http://doi.org/10.35784/iapgos.7616 *received: 22.04.2025 | revised: 10.11.2025 | accepted: 18.11.2025 | available online: 20.12.2025*

REAL-TIME COVID-19 DIAGNOSIS ON EMBEDDED IoT PLATFORMS

Elmehdi Benmalek[1,2,3], Wajih Rhalem[1], Atman Jbari[1], Abdelilah Jilbab[1], Jamal Elmhamdi[1]
[1]Mohammed V University in Rabat, National Higher School of Arts and Crafts of Rabat, Rabat, Morocco, [2]Mohammed VI University of Sciences and Health, Mohammed VI Graduate School of Health Sciences Engineering, Casablanca, Morocco, [3]Mohammed VI Center for Research and Innovation, Research Unit, Rabat, Morocco

Abstract. *COVID-19 continues to pose a persistent global health challenge, where rapid, accurate, and accessible diagnostic tools are crucial for controlling viral transmission. In this study, we present a non-invasive, embedded diagnostic system for COVID-19 detection based on chest CT scan image analysis. The acquired CT images are preprocessed to align with the input dimensions required by four lightweight convolutional neural network (CNN) architectures – ResNet-18, MobileNet, ShuffleNet, and SqueezeNet – selected for their efficiency and suitability in embedded systems. Among these, SqueezeNet achieved a classification accuracy and an f1-score of 99.1%, delivering performance comparable to the other models while offering superior computational efficiency, making it particularly well-suited for real-time, resource-constrained applications. The optimized model was deployed on an NVIDIA Jetson embedded platform to enable on-device, real-time COVID-19 detection at the edge. Diagnostic results are transmitted to the ThingSpeak cloud platform via the MQTT protocol, facilitating continuous, remote health monitoring. Experimental findings confirm the feasibility, accuracy, and real-time operational capability of the proposed embedded system for COVID-19 detection using chest CT scan images.*

Keywords: COVID-19 detection, deep learning, edge computing, embedded systems, IoT health monitoring, CT scan analysis

DIAGNOSTYKA COVID-19 W CZASIE RZECZYWISTYM NA WBUDOWANYCH PLATFORMACH IoT

Streszczenie. *COVID-19 pozostaje poważnym globalnym wyzwaniem zdrowotnym, w którym szybkie, dokładne i łatwo dostępne narzędzia diagnostyczne odgrywają kluczową rolę w kontroli rozprzestrzeniania się wirusa. W niniejszym artykule przedstawiono nieinwazyjny, wbudowany system diagnostyczny do wykrywania COVID-19 na podstawie analizy obrazów tomografii komputerowej (CT) klatki piersiowej. Uzyskane obrazy CT są wstępnie przetwarzane w celu dostosowania ich do wymiarów wejściowych wymaganych przez cztery lekkie architektury konwolucyjnych sieci neuronowych (CNN) – ResNet-18, MobileNet, ShuffleNet oraz SqueezeNet – wybrane ze względu na ich wydajność oraz przydatność w systemach wbudowanych. Spośród nich SqueezeNet osiągnęła dokładność klasyfikacji na poziomie 99%, zapewniając wyniki porównywalne z pozostałymi modelami, przy jednoczesnym zachowaniu większej efektywności obliczeniowej, co czyni ją szczególnie odpowiednią do pracy w czasie rzeczywistym w środowiskach o ograniczonych zasobach. Zoptymalizowany model został wdrożony na platformie wbudowanej NVIDIA Jetson, umożliwiając wykrywanie COVID-19 w czasie rzeczywistym na urządzeniu brzegowym. Wyniki diagnostyczne są przesyłane na platformę chmurową ThingSpeak za pośrednictwem protokołu MQTT, umożliwiając ciągły, zdalny monitoring stanu zdrowia. Wyniki eksperymentalne potwierdzają wykonalność, wysoką dokładność oraz możliwość pracy w czasie rzeczywistym proponowanego systemu wbudowanego do wykrywania COVID-19 na podstawie obrazów tomografii komputerowej klatki piersiowej.*

Słowa kluczowe: wykrywanie COVID-19, uczenie głębokie, przetwarzanie brzegowe, systemy wbudowane, zdalny monitoring zdrowia IoT, analiza obrazów CT

Introduction

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has posed unprecedented challenges to global healthcare systems. As of 2024, over 770 million confirmed cases and more than 7 million deaths have been reported worldwide [19]. In this context, early, rapid, and non-invasive diagnostic solutions remain essential for mitigating viral transmission and improving patient management outcomes. While conventional diagnostic techniques such as reverse transcription polymerase chain reaction (RT-PCR) and rapid antigen tests are widely used, they present notable limitations in terms of accessibility, cost, processing time, and sensitivity under certain clinical conditions [16]. These constraints have intensified interest in alternative, AI-driven diagnostic tools leveraging medical imaging modalities, particularly chest computed tomography (CT) scans.

Chest CT imaging has demonstrated considerable clinical value in detecting characteristic pulmonary features associated with COVID-19, including ground-glass opacities, consolidations, and bilateral infiltrates [20]. The increasing availability of public CT image datasets, combined with rapid advancements in artificial intelligence (AI) and deep learning (DL) techniques, has enabled the development of automated diagnostic systems powered by convolutional neural networks (CNNs). These models are capable of directly analyzing CT scan images to accurately classify COVID-19 positive and negative cases, offering a valuable decision-support tool to complement traditional clinical diagnostics [2–17].

In this study, we introduce a real-time, embedded COVID-19 diagnostic system based on the direct classification of chest CT scan images using lightweight, computationally efficient CNN architectures. Four models – ResNet-18 [8], MobileNet [9], ShuffleNet [21], and SqueezeNet [20] – were selected for evaluation due to their suitability for deployment in embedded healthcare applications. To enable continuous and remote health monitoring, the optimized model was deployed on an NVIDIA Jetson embedded platform, providing on-device, real-time COVID-19 detection at the edge. Diagnostic results are transmitted to a ThingSpeak cloud dashboard via the lightweight MQTT protocol, facilitating immediate, remote access to patient diagnostic data. The proposed system architecture is illustrated in Figure 1.

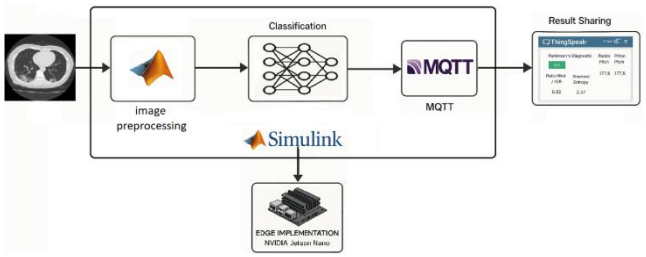


Fig. 1. System architecture

This work represents a significant contribution toward scalable, accessible, and embedded AI-powered diagnostic technologies for COVID-19 and potentially other respiratory diseases in both clinical and decentralized healthcare settings.

1. Related works

In recent years, there has been a growing interest in leveraging medical imaging modalities such as chest X-rays and CT scans for the early detection and diagnosis of COVID-19 and other respiratory diseases. Early investigations by [3] demonstrated the feasibility of identifying COVID-19 infection through the detection of characteristic pulmonary features in CT images, such as ground-glass opacities and consolidations. Building on this foundation, [6] proposed an AI-assisted COVID-19 diagnostic framework using CNN applied directly

to CT scan images, reporting high diagnostic accuracy and, in several instances, outperforming conventional clinical assessments.

With the rapid advancement of DL techniques, the focus has shifted toward optimizing CNN architectures for COVID-19 image classification tasks. Notable studies by [14-15] evaluated popular deep CNN models, including ResNet, DenseNet, and VGGNet, on large-scale, publicly available CT datasets. These works consistently reported strong classification performance in differentiating COVID-19 positive and negative cases, confirming the potential of deep learning as a supplementary diagnostic tool. Similarly, [3] conducted a comparative study of deep learning applications for COVID-19 detection using both chest X-ray and CT images, underscoring the increasing interest in non-invasive, AI-driven diagnostic systems for respiratory disease screening. Beyond imaging-based approaches, recent studies have also investigated acoustic and audio-based COVID-19 detection methods. Notably, [4] applied machine learning techniques to voice recordings for automatic COVID-19 detection, further validating the potential of AI-driven audio signal analysis for non-invasive, remote screening applications.

In parallel, the emergence of embedded AI and edge computing has prompted efforts to adapt diagnostic systems for deployment on resource-constrained hardware. Lightweight CNN architectures such as MobileNet, ShuffleNet, and SqueezeNet have gained considerable attention due to their ability to balance diagnostic accuracy with computational efficiency [8, 9]. However, the majority of existing research relies on high-performance cloud-based servers or offline processing frameworks, which are not ideal for real-time, decentralized, or resource-limited healthcare environments. Only a limited number of recent studies have explored the integration of COVID-19 detection systems into embedded platforms like Raspberry Pi and NVIDIA Jetson devices for remote diagnostic applications. Notably, [5] reviewed embedded machine learning systems for healthcare, highlighting preliminary attempts that primarily utilized chest X-ray images for COVID-19 screening on edge devices, leaving the potential of CT scan-based remote embedded diagnostics largely underexplored.

This study addresses this research gap by proposing a real-time, embedded COVID-19 diagnostic system based on the direct classification of chest CT scan images using lightweight CNN models and IoT platform. The system deploys the most efficient model on an NVIDIA Jetson embedded system, integrating MQTT-based communication for real-time transmission of diagnostic outcomes to a ThingSpeak cloud dashboard. This architecture offers a scalable, low-latency, and resource-efficient solution for continuous remote patient health monitoring, tailored specifically for COVID-19 detection from CT scan images in edge computing environments.

2. Methods

This section outlines the dataset used, preprocessing procedures, CNN-based classification framework, and the implementation of the diagnostic system on an embedded platform.

2.1. Dataset description

For this study, we utilized the publicly available HUST COVID-19 CT dataset [13], in which 19,685 CT scans categorized into three different classes:

- Non-informative CT (NiCT): 5,705 samples, with which the lung parenchyma was inadequate for diagnostic examination.
- Positive CT (pCT): 4,001 samples, showing prominent imaging features relevant to COVID-19 pneumonia.
- Negative CT (nCT): 9,979 samples, upon which no imaging features were present for COVID-19 pneumonia.

This dataset provides a comprehensive and well-annotated resource for developing and evaluating deep learning models aimed at COVID-19 detection using chest CT imaging. Representative examples of CT scan images for positive, negative, and non-informative cases are illustrated in Figure 2.

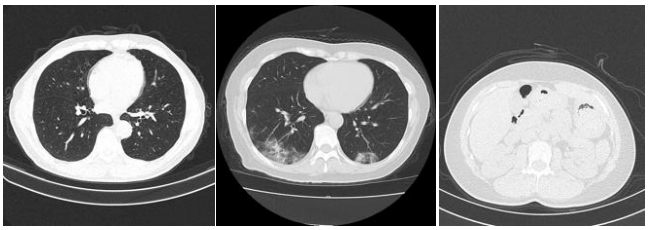


Fig. 2. CT scan images from the dataset

Representative chest CT images of every category of COVID-19 pneumonia are shown in Figure 3. Compared to the normal lung CT image of the control group, no sign of pneumonia was observed in either the right or the left lung. The mild COVID-19 pneumonia was characterized by ground-glass opacities (GGO) limited to small parts of the lungs. The mild manifestation typically presented multiple lesions with heterogeneous GGO, consolidation, and linear shadows, whereas the severe manifestation presented extensive GGO and linear consolidation in both lungs. Radiographic features in critically ill patients primarily comprised widespread consolidation, interlobular septal thickening, and the presence of air bronchograms. Suspected cases can present one or more radiographic features that simulate any one of these stages of COVID-19 pneumonia [13].

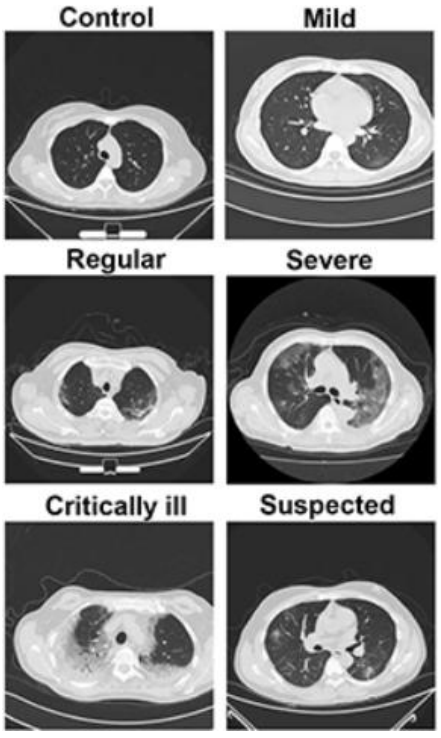


Fig. 3. Representative chest CT images for six types of cases

2.2. Image preprocessing

All CT scan images were resized to 224×224×3 pixels to meet the input size requirements of the selected CNN architectures. Image normalization was applied by scaling pixel intensity values to the (0, 1) range, which enhances training stability and accelerates convergence. To maintain a consistent evaluation of the models' baseline diagnostic performance, no data augmentation techniques were employed in this study.

2.3. Deep learning classifiers

Four lightweight CNN architectures – ResNet-18, MobileNet, ShuffleNet, and SqueezeNet – were selected based on their demonstrated effectiveness in image classification tasks and suitability for deployment on embedded systems with limited computational resources. All models were implemented using transfer learning, with pretrained weights from the ImageNet dataset. The final fully connected layer in each model was replaced with a customized layer outputting three class probabilities: COVID-19 Positive, Negative, and Non-informative CT scans.

The dataset was randomly split into 70% for training, 10% for validation, and 20% for testing. Training was conducted using the categorical cross-entropy loss function and the Adam optimizer. An early stopping mechanism, based on validation loss, was applied to prevent overfitting.

2.4. Embedded implementation on NVIDIA Jetson

Following performance evaluation, a model was selected for deployment based on its optimal balance between diagnostic performance and computational efficiency, making it particularly well-suited for real-time, embedded applications. The trained model was converted into an embedded-compatible format using MATLAB Coder and GPU Coder, generating CUDA-accelerated code optimized for the GPU resources available on the NVIDIA Jetson platform.

The embedded system was configured to acquire CT scan images from a local directory or an integrated imaging system, perform classification inference using the optimized classification model, and generate a diagnostic label according to the following encoding:

- 0 → COVID-19 Negative (nCT),
- 1 → COVID-19 Positive (pCT),
- 2 → Non-informative (NiCT).

The high-performance Jetson platform facilitated real-time image classification and diagnostic result generation directly at the edge.

2.5. Remote monitoring using MQTT and ThingSpeak

To support real-time, remote health monitoring, the embedded system was integrated with a ThingSpeak cloud dashboard via the MQTT (Message Queuing Telemetry Transport) protocol. MQTT is a lightweight communication protocol, ideal for IoT and embedded healthcare applications.

Upon classifying each CT image, the system published the diagnostic result as a numerical label (0, 1, or 2) to a designated ThingSpeak channel. This configuration enables healthcare professionals and monitoring systems to remotely visualize, log, and analyze diagnostic outcomes in real time, supporting continuous COVID-19 patient monitoring and facilitating prompt clinical decision-making.

3. Results

3.1. Classification results

The classification performance of the four evaluated CNN architectures – MobileNet, ResNet-18, ShuffleNet, and SqueezeNet – was assessed using confusion matrices, classification metrics: accuracy, recall, precision and F1-score and Receiver Operating Characteristic (ROC) curves. The confusion matrices and ROC curves achieved by each model are summarized in Table 1.

The confusion matrices provided detailed insights into the classification accuracy for the three diagnostic categories: pCT, nCT, and NiCT. Ideally, high values are expected along the diagonal elements, indicating correct classifications, with minimal off-diagonal values representing misclassifications. As illustrated in Table 1, all models demonstrated excellent classification capability, with the highest sensitivity and specificity consistently observed across the three categories. Along with the confusion matrices, the recall and the precision metrics are illustrated for each class in the column in the right and the row in the bottom of each matrix.

To compare the performance of proposed classification models, traditional evaluation measures – precision, recall, and F1-score – were used alongside overall accuracy. These measures provide a comprehensive view of the performance of the classifier on various classes.

Precision (Positive Predictive Value) according to the number of accurately predicted positive samples over all samples predicted positive. It tells us to what extent the model is capable of avoiding false positives, and is given by:

$$precision = \frac{TP}{TP+FP} \tag{1}$$

where *TP* and *FP* are true positives and false positives, respectively.

Recall (True Positive Rate or Sensitivity) quantifies the proportion of correct detection of positive instances by the model. It indicates the ability of the model to detect all the true positives, and is expressed as:

$$recall = \frac{TP}{TP+FN} \tag{2}$$

where *FN* represents false negatives.

F1-score is the harmonic mean of recall and precision, assigning equal importance to both the measures. It is particularly useful when the data set is unbalanced based on classes. F1-score is computed as:

$$F1 - score = \frac{2*recall*precision}{recall+precision} \tag{3}$$

Here, the F1-score of each class was calculated to evaluate the performance of the model to accurately distinguish the three diagnostic classes correctly. High F1-scores in all classes indicate that the trained model achieves high sensitivity and high predictive confidence, which are essential to safe medical diagnosis.

All the models achieved very high classification performance with overall accuracy values of 99.0% for MobileNet, 99.5% for ResNet-18, 99.0% for ShuffleNet, and 99.1% for SqueezeNet. ResNet-18 produced the highest accuracy (99.5%) and performed better F1-scores in all classes (NiCT: 99.1%, nCT: 99.6%, pCT: 99.6%), confirming its stability and reliability for COVID-19 detection applications.

However, considering the deployment constraints of edge or embedded devices, model efficiency and parameter counts are of utmost importance. SqueezeNet, with approximately 1.2 million parameters, achieved 99.1% accuracy and possessed symmetric F1-scores for classes (NiCT: 98.5%, nCT: 99.2%, pCT: 99.4%), for an F1-score total of 99.1% forming the best performance vs. computational expense compromise. MobileNet and ShuffleNet, with approximately 3.4 million and 2.3 million parameters, respectively, also obtained comparatively F1-score 99% and 98.6%, respectively. Although ResNet-18 was marginally more accurate, its larger parameter count (~11.7 million) and computational requirement make it an undesirable option for real-time, embedded systems.

ROC curves were plotted for each CNN model, assessing the trade-offs between sensitivity (true positive rate) and specificity (false positive rate). The Area Under the Curve (AUC) values confirmed the performance for all models, with values approaching 1.0, indicating near-perfect classification capabilities in distinguishing between COVID-19 positive, negative, and non-informative CT images.

Table 1. confusion matrices and ROC curve of all models

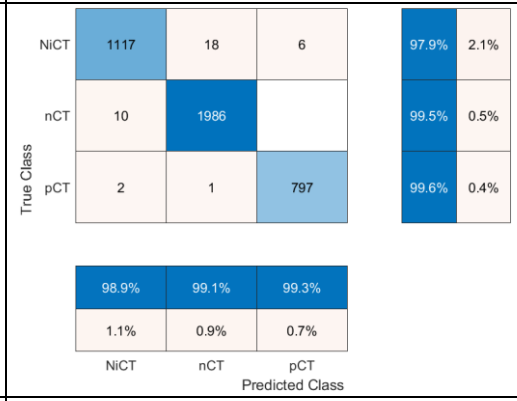
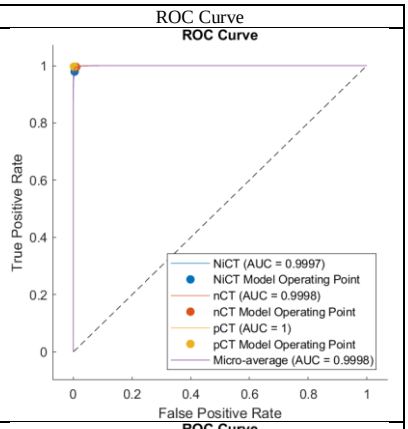
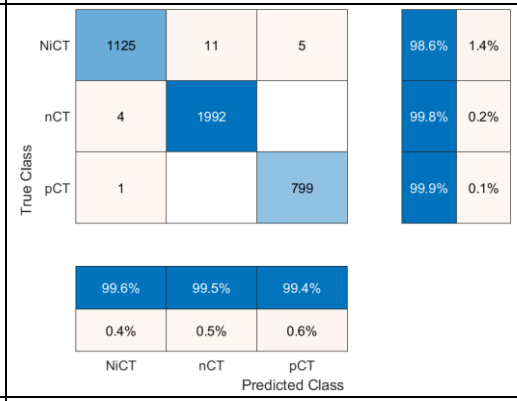
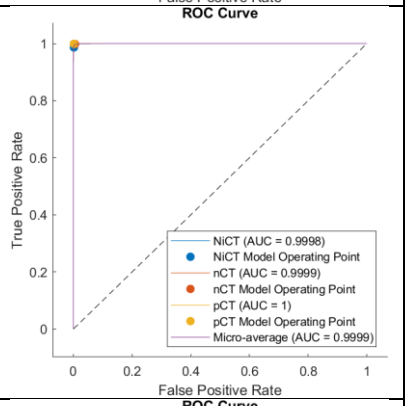
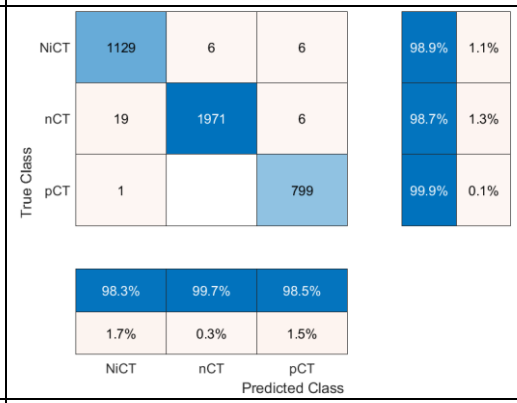
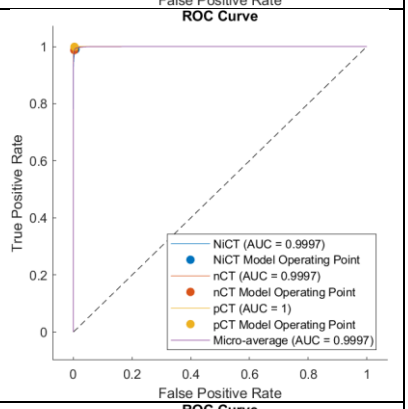
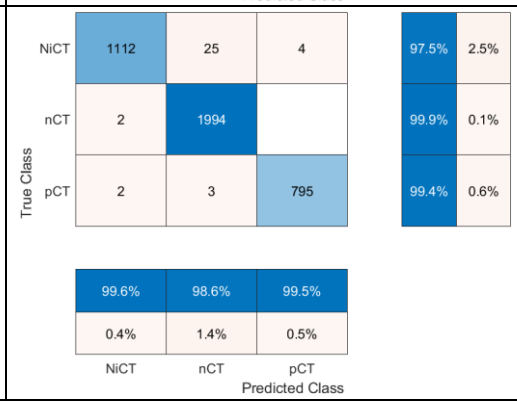
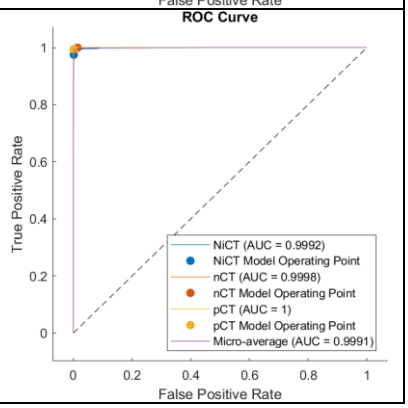
CNN	Confusion matrix	ROC Curve
Mobilenet		
Resnet18		
Shufflenet		
Squeezenet		

Table 2. Performance metrics

Classifier	Accuracy (%)	F1 Score (%)				Nbe of parameters
		NiCT	nCT	pCT	Total	
MobileNet	99	98,4	99.6	99.4	99	3.4M
ResNet-18	99.5	99.1	99.6	99.6	99.4	11.7M
ShuffleNet	99	98.6	98.2	99.2	98.6	2.3M
SqueezeNet	99.1	98.5	99.2	99.4	99.1	1.2M

3.2. Simulink model integration and embedded deployment

A comprehensive Simulink model was developed to integrate CT image acquisition, image preprocessing, deep learning classification, result visualization, and cloud-based transmission, as depicted in Figure 4. This model facilitated the real-time, embedded implementation of the proposed diagnostic system on an NVIDIA Jetson platform, leveraging MATLAB Coder and GPU Coder for CUDA-optimized embedded code generation.

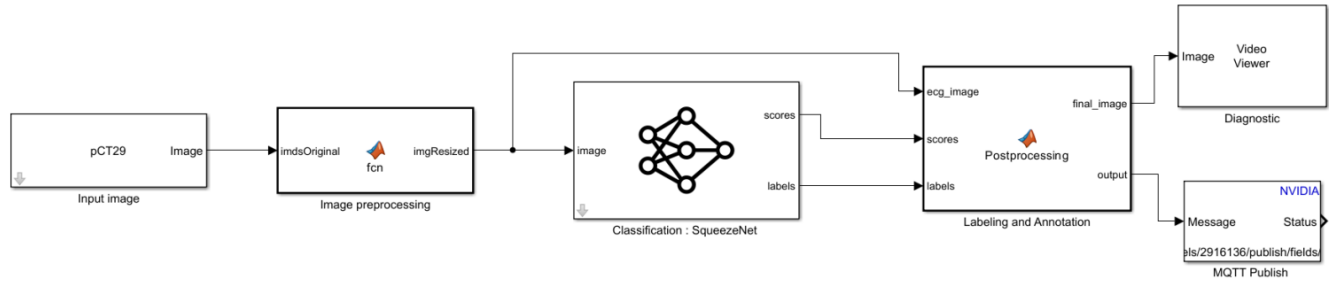


Fig. 4. Simulink model of the system

The Simulink-based system comprised the following functional components:

- CT Image Input Block: Acquired CT scan images from a local directory, imaging hardware, or simulated source.
- Image Preprocessing Block: Resized and normalized the images to 224×224×3 pixels, matching the input size requirements of the selected CNN model.
- Classification Block: Processed the image using the optimized SqueezeNet model to generate a diagnostic label:
  - 0 → COVID-19 Negative,
  - 1 → COVID-19 Positive,
  - 2 → non-informative.
- Labeling and Annotation Block: Figure 5 shows how the diagnostic label is overlaid onto the original CT image, providing instant, real-time visualization of the diagnostic outcome.
- Video Viewer Block: Displayed the labeled CT image alongside its diagnostic outcome within the embedded system interface.
- MQTT Publish Block: Transmitted the classification result to a ThingSpeak cloud platform via the MQTT protocol for real-time remote monitoring.

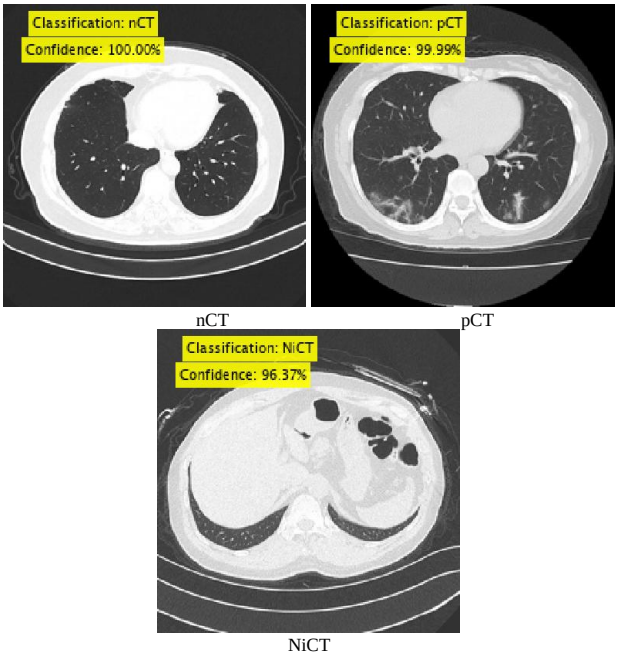


Fig. 5. Results of the classification

The inference runs were conducted on an NVIDIA Jetson Nano embedded board with an NVIDIA Tegra X1 GPU, CUDA 10.0, cuDNN 7.3, and TensorRT 5.0 acceleration. The software stack included OpenCV 3.3.1, GStreamer 1.14.5, V4L2 1.14.2-1, and SDL 1.2 for image capture and preprocessing. For deployment using SqueezeNet, the performance-tuned TensorRT engine achieved an average inference throughput of 24.6 frames per second (FPS) or an equivalent average latency of approximately 40.7 milliseconds per frame. Inference utilized approximately 720 MB of system RAM and 480 MB of GPU memory (VRAM) and averaged a 4.8 W power consumption under the Jetson Nano's 10 W performance mode. These results confirm that, with the platform's limited computational and energy resources, SqueezeNet maintains real-time performance, low memory consumption, and high energy efficiency to achieve 99.1% classification accuracy with ~1.2 million parameters only. Consequently, the model demonstrates superior embedded or portable medical imaging system applicability, allowing efficient in-device COVID-19 CT scan analysis without reliance on cloud or high-end-hardware resources.

The complete system successfully achieved real-time embedded execution, efficiently classifying incoming CT images and transmitting diagnostic outcomes to the cloud for immediate remote access. The numeric result encoding streamlined data transmission while preserving interpretability through corresponding diagnostic labels and live image annotations displayed on both the local interface and cloud-based dashboard (Figure 6).

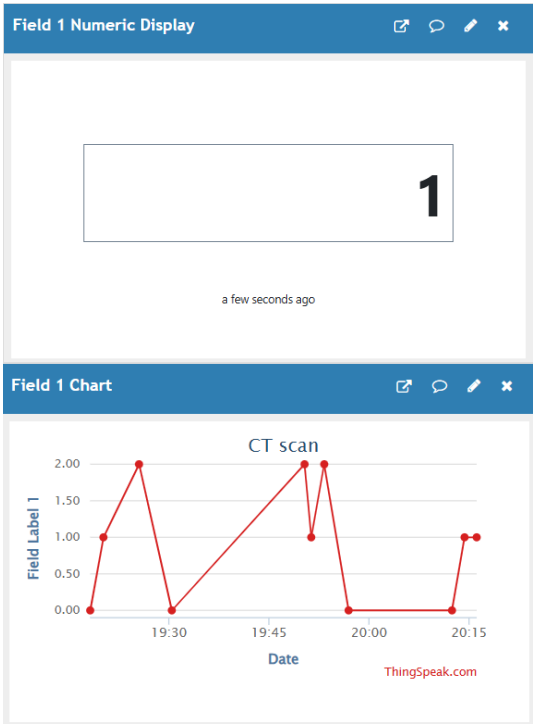


Fig. 6. Thingspeak results

4. Discussion

In this study, we proposed and validated an innovative, non-invasive system for the real-time diagnosis of COVID-19 based on CT scan image analysis. By employing deep learning techniques with lightweight CNN architectures – MobileNet, ResNet-18, ShuffleNet, and SqueezeNet – the system demonstrated highly accurate classification capabilities, reliably distinguishing between COVID-19 positive, negative, and non-informative cases, achieving classification accuracies and F1-score exceeding 99% across all models. Several works have demonstrated that with judicious architectural choices

and optimization techniques, high diagnostic accuracy can be maintained while latency and power consumption can be significantly reduced. In Table 3 we compare our findings with those of previous works.

In Lou et al.'s [11] study, the authors implemented an ultralightweight SqueezeNet model for the diagnosis of COVID-19 on the NVIDIA Jetson Nano platform. It was indicated that the optimized network achieved 97.5% classification accuracy on CT scans, while inference latency was under 100 ms per image with TensorRT acceleration. The inference power usage was between 5–10 W, demonstrating the feasibility of the Jetson Nano for real-time diagnosis under restricted energy budgets [11]. Similarly, Ulutaş et al. [18] deployed an embedded edge-optimized CNN model and achieved 96.8% accuracy for COVID-19 detection with a reduction in inference time by 30% compared to non-optimized implementations [18]. These results confirm that GPU-based embedded systems can preserve high diagnostic accuracy with near real-time performance.

On the other hand, CPU-based embedded boards such as the Raspberry Pi 4 possess higher inference latency but lower total power consumption. Ezzat et al. [7] implemented a MobileNetV2 model for pneumonia detection and achieved an inference latency of approximately 300–400 ms per image with an average power consumption of 3–4 W [7]. While the accuracy remained close to 94%, the limited computational resources of the Raspberry Pi restricted throughput, making it more suitable for low-rate or offline medical applications. Ahmed et al. (2021) also noted that post-quantization and pruning techniques can reduce latency by nearly 40% with minimal accuracy loss (<2%), highlighting the effectiveness of model compression in CPU-based inference [1].

FPGAs, on the other hand, provide a compelling balance between energy efficiency and deterministic latency. Nechi et al. [12] demonstrated an FPGA-based CNN accelerator for chest CT classification with 98.1% accuracy and inference latency of less than 10 ms while consuming only 1.2 W [12]. Similarly, Zhou et al. (2022) demonstrated an FPGA implementation of a fixed-point quantized ResNet with comparable accuracy to its floating-point equivalent but 6× more energy efficiency [22]. These works indicate the FPGA's advantage in use cases where ultra-low power and low latency are critical, but at the cost of increased development time and reduced flexibility compared to general-purpose embedded GPUs.

Overall, the papers surveyed here show that Jetson Nano provides an optimal trade-off of accuracy, latency, and usability, particularly for medical imaging use cases requiring real-time inference. Raspberry Pi remains a good option for low-power or educational deployments, and FPGA boards still have unmatched energy efficiency for fixed inference workloads. Collectively, these findings demonstrate that hardware-specific optimization – i.e., quantization, pruning, and accelerator utilization – can reduce computational cost by up to 60–80% with no significant compromise in diagnostic accuracy. Therefore, the selection of an embedded platform must balance power, latency, and precision requirements according to the target application.

A key strength of the proposed system lies in its integration of real-time inference with continuous, remote monitoring capabilities. The adoption of MQTT-based communication ensures efficient, low-latency data transmission to the cloud, allowing healthcare providers to remotely access and analyze diagnostic outcomes in real time. Clinically, this offers a non-invasive alternative to traditional COVID-19 diagnostic approaches and alleviates challenges related to hospital overcrowding, limited testing infrastructure, and delayed results in under-resourced regions.

Moreover, the embedded, scalable, and adaptable nature of the proposed system positions it as a practical solution for a wide range of telehealth and public health applications. Beyond COVID-19, the framework can be readily extended to support the detection of other respiratory diseases such as pneumonia, bronchitis, and asthma by incorporating additional medical imaging datasets and training customized classification models. Additionally, the system holds potential for longitudinal disease monitoring, enabling the assessment of disease progression over time through sequential CT imaging.

In summary, this study demonstrates the feasibility and effectiveness of an AI-powered, embedded diagnostic platform for COVID-19 detection using CT scan images, combining high diagnostic accuracy with real-time remote monitoring functionality. Its integration into edge-based healthcare infrastructures could significantly enhance early disease detection, patient management, and healthcare accessibility, particularly in remote, home-care, or resource-limited settings.

Table 3. Summary of studies implementing CNNs on embedded systems

Study	Embedded platform	Model / Task	Accuracy (%)	Inference latency	Power consumption
Lou et al. [11]	Jetson Nano	SqueezeNet for COVID-19 CT classification	97.5	<100 ms per image (TensorRT optimized)	5–10 W
Ulutaş [18]	Jetson Nano	Custom CNN for COVID-19 CT images	96.8	30% faster than baseline model	~8 W
Ezzat et al. [7]	Raspberry Pi 4	MobileNetV2 for pneumonia detection	94.0	300–400 ms per image	3–4 W
Ahmed et al. [1]	Raspberry Pi 4 / Edge CPU	Pruned & quantized CNN (general medical imaging)	92–95	40% latency reduction vs. baseline	2.5–3 W
Nechi et al. [12]	FPGA (Xilinx Zynq)	CNN for chest CT classification	98.1	<10 ms per image	1.2 W
Zhou et al. [22]	FPGA (Intel Arria 10)	Quantized ResNet for medical imaging	97.8	~8 ms per image	<2 W
this study	Jetson Nano	SqueezeNet for COVID-19 CT classification	99.1	~41 ms	4.8 W

5. Conclusion

This paper presented an embedded, real-time diagnostic platform for automated COVID-19 detection from CT scans using lightweight CNNs. SqueezeNet, among the four architectures that were evaluated, was used to deploy because it presented the best compromise between accuracy (99.1%) and computational complexity and thus was particularly well-suited for embedded inference. The model was deployed onto an NVIDIA Jetson Nano and natively integrated onto the ThingSpeak cloud platform via the MQTT protocol for real-time remote monitoring and data visualization. Two data fields were generated within the cloud interface – one for the diagnostic outcome (0 = negative CT, 1 = positive COVID-19 CT, 2 = uninformative CT) and one for the confidence score of the model – to facilitate efficient, interpretable, and low-bandwidth communication for clinical or mobile health use.

Experimental performance confirms that direct high-accuracy and low-latency medical inference based on edge devices is possible, validating the feasibility of employing AI-driven diagnostic systems in decentralized or resource-constrained healthcare environments. Beyond COVID-19 diagnosis, the framework established here presents a scalable structure for biomedical imaging based on IoT, to further migration toward connected, autonomous health systems. Future work will focus on clinical validation, cross-domain generalizability, and extension of diagnostic functionality to other respiratory illnesses such as pneumonia and bronchitis, leading to a more robust and field-deployable system.

References

[1] Ahmed S. M. et al.: Optimized Deep Neural Networks for Edge Computing Devices Using Quantization and Pruning. IEEE Access 9, 2021, 119234–119245.
[2] Aslani S., Jacob J.: Utilisation of deep learning for COVID-19 diagnosis. Clinical Radiology 78(2), 2023, 150–157.
[3] Benmalek E., Elmhamdi J., Jilbab A.: Comparing CT scan and chest X-ray imaging for COVID-19 diagnosis. Biomedical Engineering Advances 1, 2021, 100003.
[4] Benmalek E. et al.: Automatic COVID-19 detection using machine learning and voice recording. Research on Biomedical Engineering 39(3), 2023, 597–612.
[5] Biglari A., Tang W.: A review of embedded machine learning based on hardware, application, and sensing scheme. Sensors 23(4), 2023, 2131.
[6] Chaddad A., Hassan L., Desrosiers C.: Deep CNN models for predicting COVID-19 in CT and x-ray images. Journal of medical imaging 8(S1), 2021, 014502–014502.
[7] Ezzat D., Hassanien A. E., Elnakib A.: Deep Learning Models for COVID-19 Pneumonia Detection on Raspberry Pi Platform. Computers in Biology and Medicine 142, 2022, 105244.
[8] He K. et al.: Deep residual learning for image recognition. IEEE conference on computer vision and pattern recognition. 2016, 770–778 [https://doi.org/10.1109/CVPR.2016.90].
[9] Howard A. G. et al.: MobileNets: Efficient convolutional neural networks for mobile vision applications. 2017. arXiv preprint arXiv:1704.04861 [https://arxiv.org/abs/1704.04861].
[10] Iandola F. N. et al.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. 2016. arXiv preprint arXiv:1602.07360 [https://arxiv.org/abs/1602.07360].
[11] Lou L., Liang H., Wang Z.: Deep Learning-Based COVID-19 Diagnosis and Implementation in Embedded Edge-Computing Device. Diagnostics 13(7), 2023, 1329.
[12] Nechi A. et al.: FPGA-Based Deep Learning Inference Accelerator for COVID-19 CT Classification. IEEE Transactions on Biomedical Circuits and Systems 17(3), 2023, 412–422.
[13] Ning W. et al.: iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia. 2020.
[14] Pham T. D.: A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks. Scientific reports 10(1), 2020, 16942.
[15] Shah V. et al.: Diagnosis of COVID-19 using CT scan images and deep learning techniques. Emergency radiology 28, 2021, 497–505.
[16] Sharma N. et al.: Coswara: A database of breathing, cough, and voice sounds for COVID-19 diagnosis. Journal of the Acoustical Society of America, 149(1), 2023, 548–563.
[17] Shorten C., Khoshgoftaar T. M., Furht B.: Deep Learning applications for COVID-19. Journal of big Data 8(1), 2021, 1–54.
[18] Ulutaş H.: Application of a Novel Deep Learning Technique Using CT Images to Implement the COVID-19 Automatic Diagnosis System on Embedded Systems. Alexandria Engineering Journal 70, 2023, 120–130.
[19] World Health Organization: COVID-19 dashboard. 2024 [https://covid19.who.int/].
[20] Wu X. et al.: COVID-AL: The diagnosis of COVID-19 with deep active learning. Medical Image Analysis 68, 2021, 101913.
[21] Zhang X. et al.: ShuffleNet: An extremely efficient convolutional neural network for mobile devices. IEEE conference on computer vision and pattern recognition. 2018. 6848–6856 [https://doi.org/10.1109/CVPR.2018.00716].
[22] Zhou Y., Chen L., Zhao J.: Energy-Efficient FPGA Implementation of Quantized ResNet for Medical Image Classification. IEEE Access 10, 2022, 68915–68924.

Prof. Elmehdi Benmalek
e-mail: elmehdi.benmalek@um5s.net.ma

Received his Ph.D. from Mohammed V University in Rabat in 2019. He is currently the Director of the Mohammed VI Graduate School of Health Sciences Engineering in Dakhla.
His research interests include signal processing, embedded systems and machine learning.

https://orcid.org/ 0000-0003-1078-1421

Prof. Wajih Rhalem
e-mail: w.rhalem@um5r.ac.ma

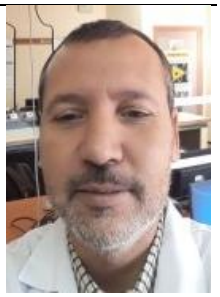
Obtained his doctoral degree in electrical engineering from Mohammed V University in Rabat. Currently, he is president of the Moroccan Society of Digital Health and professor of electrical engineering at the National School of Arts and Crafts under the University Mohammed V of Rabat. His research interests focus on artificial intelligence, digital health and bioinformatics.

https://orcid.org/0000-0001-6221-6833

Prof. Atman Jbari
e-mail: a.jbari@um5r.ac.ma

He is currently a professor at the electrical engineering department of ENSET "Ecole Normale Supérieure de l'Enseignement Technique", Mohamed V University in Rabat, Morocco. In 2009, he received his Ph.D. in computer and telecommunications from Mohammed 5 University. He is member of Electronic Systems, Sensors and Nano-biotechnology research group. His current research interests include signal processing, blind source separation and embedded electronic systems.

https://orcid.org/0000-0002-1855-2503



Prof. Abdelilah Jilbab
e-mail: a_jilbab@yahoo.fr

He is an electrical engineering professor at ENSAM (formerly known as ENSET) at Mohammed V University in Rabat, Morocco. He earned his Ph.D. in Computer and Telecommunication from Mohammed V-Agdal University, Rabat, Morocco, in February 2009. His doctoral thesis revolved around Internet content filtration, explicitly contributing to image recognition techniques based on the Principle of Maximum Entropy. His ongoing research centers on various subjects, including image processing, engineering, information systems, and artificial intelligence.

https://orcid.org/0000-0002-1577-9040

Prof. Jamal Elmhamdi
e-mail: mhamdi_jamal@yahoo.fr

Holds the position of a professor in the field of electrical engineering at the "Ecole Nationale Superieur des Arts et M´etiers" (ENSAM, previously known as ENSET), situated within Mohammed V University in Rabat, Morocco. His academic journey includes attaining a Ph.D. in electrical, electronic, and telecommunication engineering from the University of Rennes, France 1988. His scholarly pursuitsspan signal processing, communication systems, and artificial intelligence domains.

https://orcid.org/0000-0001-8219-3560