Figure 1
Functional flow diagram of the smart diagnostic ecosystem



1.1. Raspberry Pi (edge computing unit)

The Raspberry Pi Foundation created the Raspberry Pi, a tiny, reasonably priced, and energy-efficient single-board computer. Its small size and compatibility with Linux-based operating systems have made it a viable option for edge computing, especially in settings with constrained processing power. The platform is appropriate for integrating both software programs and external hardware components because it offers a variety of connectivity options, such as HDMI, USB, Ethernet, wireless communication, and GPIO interfaces. In order to support on-site diagnostic decision-making in clinical settings with limited computational infrastructure, the suggested deep learning model in this study is meticulously optimized and implemented on a Raspberry Pi to enable real-time histopathological image classification directly at the point of use. Figure 2 is the model diagram of Raspberry Pi.

The Raspberry Pi 4 Model B used in this project is powered by a quad-core Cortex-A72 processor running at 1.5 GHz. This delivers stable, efficient performance of edge-level calculations in machine learning applications. This model's system includes 4GB of LPDDR4 flash memory. This facilitates the execution of the quantized deep learning model without any memory-related issues. The entire circuit is powered by a 32GB microSD card that stores the operating system, machine models, and diagnostic information. The circuit also contains components like Wi-Fi, Bluetooth 5.0, Gigabit Ethernet, and several USB connections. It allows for multiple methods to connect in real-time for medical purposes. Its compact design and energy efficiency make it suitable for deploying lightweight AI systems in clinical settings.

2. Methodology

The pipeline of the proposed methodology included the acquisition of data and preprocessing to ensure compatibility with the architecture. Subsequently, several training techniques were employed, with adjustments to hyperparameters, to facilitate the model's learning process and minimize loss for effective classification. To further enhance classification accuracy. Following this, disease severity is indexed in a score ranging from 0 to 3 (low to high levels, respectively). Later deployment of this model into an edge device (Raspberry PI) to create a web interface to utilize in the diagnosis process by specialists. Figure 3 shows the end-to-end pipeline for AI-based lung cancer diagnosis.

2.1. Data source and data acquisition

The Lung and Colon Histopathology Images (LC25000) dataset, curated by Borkowski et al. [25], is publicly available and used in this study. The dataset contains 25,000 samples, with 15,000 related to lung cancer and 10,000 to colon cancer. The lung cancer group is divided into three types: benign, squamous cell carcinoma, and adenocarcinoma, each with 5000 samples. All images in the dataset are anonymized, comply with HIPAA regulations, have been thoroughly validated, and are freely accessible to AI researchers.

2.2. Data preprocessing and formatting

Originally, the dataset was in the JPG and RGB file format with a (728, 728) pixel size. Subsequently, they were scaled down to (224, 224) pixels to maintain a consistent aspect ratio. The pixel intensity in the image was then normalized between a value of 0 and 1 to promote a stable and quicker convergence phase in training this neural network model. Additionally, a series of image data augmentation operations were also performed, such as rotation of an image, zoom operations, cropping, inserting a patch, and both vertical and horizontal flipping. The aforesaid operations were undertaken to prevent the network from suffering from an overfitting problem and to increase the size of the dataset for training [26].

The dataset was partitioned into training and validation subsets prior to data augmentation, allocating 80% of the images for training and the remaining 20% for validation. All augmentation operations were applied exclusively to the training set to prevent information leakage into the validation data. It is noted that the LC25000 dataset does not provide patient-level

Figure 2
Raspberry Pi

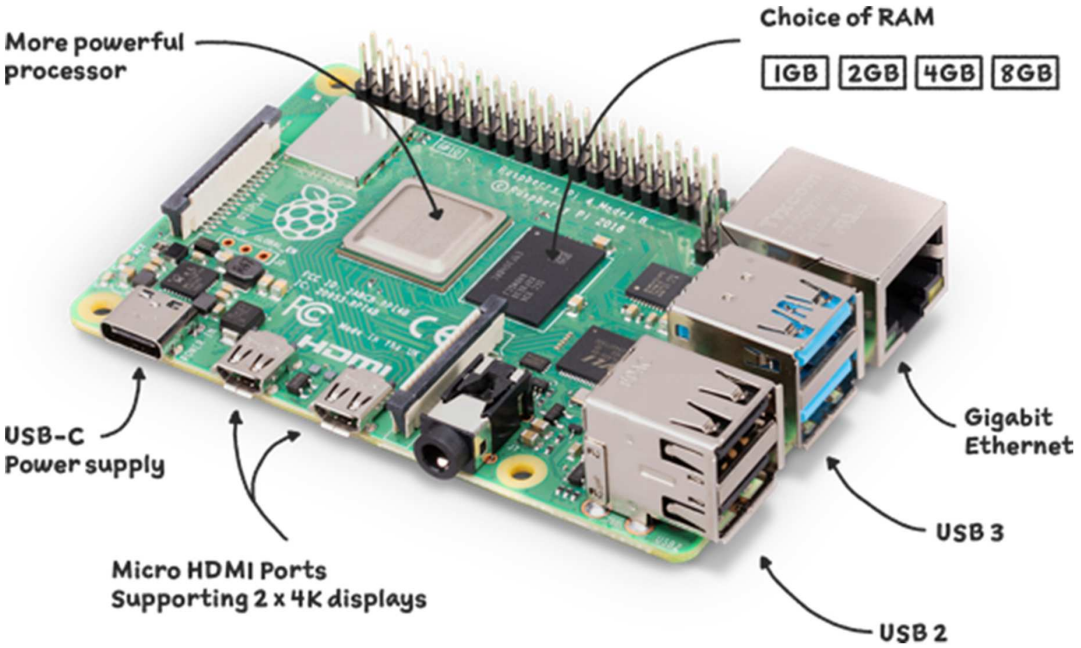


Figure 3
End-to-end pipeline for AI-based lung cancer diagnosis

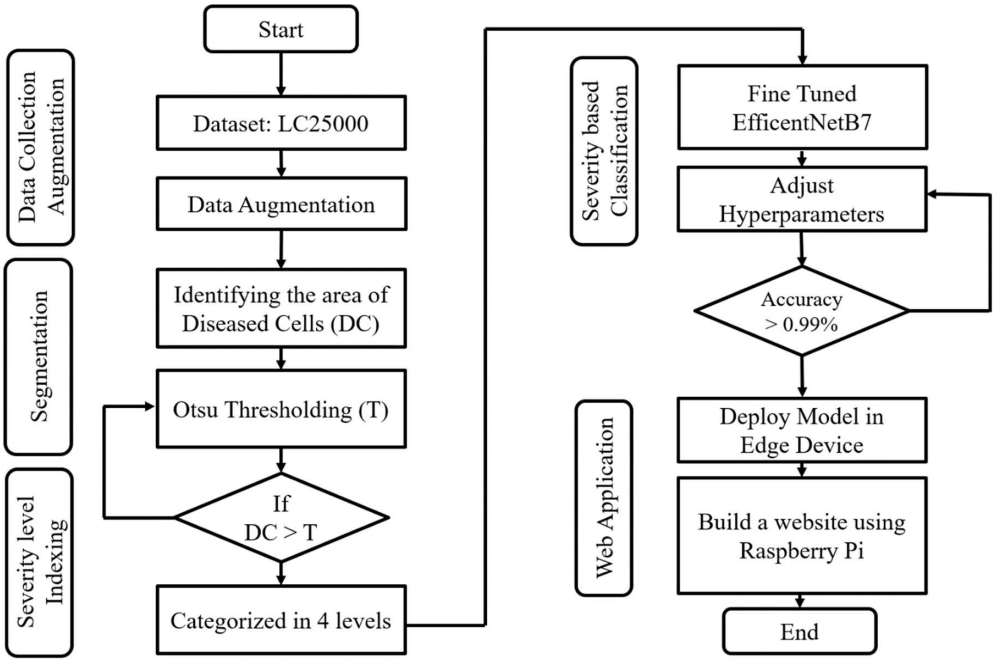
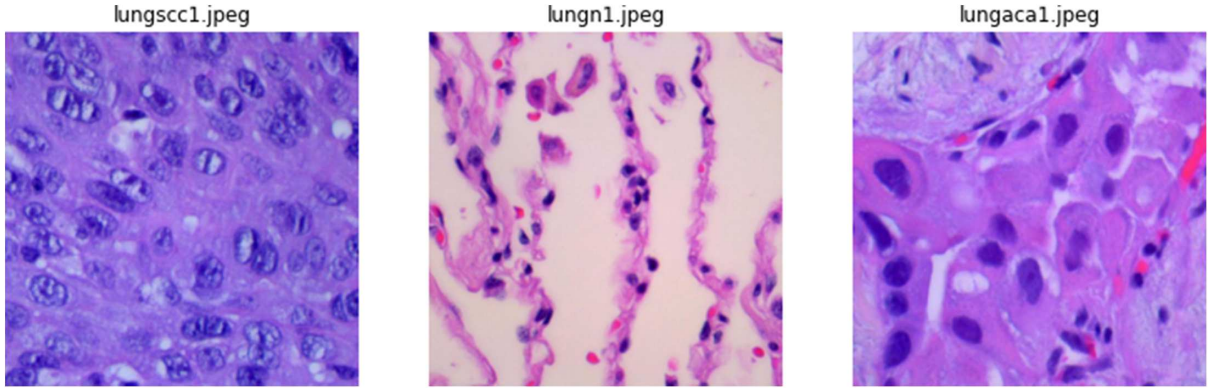


Figure 4
Three types of lung cancer images from the LC25000 dataset



identifiers; therefore, patient-wise splitting could not be enforced. This limitation is explicitly acknowledged, and the reported results are interpreted as engineering performance validation rather than clinical validation. Figure 4 displays sample lung histopathology images of adenocarcinoma, squamous cell carcinoma, and benign tissue.

2.3. Proposed method

2.3.1. Computational Severity Index

The Computational Severity Index (CSI) was introduced to identify the severity levels, which were given on a scale of 0–3, representing a risk range of low to high. This assignment, particularly important in cases of adenocarcinoma and squamous cell carcinoma, was based on the number of abnormal cells presented in the image and the area they occupied. Figure 5 provides an overview of the distribution of severity levels of disease across the dataset and displays a percentage of images assigned to each level of severity for each cancer class in the LC25000 dataset.

The severity index outcomes are demonstrated in Figure 6 for the known disease severity classes. The system comprises three steps, namely, (a) preprocessing, wherein image scaling, normalization, and conversion into grayscale are carried out; (b) segmentation (step I), wherein the background is discarded to extract tissue/cell regions that are color-coded to estimate cell sizes and the areas are calculated using the pixels; and (c) classification (step II), wherein the severity level of the disease is determined through thresholding methods into four classes denoted as 0 to 3 (low to high). The discussion of how segmentation works and the calculations performed on the pixels and threshold estimation will be presented in the next paragraph.

The different levels clearly show the strength of the disease in the patients, making it easier to prioritize the cases especially in emergencies. To differentiate between the cellular areas and the rest of the tissues in the background, the tissues are converted into grayscale before being analyzed using the Otsu thresholding method. After converting images to grayscale, Otsu thresholding is applied to segment abnormal regions, and the number

Figure 5
Severity level in each cancer type: (a) adenocarcinoma and (b) squamous cell carcinoma

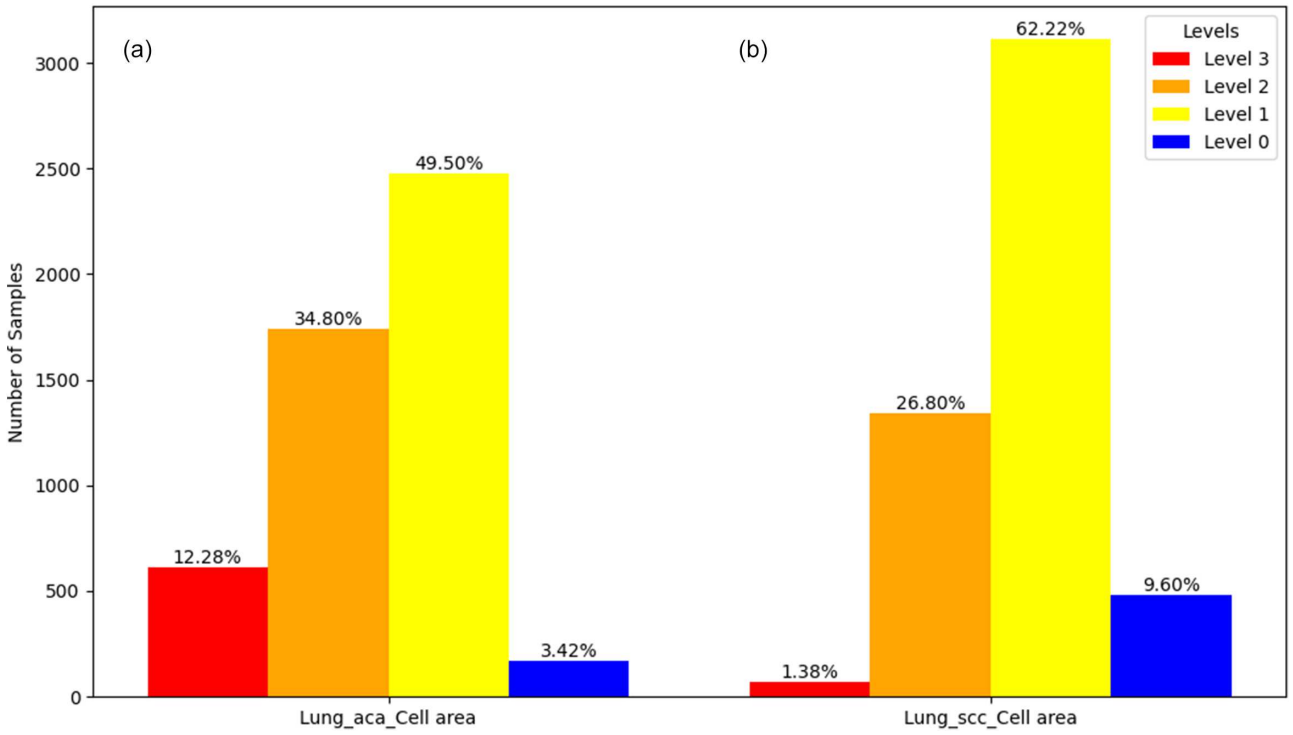
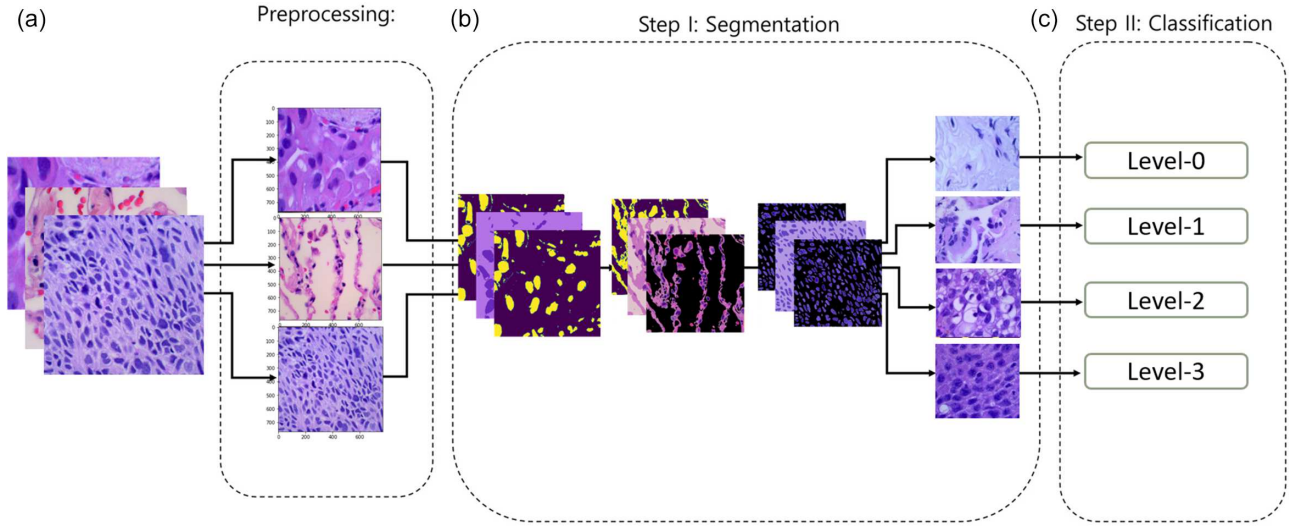


Figure 6
Corresponding lung histopathology images with sample severity index levels: (a) preprocessed images representing three types of lung cancers; (b) segmented images for adenocarcinoma, benign tissue, and squamous cell carcinoma, respectively; and (c) classification of disease severity score ranging from level 0 to level 3 (low to high)



of foreground pixels (Aw) representing abnormal tissue is computed. For samples predicted as malignant, CSI levels are assigned using fixed pixel-area thresholds: Level 0 (Aw < 100,000), Level 1 (100,000 ≤ Aw < 200,000), Level 2 (200,000 ≤ Aw < 300,000), and Level 3 (Aw ≥ 300,000). These empirically selected thresholds enable a consistent and reproducible mapping from segmented tissue extent to CSI levels, while benign samples are excluded from severity assignment.

The diseased cell regions are categorized into four severity levels using predefined threshold values based on the calculated pixel area occupancy. It is acknowledged that the proposed CSI, which relies on threshold-based segmentation and estimation of abnormal tissue area, may be influenced by factors such as staining variability, illumination differences, and slide preparation artifacts. To clearly distinguish this approach from clinically validated cancer severity or staging systems, the CSI is explicitly defined as an image-level, computational, triage-oriented indicator, rather than a biological or pathological measure of disease progression. Clinical severity assessment typically depends on complex histopathological and biological criteria beyond

pixel-area measurements; therefore, the absence of clinical validation is acknowledged as a limitation of the present study. Incorporating artifact-aware modeling and clinically informed validation is identified as an important direction for future work. For the entire LC25000 dataset, Figure 5 shows the percentage-wise distribution of images across various severity categories, with a separate breakdown for each type of cancer. A clear and thorough understanding of how disease severity is represented throughout the dataset is provided by this visualization.

2.3.2. Severity-based classification

Although many deep learning frameworks are computationally heavy, training on a large dataset such as ImageNet has been found to provide notable gains in accuracy and architectural design [27]. A typical example is the EfficientNet framework [28], which utilizes about 66 million parameters and scores an 84.4% classification rate on ImageNet. The architecture of the EfficientNet comprises CNN versions B0 to B7. As one progresses from B0 to B7, there is a notable improvement in accuracy, while there is an increase in parameters in a somewhat controlled manner. An important feature of EfficientNet is the use of the Leaky ReLU activation function rather than ReLU, which helps in enhancing learning capacity for negative-valued data points. In addition, EfficientNet differs from other models, particularly smaller versions, in its ability to perform compound scaling in which network width, network depth, and input resolution are simultaneously scaled. Figure 7 shows some of the fundamental building blocks of the EfficientNet framework.

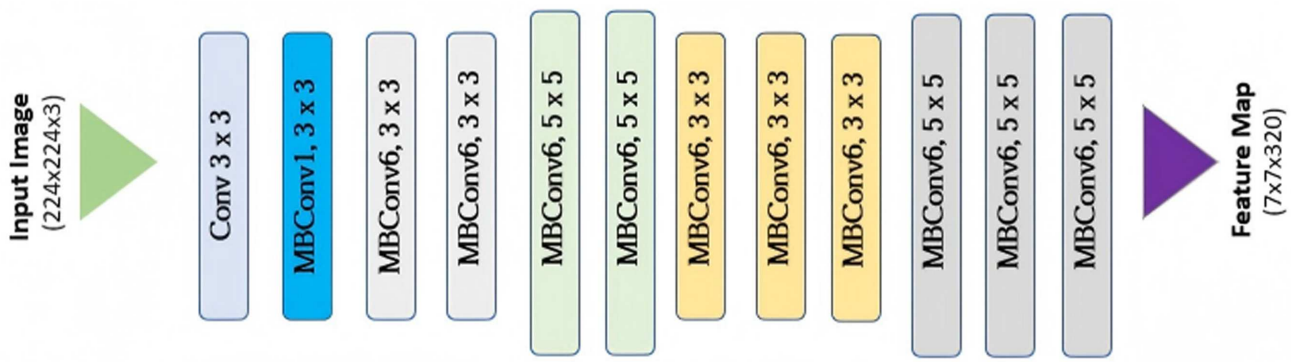
Based on histopathological imaging, this work presents a trustworthy severity-aware classification framework for lung cancer diagnosis [29]. The proposed approach categorizes the lung tissue into three severity levels: adenocarcinoma (lung_aca), squamous cell carcinoma (lung_scc), and normal tissues (lung_n). The high-resolution JPEG images are downloaded from a publicly available source, and all images are resized to $224 \times 224 \times 3$ pixels and then organized systematically into class-based folders. The images are fed into an Image Data Generator with an 80:20 split for training and validation purposes, along with categorical classification [30]. It is built upon EfficientNetB7, an efficient deep learning CNN architecture that is widely acclaimed for its compound scaling technique. This is a novel technique whereby the depth, width, and spatial resolution are all scaled simultaneously to obtain high accuracy and reduced computational requirements. To allow efficient learning and retain trustworthy and universal attributes, EfficientNetB7 was used as a frozen feature extractor. It was first pretrained on the ImageNet dataset. The better version of the proposed approach was developed by selectively using the higher severity images (Levels 2 and 3) and utilizing a total of 2500 balanced images for selective training on each type of cancer. By successfully reducing the training time and computational requirements, it was possible to successfully apply and run the model on resource-constrained Edge platforms.

The extracted features were analyzed using a custom classification head containing a global average pooling layer, flattening, and two fully connected layers with 128 and 64 neurons, respectively, activated by Leaky ReLU. A SoftMax output layer for multi-class prediction came next. A sigmoid activation was employed in a different implementation designed for binary scenarios. The model was trained for up to 50 epochs with a batch size of 128 using the Adam optimizer and categorical cross-entropy loss. To prevent overfitting, early stopping was used. To enhance generalization, the architecture design included batch normalization, Leaky ReLU activations, depth-wise separable convolutions with filters of (3×3) and (5×5), and a fixed dropout rate of 0.2. Each of the five modular subblocks was linked sequentially and scaled using a two-fold multiplication technique. The network is scaled as follows to increase the computational task by a factor of "n": depth = αn, width = βn, and size = γn, where α, β, and γ are constants. To guarantee uniform scaling across all network dimensions, EfficientNet additionally integrates a factor "ϕ," also referred to as the compound coefficient. A classification report and a normalized confusion matrix were used to further assess the model's performance, which showed excellent accuracy, precision, recall, and F1-scores in every category. Ultimately, 99.02% validation accuracy was attained by the model. 99.02%.

To discuss the hyperparameters, the model is based on the EfficientNetB7 architecture with a total of 64,433,946 parameters, of which 336,259 are trainable, and the remaining are retained as non-trainable pretrained weights. The network is optimized using the Adam optimizer with categorical cross-entropy loss and default TensorFlow learning rate settings. Training is conducted for up to 50 epochs with early stopping monitored on validation loss and a patience of 3 epochs to ensure stable convergence and mitigate overfitting.

To ensure compatibility with clinical deployment workflow and edge inference, the trained network is exported in multiple formats including HDF5, TensorFlow Saved Model, and JSON formats. An efficient inference pipeline comprising preprocessing specific to EfficientNet, followed by model-based predictions [31], and mapping of predicted results to predefined classes was adopted to generate predictions on test images. The proposed

Figure 7
Basic blocks in EfficientNet architecture



approach has proved to exhibit significant benefits over other existing models such as Baseline Network, InceptionNetV2, ResNet50, DenseNet, classical CNN, and BreastNet [32] regarding classification performance and computational complexity. The proposed system provides an effective, efficient, and understandable approach for lung cancer detection in clinical environments and situations with limited computational resources through the integration of severity-oriented domain knowledge with a deep learning architecture.

3. Real-Time Monitoring

This workflow describes a comprehensive edge-deployed medical diagnosis system, specifically designed to operate on resource-constrained devices like the Raspberry Pi. The system combines a lightweight yet powerful pretrained EfficientNetB7 model for lung histopathology classification with a Flask-based web application that provides an accessible user interface for end users such as clinicians or lab technicians.

3.1. Boot and setup

The boot and setup phase is responsible for initializing the device and configuring the system environment required to load the classification model and start the Flask-based application. The Linux-based Raspberry Pi OS and all necessary system services are loaded into memory when the Raspberry Pi is turned on, activating the hardware components. In this phase, an automated batch script set up with rc.local or systemd begins the application environment automatically at startup. To guarantee proper model execution, a specialized Python virtual environment is then turned on, preinstalled with all required libraries like EfficientNetB7 and TensorFlow Lite. The edge device can start automatically thanks to this initialization sequence, and it will be ready to perform inference as soon as it boots up.

3.2. Networking

In order for a web browser to access the diagnostic application, the networking module is responsible for establishing self-hosted or local connectivity. A local web server that runs directly on the Raspberry Pi is started by the Flask service. DNSmasq and hostapd are programs that are used to either assign an IP address that remains static or enable a device-hosted Wi-Fi hotspot. By enabling the application to be accessed through the local IP address, which is normally accessible at the local host link[1]. This removes the need for external cloud infrastructure. Because of its design, the system is perfect for use in remote clinics or rural healthcare settings where connection to the internet may be weak or unreliable.

3.3. Role-based interaction

Role-based interactions make sure that the right people have access and the right actions are taken with respect to access control and proper handling of confidential information, since the permissions are defined based on responsibilities. Doctors, technicians, and other healthcare professionals can log into the platform through any Internet-capable device from a Web browser. The permissions are provided by the system based on the below-listed set of roles such as administrator, doctor, technician, and medical staff, among others. Credentials for each user are checked during the login process, and then the functions available to a particular user based on his/her role are listed to him/her. Access to the upload of chest X-ray images and patient information is granted to authorized individuals only.

3.4. On-device AI inference (EfficientNetB7)

The primary AI inference model, which runs solely on a Raspberry Pi without a live internet connection, supports real-time disease prediction. The quantized or optimized version of the EfficientNetB7 model, which has been trained to distinguish between all lung-related illnesses such as COVID-19 and other variants, pneumonia, and tuberculosis, is loaded into it. The model provides an output after evaluating a supplied medical image, which includes a predicted disease type and a quantifiable level of confidence. Finally, without needing manual intervention, a PDF report containing information on prediction outcomes, patient identity information, and a timestamp is automatically created. The generated report and initial image are saved on this device. The report and initial image are saved on this device, which provides functionality for future clinical audits. The device is ideal for point-of-care use in disease prediction and diagnosis. The device enables quick decision-making on-site without needing a live internet connection and performs all calculations on-site.

3.5. Report output

The final stage of the system handles user sessions, including the presentation of diagnostic results and returning to a ready state. Once the model has completed inference, the client displays to the user the expected output along with providing an option to download the generated PDF report for further review or documentation. Following the results review, the user will securely log out to ensure patient data security. The application based on FLASK framework, operates in the background for front-end and is ready to launch a new session for any new login requests or images uploaded for further examination, while the edge device switches back to idle mode after the user logs out.

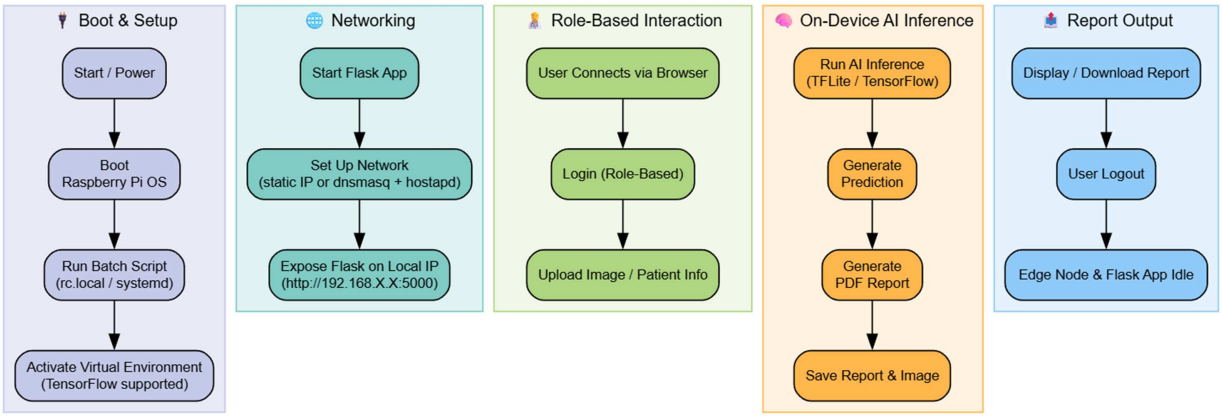
This streamlined process guarantees continuous system availability, supports multiple sequential diagnoses, and upholds the privacy and security of user interactions. Figure 8 explains the step-by-step Edge-based Raspberry PI Flask Inference Pipeline. We developed the Histopath AI Web Application, an edge-based inference system running on Raspberry Pi with Flask, for automatic prediction of lung cancer type and severity from histopathology images. The software has role-based access control. Doctors can see patient reports, patients can download their reports, and healthcare personnel can enroll patients and make appointments for them. System management and maintenance will be handled by an admin. It guarantees privacy of information, effectiveness, and real-time diagnosis, among other factors. System architecture and processes will be discussed in detail below.

3.6. HistoPath AI: A real-time edge-based web diagnostic system

Histopath AI is a web-based application using a Raspberry PI. The application is in place, and the healthcare system's security and management framework, known as role-based access, regulates user permissions according to their assigned roles within the company. Every user is given a specific role that dictates what information they can access and what they can do on the system, such as medical staff, diagnostic expert, specialist doctor, or admin. For example, medical staff are permitted to manage

[1] http://192.168.x.x:5000

Figure 8
Edge-based Raspberry Pi Flask Inference Pipeline



patient registration and schedule appointments, but they are restricted from accessing confidential diagnostic results. While specialized doctors are given access to prediction outcomes and patient medical histories so they can make clinical suggestions, diagnostic experts are responsible for submitting examination reports and highlighting abnormal observations. The administrator has the highest level of access and is able to monitor the system's performance, assign and handle user roles, oversee the entire system, and apply software updates.

Figure 9 shows the HistoPath AI application with role-based access along with tasks. This structured approach ensures data privacy, enhances security, and streamlines clinical workflows by ensuring that users only interact with the functions relevant to their responsibilities. The entire setup is connected with the monitor and the application dashboard, and the frontend display is shown in Figure 10, and the Raspberry Pi 4 Model B setup is shown in Figure 11.

4. Evaluation Metrics

4.1. Confusion matrix

For organizing the classifier's outputs in four ways: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the confusion matrix becomes the major framework for analysis. This table indicates the effectiveness of the classifier in separating the lung adenocarcinoma, squamous cell carcinoma, and benign tumors. Through analyzing these values, it is possible to get an idea about the strengths and weaknesses of the classifier model, including its ability to detect the presence of malignant areas and, vice versa, failure to detect these regions or labeling of normal tissues as tumors.

4.2. Accuracy

Accuracy means the proportion of correct predictions to all cases analyzed by the model. The proposed model proved to have an impressive prediction rate in all classes in our tests. The level of accuracy was 99% for squamous cell carcinoma and 100% for benign and adenocarcinoma classes. Calculation of the accuracy rate is

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad (1)$$

4.3. Recall/sensitivity

Recall or sensitivity indicates how well the model identifies all the positive examples. The higher the recall rate, the fewer the false negative errors; hence, the system makes very few mistakes in

Figure 9
HistoPath AI application with role-based access and tasks

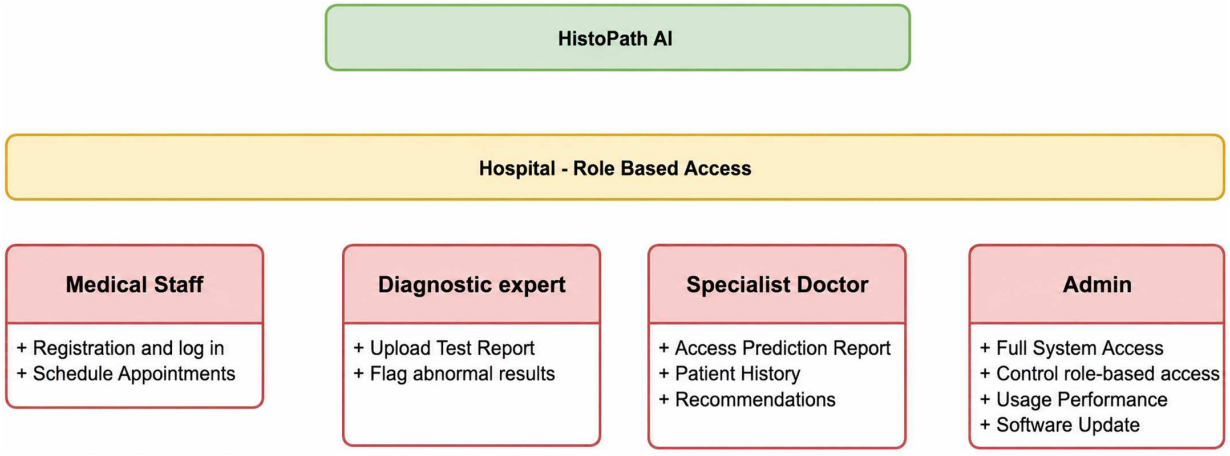
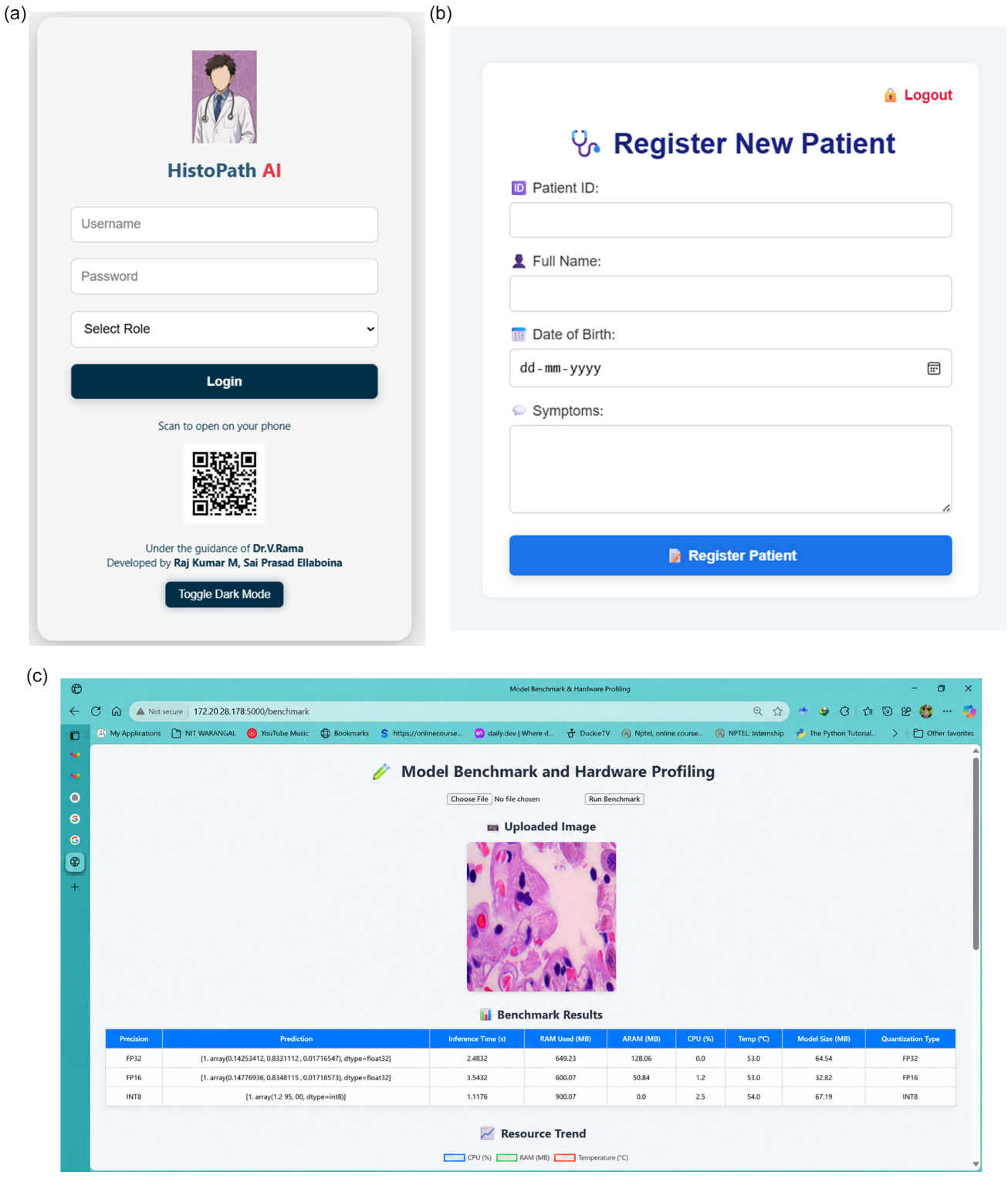


Figure 10
(a) HistoPath AI dashboard, (b) registration tab, and (c) model benchmark and hardware profiling, respectively



not recognizing positive examples. Our enhanced model exhibited exceptional sensitivity when tested: the recall of the system reached 100% for benign samples and 99% for squamous cell carcinoma and adenocarcinoma samples.

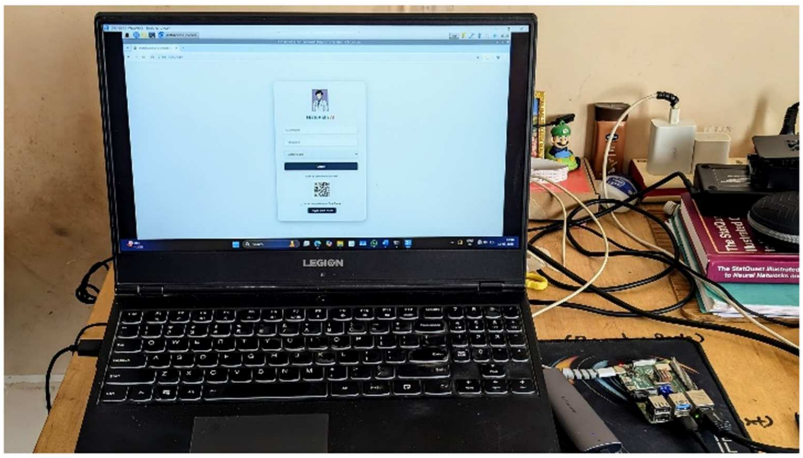
$$Recall = \frac{TP}{(TP + FN)} \quad (2)$$

4.4. Precision

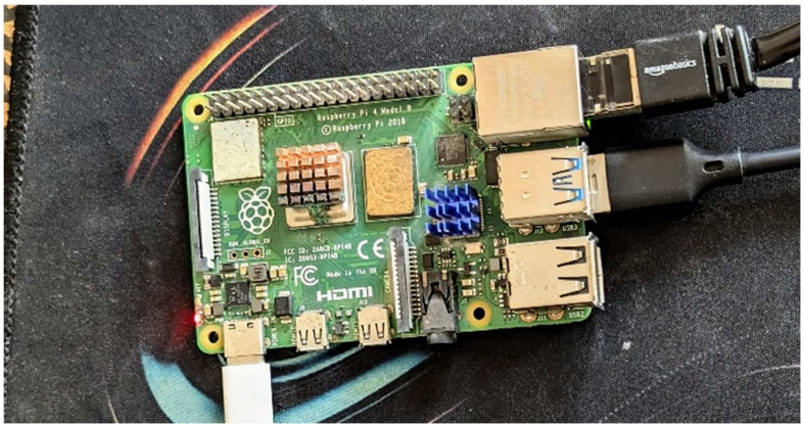
Since precision reflects the accuracy of predicted positive instances, it becomes a crucial factor especially in situations where it becomes necessary to avoid warnings from being raised. According to the experimental findings, precision was very high for all the classes, and it was achieved at 100% level for benign tumors,

Figure 11
Raspberry Pi 4 Model B setup

(a)



(b)



while it was 99% accurate for squamous cell carcinomas and adenocarcinomas. It is therefore evident that the non-tumor area was never incorrectly identified as a tumor region.

$$Precision = \frac{TP}{(TP + FP)} \tag{3}$$

4.5. F1-score

It measures the harmonic mean of the precision rate and recall rate and helps to find a balanced view of the efficacy of the classifier for datasets with unequal class distribution or where both the effects of false positives and negatives have a medical impact. The F1-scores of 100% for benign tissue and 99% for both adenocarcinoma and squamous cell carcinoma indicate that the model had been performing very well in all respects that concerned our evaluation.

$$F1score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4}$$

4.6. Interfacing time (latency)

Inferencing time, also known as inference computation, is the time a deep neural network requires to work on new data, like the images from the biopsy performed on the subjects. There are myriad reasons why a deep neural network will have a high inferencing time, some of which are the complexity, the number of layers, and the number of neurons within each layer.

4.7. Peak RAM usage

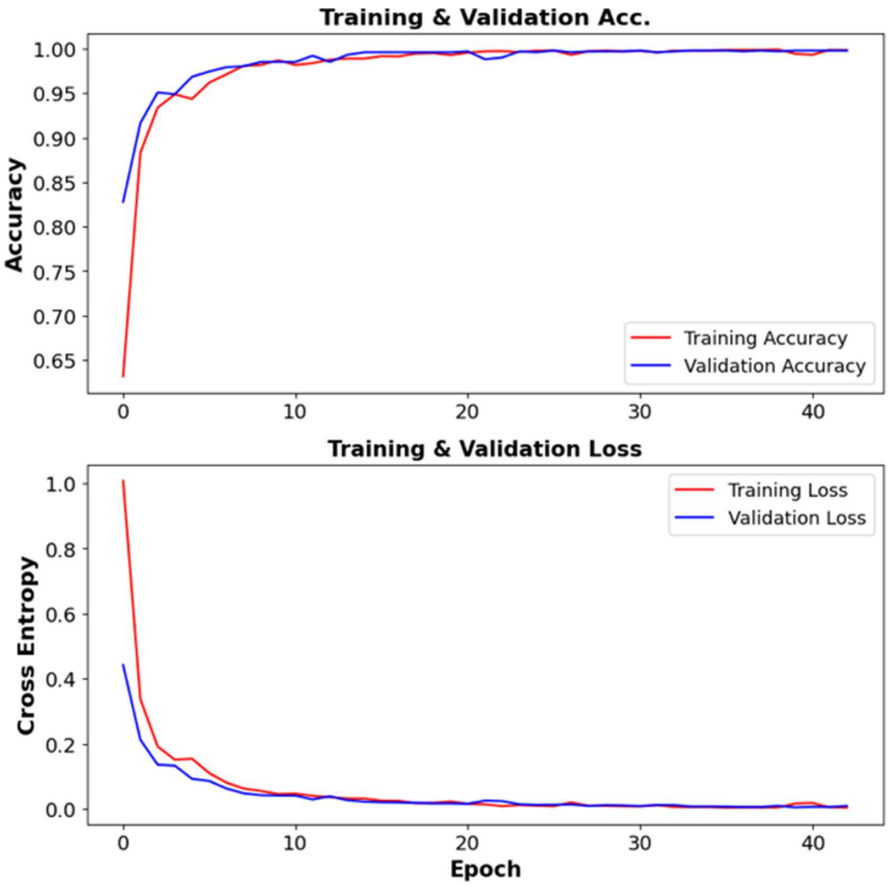
The scratch memory needed to carry out intermediate computations during runtime primarily determines how much RAM is needed to run neural network models. Flexible operation execution within the network architecture is made possible by this allocated memory. On the other hand, the storage of learned model weights and any random values produced during model initialization or configuration control peak ROM consumption. Hardware-level optimizations, which are essential for speeding up mathematical operations and are accessed whenever particular computational routines are called, are typically stored in ROM.

5. Results and Discussion

The images corresponding to the identified growth patterns were first augmented and subsequently used as input for training the network. A fixed batch size of 128 was employed, and the model was trained over 35 epochs. After training, the model performed very well, attaining accuracies of 99.85% and 99.9% in the process of validation. Figure 12 shows the plot of accuracy versus epoch and loss versus epoch for the training and validation datasets.

The performance of the model for classifying adenocarcinoma, benign tissue, and squamous cell carcinoma is illustrated

Figure 12
Plot for model accuracy and loss vs epochs images, respectively



in Table 1. The model performed exceptionally well, having scored 99% on precision, recall, and F1-score for adenocarcinoma and squamous cell carcinoma. However, the performance of benign tissue was perfect since it had achieved 100% on all three measures. In addition, Table 2 presents the corresponding class-wise details along with macro-averaged, weighted-average, and overall accuracy values.

Table 1
Recall, precision, and F1-score of model for different categories

	Performance metrics		
Cancer type	Precision	Recall	F1-score
Adenocarcinoma	99	99	99
Benign tissue	100	100	100
Squamous cell carcinoma	99	99	99

Table 2
Recall, precision, and F1-score of model for different categories

	Performance metrics		
Labeled category	Precision	Recall	F1-score
Accuracy	–	–	99
Macro average	99	99	99
Weighted average	99	99	99

Figure 13 is the illustration of the confusion matrix where the visualization of the comparison between the true label and the predicted output of the test dataset is shown. The mathematical equations used for the calculation of these metrics are explained in Section 4.

Table 3 provides an overall view of the evaluation metrics in comparison to existing research on the same dataset. The metrics include parameters like recall, F1-score, precision, and accuracy for different cancer-type categories. In this comparison, the proposed model demonstrates superior performance compared to existing models. Later, the proposed model from EfficientNetB7 for the lung classification was converted into TensorFlow Lite versions in three precision models, such as INT8, FP16, and FP32, for compatibility purposes on edge devices like Raspberry Pi.

Then the analysis compares t-distributed Stochastic Neighbor Embedding (t-SNE) visualizations of feature embeddings across three precision formats—INT8, FP16, and FP32—using 120 lung image samples. Figure 14 shows the t-SNE Feature maps for three TensorFlow Lite models (INT8, FP32, FP16). Although INT8 provides efficiency for edge devices, its reduced feature detail results in limited class separation. FP16 is appropriate for low-power devices because it enhances class distinction while balancing resource usage. For high-resource clinical settings requiring the highest level of diagnostic precision, FP32 offers the most precise and well-separated feature clusters.

The EfficientNetB7 model's performance under INT8, FP16, and FP32 precision settings is examined in this assessment. The INT8 variant is especially well-suited for real-time execution on

Figure 13
True label vs predicted label in the confusion matrix for different image categories for validation images

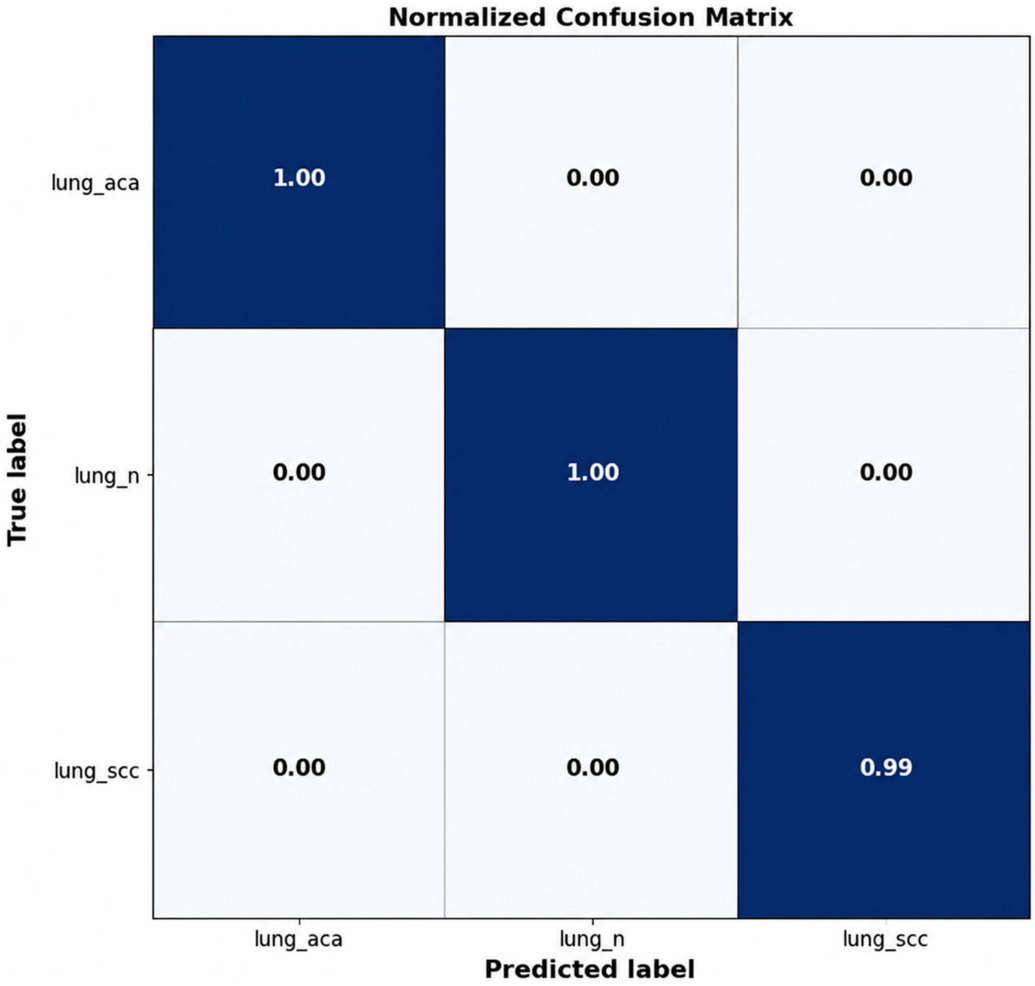


Table 3
Comparison of performance metrics with other network architectures

	Performance metrics			
Author	Precision	Recall	F1- score	Accuracy
Hatuwal and Thapa [12]	0.96	0.95	0.96	0.96
Mehmood et al. [15]	0.97	0.94	0.97	0.98
Mangal et al. [16]	–	–	–	0.96
Masud et al. [17]	0.96	0.96	0.96	0.96
Proposed model	0.99	1.00	0.99	0.99

edge devices, with the lowest inference latency of 1.17 s among the three. Remarkably, the INT8 model also consumed the most RAM (1927.53 MB), suggesting that lower runtime memory usage is not always the outcome of quantization. Unlike the FP16 model, which experienced a discernible memory spike of 187.27 MB, its memory usage was comparatively constant, with only minor fluctuations (ΔRAM = 6.38 MB). As low-precision computations require increased parallelism, the CPU usage was at its peak for INT8 precision configuration at 3.8%. Apparently, the quantization process does not add up to the thermal burden since all precision levels were kept in temperature between 47 and 48 degrees Celsius with a similar energy response. Further, compared to FP16 precision, an estimated 47% of storage memory savings was recorded when using INT8. The overall benchmarking results for the three TensorFlow Lite configurations are compiled in Table 4 and shown the same in bar graphs individually in Figure 15.

The relationship between inference time and RAM usage is shown in Figure 16. With the shortest inference time of 1.17 s and very little variation in memory consumption (ΔRAM = 6.38 MB), the INT8 variant performs the most evenly among the models that were evaluated. It is especially well-suited for deployment

Figure 14
t-SNE feature maps for three TensorFlow Lite models: (a) INT8, (b) FP32, and (c) FP16

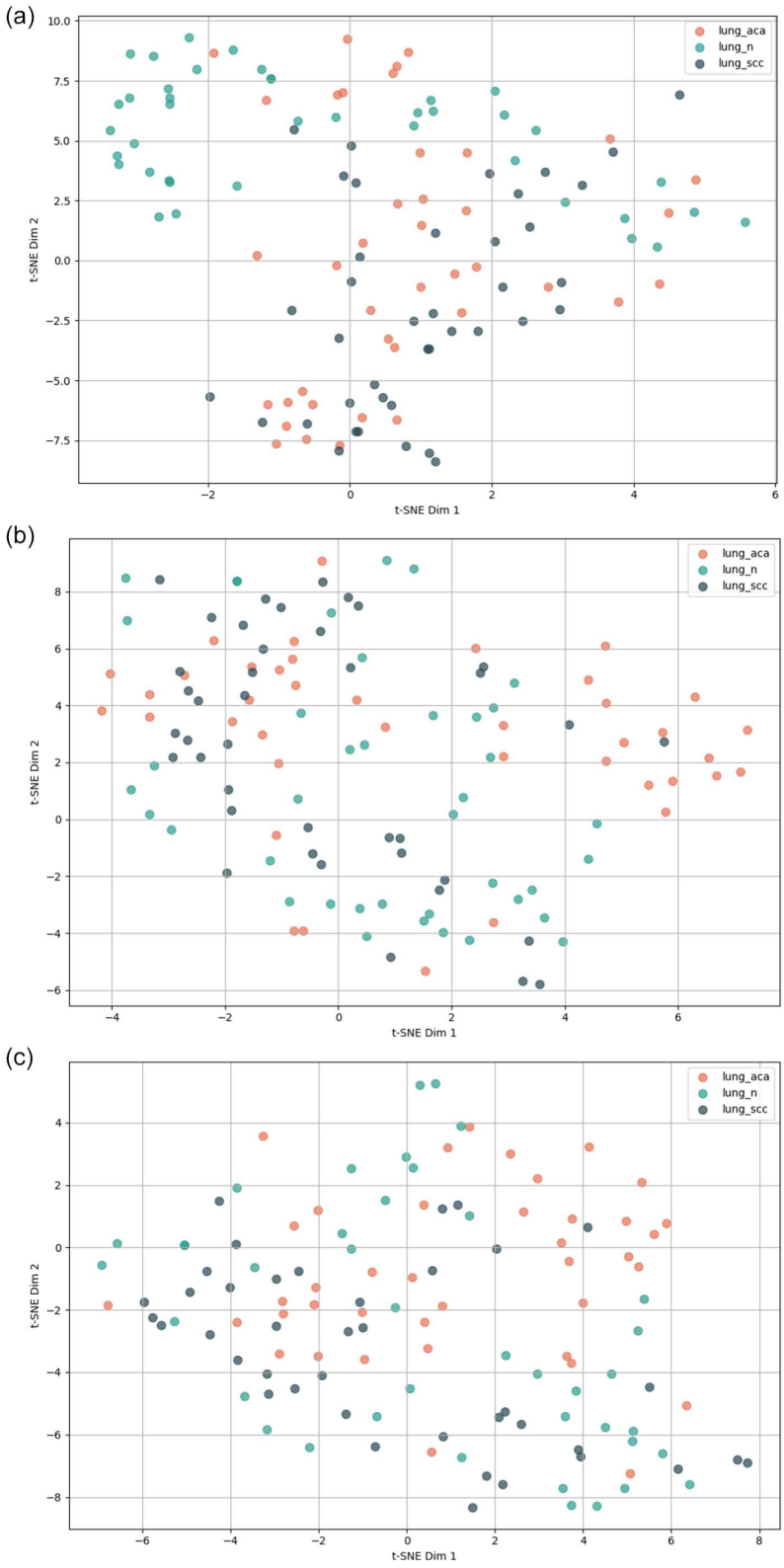
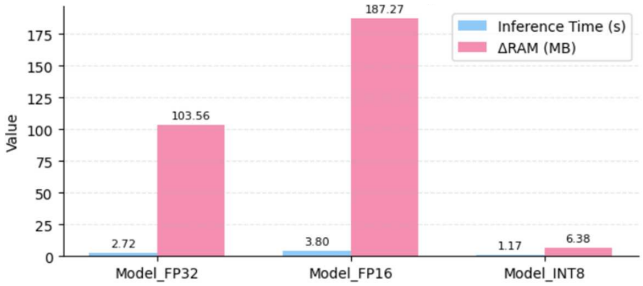


Table 4
Overall benchmark comparisons across three TensorFlow Lite versions

Model	FP32	FP16	INT8
Inference time (s)	2.7212	3.803	1.1745
RAM (MB)	1733.89	1921.15	1927.53
DRAM (MB)	103.56	187.27	6.38
CPU (%)	0	1.2	3.8
Temp(°C)	47.2	48.2	47.7
Model size (MB)	64.54	122.32	67.19

Figure 15
Various hardware performance parameters vs TensorFlow Lite versions

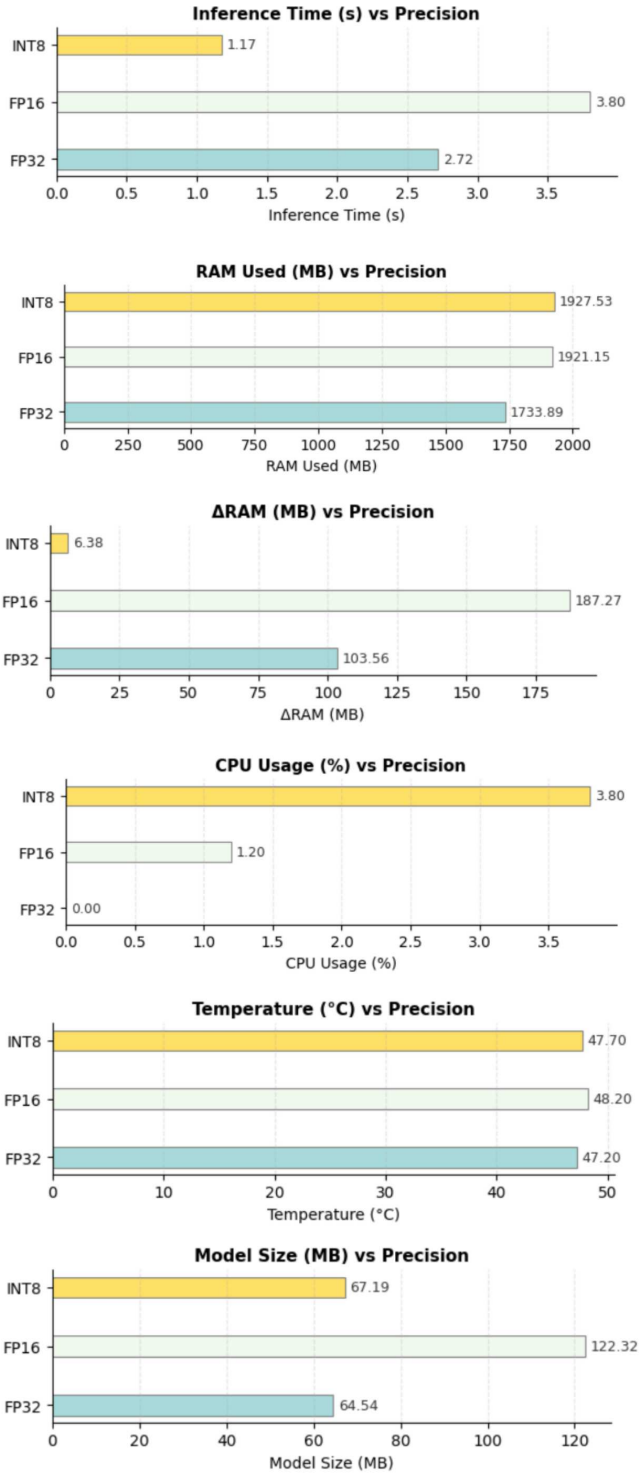


on systems with constrained memory resources because of these features.

6. Limitations and Future Work

Although the suggested system allows for effective, real-time lung histopathology classification on embedded hardware, it has a number of drawbacks that should direct further investigation. First, the dataset scope is limited to three lung tissue categories, warranting expansion to cover more histopathological classes and rare subtypes for broader clinical relevance. Second, the lack of explainability tools like Grad-CAM or attention heatmaps limits diagnostic transparency, highlighting the need for explainable AI integration. Third, the system lacks integration with Electronic Health Records, suggesting future support for standards like HL7/FHIR with secure data synchronization. Moreover, as the present pipeline relies on just one image in one inference cycle, the integration of batch processing or autonomous processing would prove immensely scalable. Although general performance analysis is done, the absence of power consumption analysis in the present system proves important, especially in the case of batteries or areas with reduced accessibility. The platform also needs better security functionalities, including support for HTTPS communication, JSON Web Token (JWT)-based authentication, and overall audit logging, in order to satisfy the requirements of overall clinical compliance. The absence of general clinical validation as well as usability studies proves the requirement of an association with healthcare setups in order to validate the overall diagnosis accuracy and ensure a seamless integration with the general healthcare setup. The overall transition of the platform toward a reliable and usable AI-based diagnostic solution in the healthcare domain would be facilitated by addressing the above-mentioned issues.

Figure 16
RAM vs inference time



7. Conclusion

This work introduces HistoPath AI, an edge-enabled diagnostic framework for the classification of lung cancer on histopathology images, considering technological feasibility and deployment in real time in resource-constrained environments. The proposed approach integrates an EfficientNetB7-based deep learning model with FP32, FP16 & INT8 quantization techniques

for optimization, and deployed on Raspberry Pi 4 Model B. The experimental results, which demonstrate good classification performance across the classes of benign tissue, adenocarcinoma, and squamous cell carcinoma, are corroborated by the confusion matrix and evaluation metrics. From an engineering perspective, the benchmarking analysis confirms that INT8 and FP16 quantization greatly reduce inference latency, memory usage, and thermal load, validating the proposed system's deployment readiness for edge-based operation.

In addition, a CSI derived from image-level signal processing is presented in the study to quantify the spatial extent of tissue abnormalities regions. The CSI is intended as a triage-based computational indicator to support process assistance and prioritization rather than as a clinically validated measure of disease severity. Although the reported performance metrics demonstrate strong computational capability and engineering robustness, claims of clinical reliability are currently precluded due to the dataset's lack of patient-level identifiers and the resulting challenges in implementing patient-wise data splitting. Consequently, the findings are interpreted as engineering performance validation, and any potential impact on patient outcomes remains hypothetical. Clinical applicability and outcome-level benefits will require future patient-level evaluation and formal clinical studies.

In summary, this study provides an excellent technical basis for the design of an AI diagnostic system in real time while maintaining privacy and being edge-centric. The areas that future research will address include clinically based evaluation of the solution, severity prediction models considering artifacts, inclusion of explainable AI concepts, and hardware acceleration through Field-Programmable Gate Array (FPGA)-based solutions like Xilinx Kria.

Ethical Statement

The lung cancer imaging data analyzed in this study were sourced from a publicly available dataset (Borkowski et al. [25]). The dataset contains anonymized data and is openly accessible. Therefore, ethical approval and informed consent were not required for this study.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available at https://doi.org/10.48550/arXiv.1912.12142, reference number [25].

Author Contribution Statement

Rajkumar Maharaju: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Sai Prasad Ellaboina:** Conceptualization, Methodology, Software, Validation, Resources. **Rama Valupadasu:** Conceptualization, Writing – review & editing, Supervision.

References

[1] American Cancer Society. (n.d). *What causes lung cancer?* https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/what-causes.html

[2] Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *74*(3), 229–263. https://doi.org/10.3322/caac.21834

[3] Redlich, J.-P., Feuerhake, F., Weis, J., Schaadt, N. S., Teuber-Hanselmann, S., Buck, C., . . . , & Homeyer, A. (2024). Applications of artificial intelligence in the analysis of histopathology images of gliomas: A review. *npj Imaging*, *2*(1), 16. https://doi.org/10.1038/s44303-024-00020-8

[4] Detterbeck, F. C., Boffa, D. J., Kim, A. W., & Tanoue, L. T. (2017). The eighth edition lung cancer stage classification. *Chest*, *151*(1), 193–203. https://doi.org/10.1016/j.chest.2016.10.010

[5] Tsao, M. S., Rosenthal, A., Nicholson, A. G., Detterbeck, F., Eberhardt, W. E. E., Lievens, Y., . . . , & Asamura, H. (2025). The International Association for the Study of Lung Cancer Staging Project: The database and proposal for the revision of the staging of pulmonary neuroendocrine carcinoma in the forthcoming ninth edition of the TNM classification for lung cancer. *Journal of Thoracic Oncology*, *20*(7), 856–870. https://doi.org/10.1016/j.jtho.2025.01.013

[6] Kish, K. (2026). Biology and treatment options for non-small cell lung cancer. In J. T. Fathi & M. F. Mortman (Eds.), *Lung cancer navigation and care: A comprehensive guide for navigators and allied health professionals* (pp. 189–204). Springer. https://doi.org/10.1007/978-3-032-02200-4_15

[7] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., . . . , & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, *25*(8), 1301–1309. https://doi.org/10.1038/s41591-019-0508-1

[8] Ristanoski, G., Emery, J., Gutierrez, J. M., McCarthy, D., & Aickelin, U. (2022). AI based cancer detection models using primary care datasets. *Journal of Advances in Information Technology*, *13*(2), 192–197. https://doi.org/10.12720/jait.13.2.192-197

[9] Shamshirband, S., Fathi, M., Dehzangi, A., Chronopoulos, A. T., & Alinejad-Rokny, H. (2021). A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, *113*, 103627. https://doi.org/10.1016/j.jbi.2020.103627

[10] Dritsas, E., & Trigka, M. (2022). Lung cancer risk prediction with machine learning models. *Big Data and Cognitive Computing*, *6*(4), 139. https://doi.org/10.3390/bdcc6040139

[11] Wei, J. W., Tafe, L. J., Linnik, Y. A., Vaickus, L. J., Tomita, N., & Hassanpour, S. (2019). Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports*, *9*(1), 3358. https://doi.org/10.1038/s41598-019-40041-7

[12] Hatuwal, B. K., & Thapa, H. C. (2020). Lung cancer detection using convolutional neural network on histopathological images. *International Journal of Computer Trends and Technology*, *68*(10), 21–24. https://doi.org/10.14445/22312803/%20IJCTT-V68I10P104

[13] Roy, S., Menapace, W., Oei, S., Luijten, B., Fini, E., Saltori, C., . . . , & Demi, L. (2020). Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. *IEEE Transactions on Medical Imaging*, *39*(8), 2676–2687. https://doi.org/10.1109/TMI.2020.2994459

[14] Šarić, M., Russo, M., Stella, M., & Sikora, M. (2019). CNN-based method for lung cancer detection in whole slide histopathology images. In *4th International Conference on Smart and Sustainable Technologies*, 1–4. https://doi.org/10.23919/SpliTech.2019.8783041

[15] Mehmood, S., Ghazal, T. M., Khan, M. A., Zubair, M., Naseem, M. T., Faiz, T., & Ahmad, M. (2022). Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing. *IEEE Access*, *10*, 25657–25668. https://doi.org/10.1109/ACCESS.2022.3150924

[16] Mangal, S., Chaurasia, A., & Khajanchi, A. (2020). *Convolution neural networks for diagnosing colon and lung cancer histopathological images*. *arXiv Preprint: 2009.03878*

[17] Masud, M., Sikder, N., Nahid, A.-A., Bairagi, A. K., & AlZain, M. A. (2021). A machine learning approach to diagnosing lung and colon cancer using a deep learning-based classification framework. *Sensors*, *21*(3), 748. https://doi.org/10.3390/s21030748

[18] Gabralla, L. A., Hussien, A. M., AlMohimeed, A., Saleh, H., Alsekait, D. M., El-Sappagh, S., . . . , & Refaat Hassan, M. (2023). Automated diagnosis for colon cancer diseases using stacking transformer models and explainable artificial intelligence. *Diagnostics*, *13*(18), 2939. https://doi.org/10.3390/diagnostics13182939

[19] Kumar, N., Sharma, M., Singh, V. P., Madan, C., & Mehandia, S. (2022). An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images. *Biomedical Signal Processing and Control*, *75*, 103596. https://doi.org/10.1016/j.bspc.2022.103596

[20] Chhillar, I., & Singh, A. (2024). A feature engineering-based machine learning technique to detect and classify lung and colon cancer from histopathological images. *Medical & Biological Engineering & Computing*, *62*(3), 913–924. https://doi.org/10.1007/s11517-023-02984-y

[21] Provath, M. A.-M., Deb, K., Dhar, P. K., & Shimamura, T. (2023). Classification of lung and colon cancer histopathological images using global context attention based convolutional neural network. *IEEE Access*, *11*, 110164–110183. https://doi.org/10.1109/ACCESS.2023.3321686

[22] Bhosale, Y. H., & Patnaik, K. S. (2022). IoT deployable lightweight deep learning application for COVID-19 detection with lung diseases using Raspberry Pi. In *2022 International Conference on IoT and Blockchain Technology*, 1–6. https://doi.org/10.1109/ICIBT52874.2022.9807725

[23] Alghareb, F. S., & Hasan, B. T. (2025). Leveraging multithreading on edge computing for smart healthcare based on intelligent multimodal classification approach. *Computerized Medical Imaging and Graphics*, *124*, 102594. https://doi.org/10.1016/j.compmedimag.2025.102594

[24] Xu, Y., Khan, T. M., Song, Y., & Meijering, E. (2025). Edge deep learning in computer vision and medical diagnostics: A comprehensive survey. *Artificial Intelligence Review*, *58*(3), 93. https://doi.org/10.1007/s10462-024-11033-5

[25] Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A., & Mastorides, S. M. (2019). *Lung and colon cancer histopathological image dataset (LC25000)*. arXiv. https://doi.org/10.48550/arXiv.1912.12142

[26] Shafiq, M., & Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied Sciences*, *12*(18), 8972. https://doi.org/10.3390/app12188972

[27] Maharaju, R., & Valupadasu, R. (2025). Advanced cloud framework for lung cancer diagnosis and severity evaluation in deep learning. In *2025 8th International Conference on Information and Computer Technologies*, 423–428. https://doi.org/10.1109/ICICT64582.2025.00072

[28] Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, *139*, 10096–10106.

[29] Sarker, I. H. (2021). Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, *2*(6), 420. https://doi.org/10.1007/s42979-021-00815-1

[30] Maharaju, R., & Valupadasu, R. (2023). Lung cancer classification and prediction of disease severity score using deep learning. In *2023 6th International Conference on Information and Computer Technologies*, 100–104. https://doi.org/10.1109/ICICT58900.2023.00024

[31] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., . . . , & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

[32] Reem, S. Z., Sumon, S. M. A., Howlader, A., & Sarker, M. (2024). A deep learning strategy for accurate lung cancer subtype classification using convolutional neural networks. In *2024 13th International Conference on Electrical and Computer Engineering*, 585–590. https://doi.org/10.1109/ICECE64886.2024.11025053

How to Cite: Maharaju, R., Ellaboina, S. P., & Valupadasu, R. (2026). HistoPath AI: An Edge-Powered Web Platform for Real-Time Lung Cancer Diagnosis. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62028130