Analysis of Contrast-Enhanced Triphasic CT Scans Using Convolution Neural Networks on NVIDIA Jetson Nano

Angelique P. Bagtas[1], Khaye Q. Arellano[1], Eliezer Gale C. Bador[1], Bianca D. Duyag[1], Albert V. Josef Jr.[1],Janine Caile M. Polecina[1], Justin Ryan L. Tan[2]

[1]*Department of Electronics Engineering, Technological University of the Philippines, Manila, Philippines*
[2]*Department of Internal Medicine, Section of Gastroenterology, Chinese General Hospital and Medical Center*

{angelique.bagtas, khaye.arellano, eliezergale.bador, bianca.duyag, albert.josef, janinecaile.polecina}@tup.edu.ph}@tup.edu.ph

***Abstract*— This research paper presents an in-depth analysis of the application of Convolutional Neural Networks (CNNs) for the evaluation of contrast-enhanced triphasic CT scans in the context of focal liver lesion (FLL) detection. The study focuses on the utilization of automated machine learning (AutoML), specifically the Roboflow train, for image segmentation and classification tasks related to FLLs. The research aims to assess the accuracy and effectiveness of AutoML-based applications through internal validation, comparative analysis, and prototype testing. The findings reveal that the AutoML using Roboflow Train achieved a 100% accuracy rate based on the matched diagnosis of the AutoML to the official CECT scan report, indicating its outperformance of the ResUNet algorithm in image segmentation and classification of FLLs. The study's comprehensive evaluation metrics, including mean Average Precision (mAP) and confusion matrices, demonstrate the robustness and potential of the Roboflow train in accurately detecting and segmenting FLLs. Furthermore, the comparative analysis between AutoML and the ResUNet model provides valuable insights into the potential of AutoML-based applications for medical image analysis, highlighting its comparable accuracy to traditional classifiers and its potential for future adoption in clinical settings. The research contributes to the advancement of computer-aided diagnosis in the medical field, emphasizing the significance of accurate detection and classification of liver lesions for improved patient care and management.**

Keywords—Automated Machine Learning (AutoML), Contrast-Enhanced Computed Tomography (CECT) Scan, Convolutional Neural Networks (CNNs), Focal Liver Lesions (FLLs), Image Classification, Image Segmentation, Mean Average Precision (mAP), Medical Image Analysis, , Roboflow Train.

I. INTRODUCTION

In the realm of data science, computer vision is a widely explored area, with Convolutional Neural Networks (CNNs) emerging as the forefront technique. CNNs are highly favored among various neural network architectures and are commonly applied in processing image data. They excel in tasks like image classification, object detection, and image recognition, thus finding extensive use in artificial intelligence research, particularly for constructing image classifiers [1]. CNNs typically consist of tens or even hundreds of layers, each trained to identify specific features within an image. During training, each image undergoes filtering at various resolutions, with the resulting convolved images serving as inputs for subsequent layers [2]. It is now essential to use deep learning techniques in many spheres of life. The top of the list is the medical industry, where fast testing and accurate diagnosis can save a lot of time and lives. Nevertheless, most medical data is not readily available in sufficient quantities to be utilized with conventional deep learning methods. Finding alternative approaches to handling small datasets is crucial.

ResUNet, a combination of UNET and RESNET models, constitutes a deep architecture leveraging residual connections for enhanced performance. Comprising encoder, decoder, and bridge components, it relies on convolutional layers instead of basic building elements. Unlike conventional methods employing pooling for feature size reduction, the encoder utilizes a stride of 2 in its initial convolution block. Each decoding unit is preceded by a transfer layer facilitating feature exchange between encoder and decoder segments. Finally, a 1 × 1 convolution with Sigmoid activation generates the segmentation map [3]. ResUNet integrates residual connections to enable deeper network structures, addressing challenges like the vanishing gradient problem in deep learning. Additionally, ResUNet's skip connections preserve spatial details, enabling precise identification of anatomical structures and anomalies in medical images [4]. Conversely, AutoML aims to automate model selection, combination, and parameter setting for optimal performance on specific tasks or datasets. It streamlines the development of tailored image recognition models using techniques such as neural architecture search and transfer learning. Automated machine learning (AutoML) has gained significant research interest for automating the construction of machine learning pipelines within predefined computational constraints. AutoML encompasses various pipeline aspects, with particular focus on neural architecture search (NAS) in recent studies, aiming to identify optimal neural structures within defined search spaces and computational limits [5]. The study aims to evaluate the accuracy of Automated machine learning (AutoML), particularly the Roboflow train, in image segmentation and classification tasks associated with focal liver lesions in comparison to ResUnet, a popular tool in medical imaging.

II. BACKGROUND OF THE PROBLEM

Liver diseases can be caused by various types of masses, such as cysts, abscesses, and tumors. Liver lesions can be diagnosed and treated using various imaging modalities, such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound (US). These techniques can help segment the liver and the lesions from the surrounding abdominal organs and tissues. However, this task is challenging due to the variability in the shape, size, location, and appearance of the liver and its lesions. Moreover, some lesions may have similar intensity values or boundaries as the normal liver tissue, making them hard to distinguish. Therefore, many researchers have been exploring the use of deep learning algorithms, especially CNN, to improve the accuracy and efficiency of liver lesion segmentation and classification from abdominal CT images. CNN can learn high-level features from the images through multiple convolutional layers and perform classification or segmentation tasks using fully connected layers or up sampling layers. CNN have shown promising results in various medical image analysis applications, such as tumor detection, organ segmentation, and disease diagnosis [6].

The advent of advanced imaging technologies has significantly transformed the medical field, particularly in the diagnosis and management of liver diseases. Focal liver lesions

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

(FLLs) are a common finding in radiological practice, and their accurate classification is crucial for patient management. Triphasic contrast-enhanced computed tomography (CECT) scans are widely used for the detection and characterization of FLLs [7]. However, the interpretation of these scans requires a high level of expertise and can be time-consuming. Automated machine learning (AutoML), specifically the Roboflow train, has shown promise in performing image segmentation and classification tasks. It has the potential to expedite the process and improve the accuracy of FLL detection and classification [8]. However, its performance compared to established models like ResUnet is yet to be thoroughly evaluated [9]. Moreover, the effectiveness of AutoML-based applications needs to be validated internally to ensure their reliability and applicability in real-world settings [10]. This study aims to address these gaps by evaluating the accuracy of the Roboflow train compared to ResUnet for image segmentation and classification tasks related to FLLs and testing the effectiveness of the AutoML-based application through internal validation.

III. OBJECTIVES

1. To evaluate the accuracy of automated machine learning (AutoML), specifically the Roboflow train, compared to ResUnet, for image segmentation and classification tasks related to focal liver lesions.
2. To test and validate the effectiveness of AutoML-based application through internal validation.

IV. REVIEW OF RELATED LITERATURE

Liver lesions are diagnosed and treated using MRI, CT, and ultrasound. However, it's challenging due to variations in size, shape, and appearance. Deep learning, especially CNN, is being explored to enhance accuracy and efficiency in segmenting and classifying liver lesions from CT scans. CNN can learn features and perform tasks like segmentation and classification, showing promise in medical image analysis.

Manjunath et al. [11] focused on using deep learning, specifically a modified ResUNet model, to automatically segment liver lesions from CT scans. This approach outperformed traditional methods, achieving high accuracy in liver and tumor segmentation. The proposed method demonstrates promise for improving liver disease diagnosis.

Sabir et al. [12] used the 3D-IRCADb01 dataset, ResU-Net architecture is developed for CT scans. It incorporates residual blocks and U-Net structure for improved information extraction. Image pre-processing techniques are applied before input.

N Nanda Prakash et al. [13] gathered liver CT scan images from Kaggle that were used for training. Pre-processing involved region-growing segmentation, and DenseNet CNN was used for training. Real-time test images (10,000 samples) from Government General Hospital Vijayawada were verified with the proposed DenseNet CNN for liver lesion diagnosis.

R. Murugesan & K. Devaki [14] proposed a new semantic segmentation technique called UNet++ for extracting liver lesions from CT images. Additionally, a hybrid approach combining the Chaotic Cuckoo Search algorithm and AlexNet is used for feature extraction and classification of liver lesions. Evaluation is conducted using the LiTS database, and results are measured using the Dice similarity coefficient and correlation coefficient.

Zhou et al. [15] developed a hierarchical CNN framework for detecting and classifying focal liver lesions (FLLs) in multi-phasic CT scans. With 616 nodules, our model outperformed other CNNs and human experts. It categorizes FLLs into malignant/benign and then into specific lesion types.

TABLE 1. CNN-BASED MEDICAL IMAGE ANALYSIS

Authors	**Similarities**	**Difference**	**Accuracy Rate**
Manjunath et al. [11]	Using CNN for liver tumor classification and segmentation in CT scan images.	The ResUNet model was employed for detection, resulting in successful DSC analysis.	98.59%
Sabir et al. [12]	Liver tumor segmentation and lesion recognition from the liver using CNN.	The ResUNet architecture was employed for tumor segmentation within the liver.	99.1%
N Nanda Prakash et al. [13]	Detection of occurrence of liver lesion into the CT scan utilizing CNN	DenseNet CNN model was used for liver lesion diagnosis.	98.34%
R. Murugesan & K. Devaki [14]	Utilized an algorithm to extract features and segment liver lesions.	UNet++ handled segmentation while additional CNN models were utilized for feature extraction.	99.2%
Zhou et al. [15]	Automatic detection and classification of Focal Liver Lesions CNN-based.	Malignant lesions and benign lesions were divided into sets and analyzed.	82.8%

V. METHODOLOGY

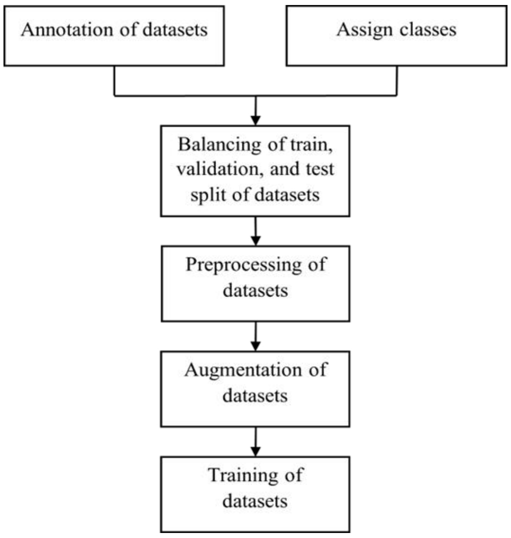

Annotation of datasets
Assign classes
Balancing of train, validation, and test split of datasets
Preprocessing of datasets
Augmentation of datasets
Training of datasets

Fig. 1. Training Process Flow

Figure 1 outlines the training process flow for the dataset, consisting of CECT images.

Initially, the collected CECT images undergo annotation, where FLLs are highlighted and assigned to their respective classes. This annotation process provides the algorithm with the knowledge needed to classify classes based on their differences.

Subsequently, the dataset is balanced and divided into training, test, and validation sets. This standard machine learning practice ensures that the model goes beyond memorizing training data and learns underlying patterns applicable to new, similar data. The training set is utilized to train the machine learning model, allowing it to discern patterns and relationships within the data. The validation set evaluates the model's performance during training, aiding in hyperparameter adjustments to prevent overfitting.

After training and tuning with the training and validation sets, the model's performance is assessed on entirely new and unseen data using the test set. This step provides an unbiased estimate of the model's real-world performance.

Following this, dataset preprocessing occurs, involving steps to create a clean, well-structured, and appropriately formatted dataset that facilitates effective training and generalization of the machine learning model. After preprocessing is the augmentation of the dataset, artificially expanding its size through various transformations to enhance the model's generalization and robustness.

Lastly, in the training phase, the dataset passes through the model to generate predictions. The model compares its predictions with actual labels using the chosen loss function to calculate the error. The optimizer is then employed to adjust the model's weights and biases, minimizing the loss function. Iterative passes through the training dataset (epochs) refining the model, completing the training process.

A. *Automated Machine Learning (AutoML)*

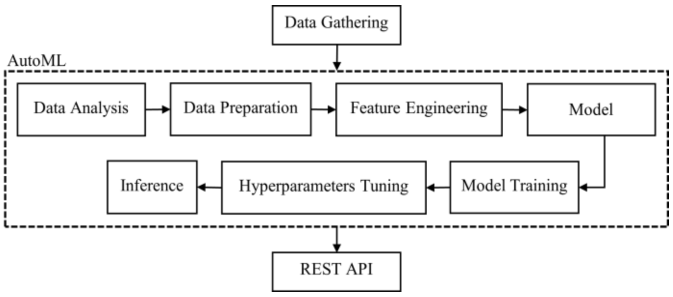


Fig. 2. AutoML Workflow

Figure 2 shows the process of AutoML workflow. AutoML is the process of automating the steps involved in building and deploying machine learning models. It consists of data analysis, data preparation, feature engineering, model selection, hyperparameter tuning, and inference or model deployment. Data analysis and data preparation are the steps where the user defines the problem and collects and labels the data. Feature engineering is the step where the user transforms the data to make it more suitable for the model. Model selection and hyperparameter tuning are the steps where AutoML chooses and optimizes the best model for the data. Inference, or model deployment, is the step where the user deploys the model and makes predictions on new data. This process is iterative and can be improved by using feedback from the inference data.

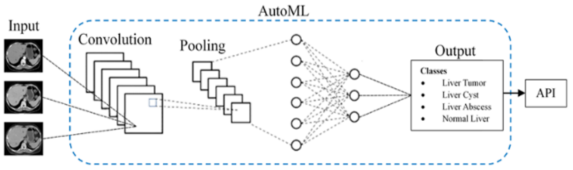


Fig. 3. AutoML Architecture

Figure 3 illustrates the process of AutoML architecture. The first step in the AutoML architecture is to acquire and prepare the data. This involves collecting images, specifically the CECT scan images from different phases. The images then need to be labeled in the proper annotation format. Inside the AutoML process is pooling and convolution are two operations that are commonly used for computer vision tasks. Pooling is a technique that reduces the spatial size of the feature maps by summarizing the presence of features in patches of the feature map. This makes the model more robust to small changes in the input image and reduces the number of parameters and computations. While Convolution is a technique that applies a filter (also called a kernel) to the input image or feature map and produces a new feature map that represents the presence of a specific pattern in the input. It can be used to detect edges, corners, shapes, and other features in the input. The next step is to train a model. This is a process where the computer formulates an algorithm based on the training images to apply to new images that it has never seen before. The model is then evaluated based on how well it performs during predictions. In this case, AutoML automates the process of model selection. It identifies the most suitable model for the dataset and optimizes that model's performance against the data provided. Once a suitable model has been selected and trained, it is deployed to a compute instance and an API endpoint is set up. This API endpoint can be used to receive predictions based on what a model is seeing in the production images.

B. *ResUNet*

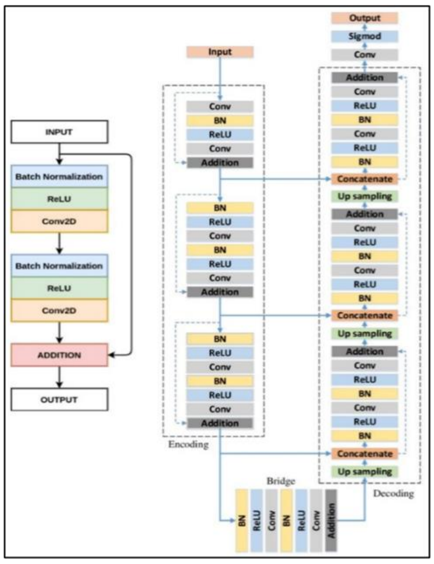


Fig. 4. ResUNet Workflow

Figure 4 shows the process of ResUNet workflow. The ResUNET is a segmentation encoder-decoder architecture that uses a deep residual U-Net. The encoder-decoder design of ResUNet incorporates three-way coding. ResUNet combines the U-Net and Deep Residual Learning architectures. ResUNet is a fully convolutional neural network with few parameters that aims for high performance. It improves on the existing U-Net design. The ResUNet is made up of an encoding network, a decoding network, and a bridge that connects the two. The ResUNet, like U-Net, connects various networks via a gateway, an encoding connection, and a decoding mechanism.

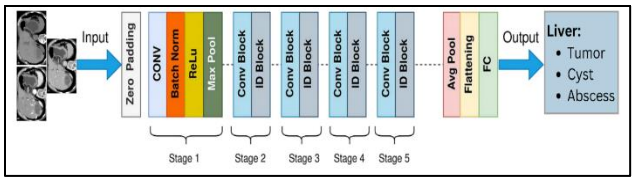


Fig. 5. ResUNet Architecture

Figure 5 illustrates the process of ResUNet architecture. In the detection of FLL, 2D and 3D images in 2D or 3D format are used for feature extraction during the liver segmentation step. Following that, box sampling is useful for dividing image information into small portions. After that, mass segmentation might be used to extract the damaged area of the liver. CNN can be thought of as programmed feature extractors from provided data. While using a pixel vector calculation loses a lot of spatial collaboration between pixels, CNN successfully uses contiguous pixel data to viably down sample the images first by convolution and then uses a forecast layer toward the end. Even tiny tumors are detectable using our CNN-based diagnostic technology.

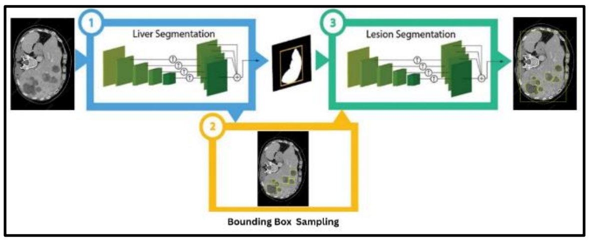


Fig. 6. ResUNet Segmentation

Figure 6 shows the process of ResUNet segmentation process. At the dataset, pruning used CNN (ResUNet) for segmentation of the liver. ResUNet trained on both CECT scans and Liver masks to detect the Region of Interest (ROI) from the neighboring organs. For tumor segmentation of the liver used the ResUNet, after extracting the ROI and trained on CECT scans of the Liver. The ResUNet is a traditional hybrid between the ResNet and UNet models. ResUNet exchanges the convolutional chunks with residual blocks giving us the benefits of both models. Training of the CNN is ease to use with residual each block and avoids the connections between poor and high levels of the network also leads to minor training parameters each residual unit. Each Block of Residual contains the Two 3x3 convolution blocks which are Batch Normalization Layer, ReLU activation layer, and Convolutional layer with Identity mapping. Figure 6 is showing the complete picture of CNN implementation on liver tumor images.

C. Data Collection

All abdominal CECT scan examinations undergo de-identification and anonymous analysis. Images from CECT scans are retrospectively retrieved from radiology department of JJASGH. These images, stored in Digital Imaging and Communications in Medicine (DICOM) format is converted to JPG or JPEG formats by taking a screenshot of the CECT scan, which are compatible with the system. Definitive diagnoses of FLLs are confirmed using pathology reports and/or official CECT scan reports. The collected data includes patients aged 18 years or older who have undergone CECT abdominal scans, with FLLs identified through radiologic or histopathologic evidence.

After the data has been collected, it will be divided into three (3) sets: a.) Training Set, b.) Testing Set, and c.) Validation Set, one for each process in the deployment method. All images obtained for the training set are be used to train in the AI system in distinguishing between liver lesions and normal organ structures. An equal number of normal images without FLLs are randomly selected in a 1:1 ratio to serve as negative controls. This balanced distribution of image types facilitates the AI's training process, enhancing its ability to accurately identify FLLs while reducing the occurrence of false positives.

D. Dataset Classes

a.) Focal Liver Lesions

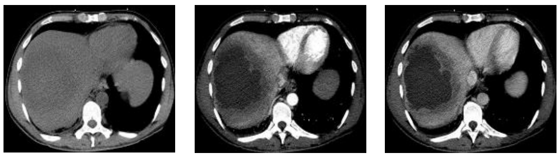

(a) (b) (c)

Fig. 7. Liver Abscess (a) unenhanced, (b) arterial, (c) portal venous.

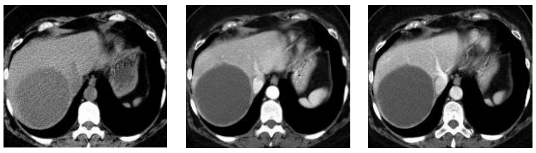

(a) (b) (c)

Fig. 8. Liver Cyst (a) unenhanced, (b) arterial, (c) portal venous

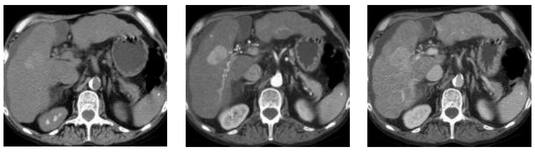

(a) (b) (c)

Fig. 9. Liver Tumor (a) unenhanced, (b) arterial, (c) portal venous

b.) Normal Liver

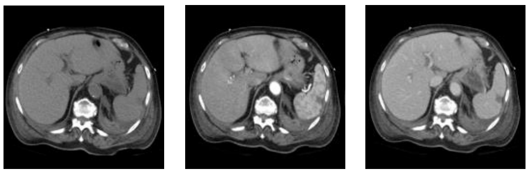

(a) (b) (c)

Fig. 10. Normal Liver (a) unenhanced, (b) arterial, (c) portal venous

The figures show the actual image of the four (4) classifications of the liver covered in this study: Liver Abscess, Liver Cyst, Liver Tumor, and Normal Liver. Each classification is presented in three (3) phases: (a) Unenhanced, (b) Arterial, and (c) Portal Venous. The unenhanced phase establishes a baseline by capturing images before the contrast agent is introduced. Subsequent phases include the arterial phase, assessing vessels and organ abnormalities 20–35 seconds after contrast injection, and the portal venous phase, which provides balanced enhancement for solid organs, bowel, and vascular structures around 70–80 seconds after injection. These phases enable doctors to assess various aspects of the liver organ in a CECT scan, contributing to its classification assessment [16].

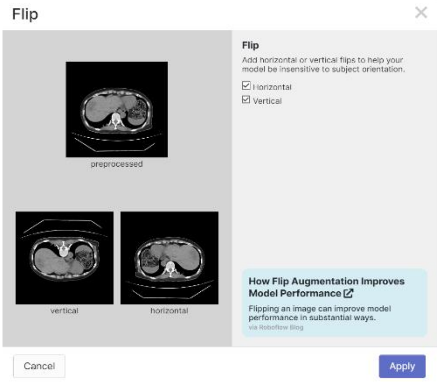
E. Data Distibution

TABLE 2. TOTAL NUMBER OF RAW DATASETS

Focal Liver Lesions	**No. of Datasets**
Tumor	201
Cyst	189
Abscess	189
Normal	252
Total	**831**

a.) Augmentation Techniques

1.) Flip



Flip

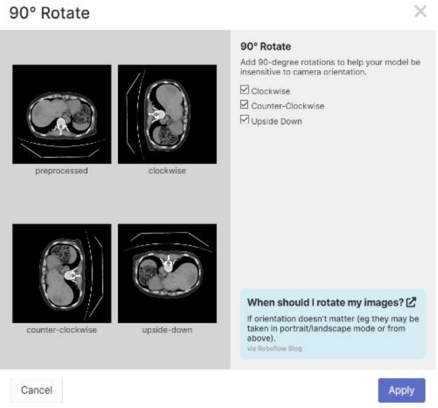
Add horizontal or vertical flips to help your model be insensitive to subject orientation.

☑ Horizontal
☑ Vertical

How Flip Augmentation Improves Model Performance

Flipping an image can improve model performance in substantial ways.
via Roboflow Blog

2.) Rotate



90° Rotate

Add 90-degree rotations to help your model be insensitive to camera orientation.

☑ Clockwise
☑ Counter-Clockwise
☑ Upside Down

When should I rotate my images?

If orientation doesn't matter (eg they may be taken in portrait/landscape mode or from above).
via Roboflow Blog

Fig. 11. Augmentation Techniques

Image augmentation is the process of altering training images to create a synthetic dataset that is larger than your original dataset and, hopefully, improves the model's downstream performance. Two techniques were used for augmentation: flipping and 90-degree rotation. These techniques expanded the dataset while preserving CT scans' inherent contrast characteristics across different phases, which was critical for maintaining pixel-level accuracy in instance segmentation. Other augmentation methods were excluded due to their potential to reduce model accuracy, given the pixel-based nature of instance segmentation and the wide range of contrast levels in CT scans.

TABLE 3. TOTAL NUMBER OF AUGMENTED DATASETS

Training Set	**Testing Set**	**Validation Set**	**Total**
4,868	1,315	613	6,796

Table 2 and 3 show the total number of raw datasets per FLL and the total number of annotated and augmented datasets, respectively, trained in AutoML.

TABLE 4. TOTAL NUMBER OF PATIENTS PER FLL

	Clinical datasets (Liver Atlas, Radiopaedia)	**Institutional datasets (JJASGH)**	**Total number of patients**
Liver Cyst	5	1	6
Liver Abscess	5	1	6
Liver Tumor	15	1	16
Normal Liver	5	3	8
Overall	**30**	**6**	**36**

TABLE 5. SUMMARY OF DATASET DISTRIBUTION

	Training Set	**Testing Set (half of the total number of patients)**	**Validation Set (institutional datasets)**	**Total number of datasets**
Liver Cyst	150 (50 per phases)	9 (3 patients x 3 phases)	30 (1 patient with 10 trials x 3 phases)	189
Liver Abscess	150 (50 per phases)	9 (3 patients x 3 phases)	30 (1 patient with 10 trials x 3 phases)	189
Liver Tumor	150 (50 per phases)	21 (7 patients x 3 phases)	30 (1 patient with 1 0 trials x 3 phases)	201
Normal Liver	150 (50 per phases)	12 (4 patients x 3 phases)	90 (3 patient with 10 trials x 3 phases)	252
Overall	**600**	**51**	**180**	**831**

Table 4 and Table 5 shows the total number of patients per FLL, and the summary of dataset distribution used as training set, testing set, and validation set.

F. Testing and Evaluation

This study necessitates consultation with doctors to review and validate datasets, ensuring the accuracy of information fed into the programming. Doctors' input is crucial to prevent misdiagnosis and guarantee high-quality results. In the testing of the device, it utilized the internal test set under medical expert supervision to verify the accuracy and effectiveness of both the device and the website.

The device is specifically designed to detect focal liver lesions by analyzing segmented CECT scan images across different phases: unenhanced, arterial, and portal venous. CECT scan images that are used are retrospectively retrieved from the Picture Archiving and Communication System (PACS) of Justice Jose Abad Santos General Hospital (JJASGH) and confirmed by official CECT scan report.

Throughout the two-month deployment, all images underwent training and testing. The generated results were based on the device's findings, but the final diagnosis remained subject to approval by a Radiologist or Gastroenterologist.

G. ResUNet Metrics

ResUNet metrics are used to measure how accurate the performance of the model is. Some metric that can be obtained through this study is IoU. This helps evaluate how well the model separates and identifies different instances of objects within an image, making it a vital metric for tasks requiring fine-grained object separation. As with any machine learning model, image classification models necessitate a range of metrics to gauge their accuracy. Among these metrics, IoU (Intersection over Union) holds significance, particularly in assessing segmentation models. Referred to as Jaccard's Index, IoU quantifies the model's ability to differentiate objects from their backgrounds within an image. It finds widespread application in various computer vision domains, including medical imaging.

a.) Intersection over Union (IoU)

Intersection over Union (IOU) serves as a key performance measure employed in assessing the accuracy of annotation, segmentation, and object detection algorithms. It measures the degree of overlap between the predicted bounding box or segmented area and the actual bounding box or annotated region within a dataset. Essentially, IOU is crucial for quantifying the agreement between predicted and ground truth regions in tasks such as object detection and segmentation. IoU, or Intersection over Union, calculates the ratio of the intersection of two boxes' areas to their combined areas. Both the ground truth bounding box and the predicted bounding box cover the union area. A high IoU score indicates substantial overlap between the predicted and ground truth boxes, while a low overlap leads to a lower IoU score. A score of 1 signifies a perfect match between the predicted box and the ground truth box, while a score of 0 indicates no overlap between the boxes. IoU values below 0.5 are typically considered poor. Values between 0.5 and 0.75 are often considered average.

Lastly, values above 0.75 are generally considered excellent. This indicates a high degree of overlap between the predicted and ground truth bounding boxes, suggesting accurate image classifiction and segmentation [17].

To calculate the overlap between the ground-truth bounding box and the predicted bounding box in the numerator, it is mathematically expressed as:

$$Intersection\ over\ Union\ (IoU) = \frac{|A \cap B|}{|A| \cup |B|}$$

But for binary classification, it is written as:

$$Intersection\ over\ Union\ (IoU) = \frac{TP}{TP + FN + FP}$$

Where:
TP = True Positive
FN = False Negative
FP = False Positive

H. *Roboflow Train Metrics*

Precision, recall, and mAP are essential metrics for evaluating instance segmentation models, each addressing distinct aspects of performance. Precision measures the model's accuracy in identifying the correct instances, prioritizing the avoidance of false positives. In contrast, recall measures the model's ability to detect all relevant instances, prioritizing the minimization of false negatives. mAP combines both measures to provide a comprehensive assessment of overall performance. The choice of which metric to prioritize depends on the specific application and the relative costs of false positives and false negatives.

a.) Mean Average Precision (mAP)

The measure of accuracy for roboflow train models is evaluated using mean average precision (mAP). The higher the mAP, the better the model is performing. While a high confidence result value or high mAP is generally a good sign, it's crucial to validate the model's predictions in the context of the specific application. The mean of average precision (mAP) values is calculated over recall values from 0 to 1. Its score of 0 indicates retrieval of no relevant objects. For all other scenarios, MAP ranges between 0 and 1, with a higher score reflecting excellent ranking performance, approaching perfection as it nears 1 [18].

The mAP is calculated by finding Average Precision (AP) for each class and then average over a number of classes.

$$mAP = \frac{1}{N} \sum_{k=1}^{k=N} AP_k$$

Where:
AP_k = Average Precision of class k
N = Number of Classes

b.) F1 Score

The F1 score is a key metric in machine learning used to evaluate a model's accuracy. It integrates the precision and recall scores of a model. Precision quantifies the proportion of correct "positive" predictions out of all positive predictions made by the model. Recall, on the other hand, assesses the proportion of actual positive class samples in the dataset that the model correctly identified. The F1 score is calculated as the harmonic mean of precision and recall, meaning that maximizing the F1 score requires optimizing both precision and recall. Consequently, the F1 score is widely favored by researchers for model evaluation alongside accuracy[19].

The F1 score is calculated as the harmonic mean of the precision and recall scores, as shown below. It ranges from 0-100%, and a higher F1 score denotes a better-quality classifier.

For Precision, the formula used is:

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

And for Recall, the formula used is:

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

Thus, the formula used for F1 score is:

$$F1\ Score = \frac{2}{\frac{1}{Precisio} + \frac{1}{Recall}} = \frac{2\ \times Precision\ \times Recall}{Precision + Recall}$$

In terms of the basic four elements of the confusion matrix, by replacing the expressions for precision and recall scores in the equation above, the F1 score can also be written as follows:

$$F1\ Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Where:
TP = True Positive
FP = False Positive
FN = False Negative

c.) Confusion Matrix

A confusion matrix is a numerical representation that indicates where a model makes errors. It provides a class-wise distribution of the predictive performance of a classification model, mapping predictions to the original classes of the data. Confusion matrices can only be used in supervised learning frameworks where the output distribution is known. They enable the calculation of a classifier's accuracy, both overall and class-wise, and help compute other important metrics used to evaluate models. When computed for the same test set using different classifiers, a confusion matrix can compare their relative strengths and weaknesses. This comparison can inform decisions about combining classifiers (ensemble learning) to achieve optimal performance [20].

I. *Accuracy Testing*

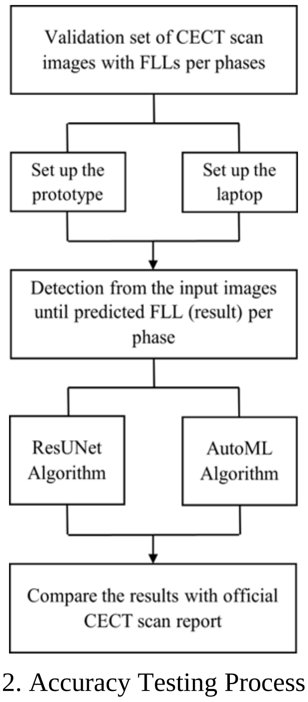

Validation set of CECT scan images with FLLs per phases
Set up the prototype
Set up the laptop
Detection from the input images until predicted FLL (result) per phase
ResUNet Algorithm
AutoML Algorithm
Compare the results with official CECT scan report

Fig. 12. Accuracy Testing Process Flow

Figure 12 depicted designed to evaluate the accuracy of the prototype, focusing specifically on the detection of FLLs.

Researchers conducted a comparative analysis between the CECT image detection results produced by the device and based on the official CECT scan report, validated by a gastroenterologist, to assess the accuracy. The evaluator indicated concordance between the prototype's diagnosis and that of the official CECT scan report by marking a check symbol (✓), while any discrepancies were indicated by an "X." This analysis aimed to assess the prototype's accuracy in segmenting and classifying lesions within CECT scan images. Moreover, the comparison was conducted to evaluate the accuracy of the AutoML detection algorithm, which is the primary focus of this study, in contrast with the ResUnet algorithm, a prevalent method in medical image segmentation. This comparison aimed to ascertain the optimal technique for lesion classification in CECT scans. Additionally, the study sought to determine whether employing solely the AutoML algorithm on both the prototype device and the laptop would yield more precise results in detecting FLLs, thereby prompting the comparison between the two devices. The comparisons conducted in this study serve to comprehensively evaluate algorithmic performance, identify optimal techniques, explore device-specific considerations, and provide practical insights for clinical application, thereby advancing the field of medical image analysis.

Drawing from insights gained in previous study of Ka Wing Wan et, al. [21] and considering the number of datasets utilized in the current research, it is suggested that the target accuracy for this study be set at 86%. The formula used to obtain the percent accuracy (%) of the device per FLL is:

$$\% = \frac{Matched\ Diagnosis\ Score\ per\ FLL}{No.of\ Patients\ x\ No.\ of\ Phase\ per\ FLL}\ x\ 100$$

The formula used to get the overall percent accuracy of the device (A) is:

$$A = \frac{\frac{Matched\ Diagnosis\ Score\ per\ FLL}{No.of\ Patients\ x\ No.\ of\ Phase\ per\ FLL}}{No.of\ FLL}\ x\ 100$$

J. Accuracy Validation

A validation set which comprises of one hundred-eighty (180) CECT images of patients is utilized. Dividing into one (1) patient per FLL and three (3) patients for normal liver, with ten (10) sample for each phase, making a total of thirty (30) images per patient. The purpose of this dataset is to assess accuracy and speed in preparation for the prototype's actual deployment in clinical settings.

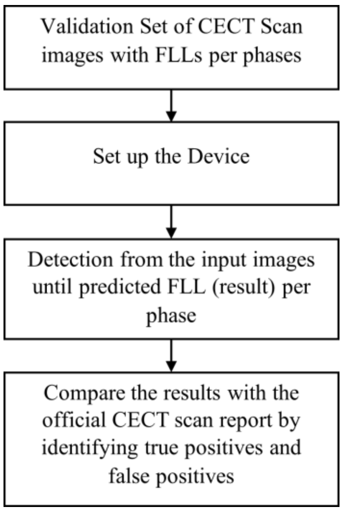

Validation Set of CECT Scan images with FLLs per phases
Set up the Device
Detection from the input images until predicted FLL (result) per phase
Compare the results with the official CECT scan report by identifying true positives and false positives

Fig. 13. Validation of the Device Process Flow

Figure 13 illustrates the procedural flow for validating the accuracy of the prototype by conducting benchmarking accuracy evaluations between the device's detection outcomes and the diagnoses rendered by Gastroenterologists across various tests administered with their patients. The following steps outline the activities to be undertaken to thoroughly evaluate the prototype's accuracy:

1. Verification of Accuracy: The individuals involved in the deployment process will operate the device to ensure its accuracy.
2. Comparative Analysis: The detection results obtained from the device's validation set will be compared with the diagnoses provided by the gastroenterologists.
3. Assessment Criteria: If the device’s results correspond with the diagnoses by the gastroenterologists, it will be deemed as a true positive (TP) analysis; otherwise, it will be considered a false positive (FP). It is crucial to note that for each phase, the prototype’s result must align with the professionals’ diagnosis. Any mismatch in a phase will result in a false positive (FP) analysis.

This testing protocol is specifically designed to evaluate the prototype's accuracy in accurately segmenting and identifying lesions within CECT scan images processed by the system.

For the analysis:

True positive (TP) = The number of CECT images from the validation set that were correctly identified in each phase per FLL when compared to the diagnoses of the gastroenterologists.

False positive (FP) = The number of CECT images from the validation set that were incorrectly identified in each phase per FLL when compared to the diagnoses of the gastroenterologists.

To obtain the accuracy of the validation, proportion of true positive and true negative must be calculated. Mathematically, this can be stated as:

$$Accuracy\ = \frac{TP}{TP + FP}$$

K. Statistical Treatment

For the accuracy which is also part of the first objective the statistical treatment that is used are the Mann–Whitney U test and Wilcoxon W test. The Mann–Whitney U test and the Wilcoxon W test are both non-parametric tests used to compare differences between two independent samples, but they have some differences in their application and the specifics of what they measure. The Mann-Whitney U test is employed to compare differences between two independent groups when the dependent variable is either ordinal or continuous but not normally distributed, focusing on the ranks of the observations. [22]. The Mann–Whitney U test is often considered the nonparametric equivalent of t-test for independent samples, but this comparison may be somewhat oversimplified [23]. The Wilcoxon W test on the other hand, is the same with Mann-Whitney U test but it focuses on the sum of ranks. It is a nonparametric statistical test that compares two paired groups. The tests essentially calculate the difference between sets of pairs and analyze these differences to establish if they are statistically significantly different from one another [24]. The comparison of accuracy is part of the first objective, and it is divided into two (2) categories: Accuracy of AutoML using Wi-Fi (Prototype versus Laptop) and Accuracy of Laptop (AutoML versus ResUNet).

The Mann-Whitney U test is used to compare the accuracy of AutoML in prototype versus laptop. It compares the accuracy of AutoML on different hardware platforms without making strict assumptions about the distribution of the accuracy scores. There are two separate sets of accuracy scores from running the AutoML system on a prototype and on a laptop. This makes Mann-Whitney U test suitable method for this type of comparison.

For comparing the accuracy of AutoML and ResUNet run on a laptop, the Wilcoxon W test is used. The Wilcoxon W test does not assume that the data follows a normal distribution. Since the accuracy measurements from AutoML and ResUNet might not be normally distributed, this test is more appropriate than parametric tests such as the paired t-test. The Wilcoxon W test is particularly suitable for non-normally distributed, paired data, and it is less prone to the effects of outliers and small sample sizes. This makes it a reliable choice for assessing whether there is a statistically significant difference in accuracy between the AutoML and ResUNet algorithms when evaluating using a laptop.

The general formula for the Mann–Whitney U test is:

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i$$

Where:
U = Mann-Whitney U test
N_1 = sample size one
N_2= Sample size two
R_i = Rank of the sample size
The general formula for the Wilcoxon W test is:

$$z = \frac{W - 0.5}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}$$

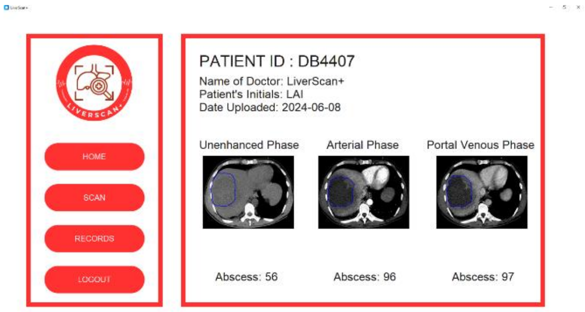
$$W = \left|\sum [sgn\,(x_2 - x_1) \cdot R]\right|$$

Where:
W = Wilcoxon W test
z = z-score

VI. RESULTS AND DISCUSSION

A. Dataset Sample Predictions

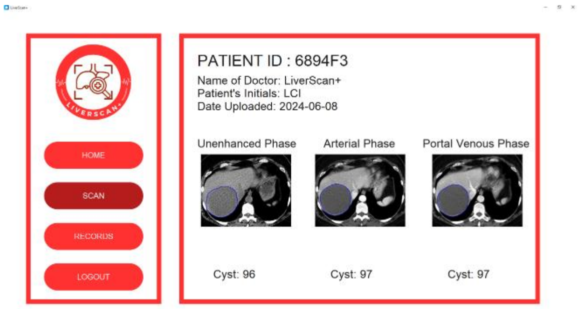
1. Focal liver Lesion



PATIENT ID : DB4407
Name of Doctor: LiverScan+
Patient's Initials: LAI
Date Uploaded: 2024-06-08

Unenhanced Phase	Arterial Phase	Portal Venous Phase
Abscess: 56	Abscess: 96	Abscess: 97

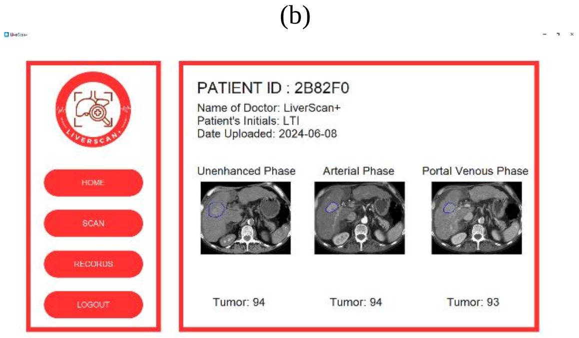
(a)



PATIENT ID : 6894F3
Name of Doctor: LiverScan+
Patient's Initials: LCI
Date Uploaded: 2024-06-08

Unenhanced Phase	Arterial Phase	Portal Venous Phase
Cyst: 96	Cyst: 97	Cyst: 97

(b)

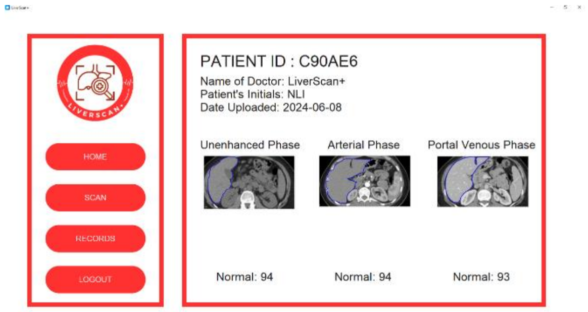


PATIENT ID : 2B82F0
Name of Doctor: LiverScan+
Patient's Initials: LTI
Date Uploaded: 2024-06-08

Unenhanced Phase	Arterial Phase	Portal Venous Phase
Tumor: 94	Tumor: 94	Tumor: 93

(c)

2. Normal Liver



PATIENT ID : C90AE6
Name of Doctor: LiverScan+
Patient's Initials: NLI
Date Uploaded: 2024-06-08

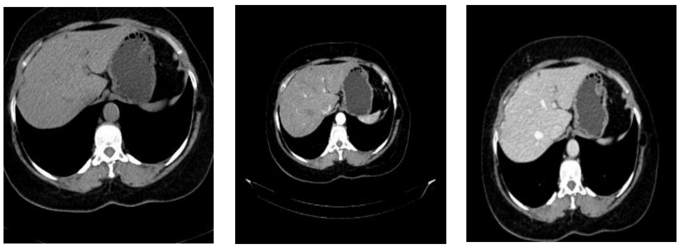
Unenhanced Phase	Arterial Phase	Portal Venous Phase
Normal: 94	Normal: 94	Normal: 93

(d)

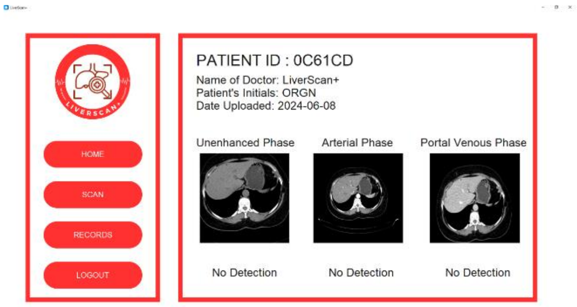
Fig. 14. Test Images (a) Abscess, (b) Cyst, (c) Tumor, (d) Normal

Figure 14 shows that the device was tested using images obtained from clinical datasets. The testing set includes CECT scans of normal livers as well as livers with tumors, cysts, and abscesses. Despite variations in image quality, the system correctly detected focal liver lesions across these different conditions, demonstrating its robustness and reliability.

3. Random Images and Other Type of Organ Disease



(a) (b) (c)

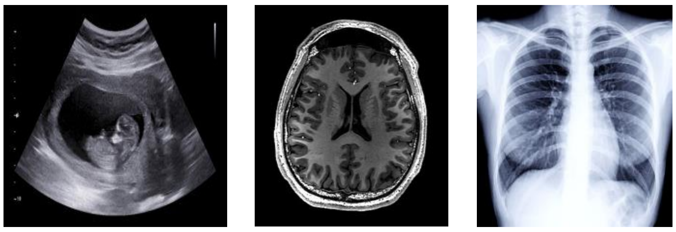


PATIENT ID : 0C61CD
Name of Doctor: LiverScan+
Patient's Initials: ORGN
Date Uploaded: 2024-06-08

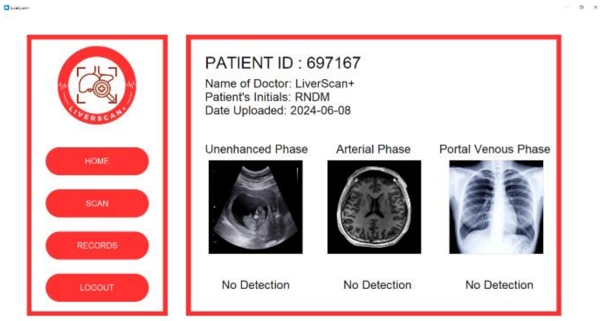
Unenhanced Phase	Arterial Phase	Portal Venous Phase
No Detection	No Detection	No Detection

(d)

Fig. 15. Other Organ Disease (a) unenhanced, (b) arterial, (c) portal venous, (d) system's prediction Note: From Radiopaedia.org, by Siocha C., 2023, (http://surl.li/uisjt)



(a) (b) (c)



PATIENT ID : 697167
Name of Doctor: LiverScan+
Patient's Initials: RNDM
Date Uploaded: 2024-06-08

Unenhanced Phase	Arterial Phase	Portal Venous Phase
No Detection	No Detection	No Detection

(d)

Fig. 16. Random Images (a) Ultrasound, (b) MRI, (c) X-Ray, (d) system's prediction

Note: (a) From Jewel Women’s Center, 2022, (https://jewelwc.org/blog/early-pregnancy-ultrasound/) (b) From Wikipedia; Wikimedia Foundation, (https://en.wikipedia.org/wiki/Magnetic_resonance_imaging_of_the_brain) (c) From Harvard Health; Harvard Health, 2019, (https://www.health.harvard.edu/staying-healthy/x-rays-a-to-z)

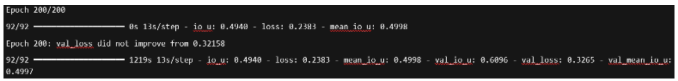
Figure 15 and 16 display the image that demonstrates the model's specificity by showing its performance with non-CECT scans, such as ultrasound, MRI, and X-ray images of the liver and other organs. The model fails to detect lesions in these random images, highlighting its training focus on CECT scan images. This specificity ensures the model's reliability and effectiveness in diagnosing focal liver lesions exclusively with CECT scans.

B. *Evaluation Metrics*

Evaluation metrics were employed to evaluate the performance of models in tasks like image classification and segmentation.

a) ResUNet Metrics

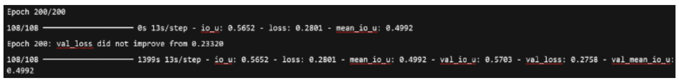
The researchers used the Intersection over Union (IoU) metric to evaluate the performance of training datasets in the ResUNet model, which determines the overlap between predicted and ground truth segmentation masks, thus providing valuable data on the model's accuracy. IoU scores are classified as follows: IoU = 1 indicates a perfect match, IoU = 0 indicates no overlap, IoU < 0.5 signifies poor overlap, 0.5 ≤ IoU < 0.75 denotes average overlap, and IoU > 0.75 represents excellent overlap.



```
Epoch 200/200
92/92 ━━━━━━━━━━━━━━━━━━━━ 0s 13s/step - io_u: 0.4940 - loss: 0.2383 - mean_io_u: 0.4998
Epoch 200: val_loss did not improve from 0.32158
92/92 ━━━━━━━━━━━━━━━━━━━━ 1219s 13s/step - io_u: 0.4940 - loss: 0.2383 - mean_io_u: 0.4998 - val_io_u: 0.6096 - val_loss: 0.3265 - val_mean_io_u:
0.4997
```

Fig. 17. IoU of Unenhanced Phase

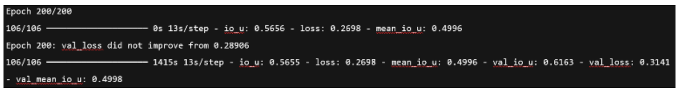
Figure 17 displays the IoU value for the unenhanced phase across all classes. The findings reveal that the IoU value peaks at 0.4940 in the 200th epoch. This indicates that there is a poor overlapping between the predicted box and the ground truth box – indicating that the there is a potential limitation in performance, signifying possible difficulties in accurately detecting certain inputs.



```
Epoch 200/200
108/108 ━━━━━━━━━━━━━━━━━━━━ 0s 13s/step - io_u: 0.5652 - loss: 0.2801 - mean_io_u: 0.4992
Epoch 200: val_loss did not improve from 0.23320
108/108 ━━━━━━━━━━━━━━━━━━━━ 1399s 13s/step - io_u: 0.5652 - loss: 0.2801 - mean_io_u: 0.4992 - val_io_u: 0.5703 - val_loss: 0.2758 - val_mean_io_u:
0.4992
```

Fig. 18. IoU of Arterial Phase

Figure 18 shows the IoU value during the training of the Arterial Phase using ResUNet, encompassing all classes. The results indicate that the IoU value reaches 0.5652 by the 200th epoch. This indicates that there is an average overlapping between the predicted box and the ground truth box. This value is deemed average in terms of accurately detecting the input.



```
Epoch 200/200
106/106 ━━━━━━━━━━━━━━━━━━━━ 0s 13s/step - io_u: 0.5656 - loss: 0.2698 - mean_io_u: 0.4996
Epoch 200: val_loss did not improve from 0.28906
106/106 ━━━━━━━━━━━━━━━━━━━━ 1415s 13s/step - io_u: 0.5655 - loss: 0.2698 - mean_io_u: 0.4996 - val_io_u: 0.6163 - val_loss: 0.3141
- val_mean_io_u: 0.4998
```

Fig. 19. IoU of Portal Venous Phase

Figure 19 illustrates the IoU value during the training phase in the Portal Venous Phase, encompassing all classes. The results indicate that the IoU value reaches 0.5655 by the 200th epoch. This indicates that there is an average overlapping between the predicted box and the ground truth box. This reflects the average classification performance across the input classes.

TABLE 6. SUMMARY OF IOU VALUES PER PHASE

Phases	**IoU Value in 200th Epoch**	**Conclusion**
Unenhanced	0.4940	Poor
Arterial	0.5652	Average
Portal Venous	0.5655	Average

Table 6 summarizes the IoU values from training using the ResUNet algorithm. The results indicate that the algorithm's performance in accurately detecting the classes ranged from poor to average. Three models were created based on the different phases to avoid any confusion during training process.

b) Roboflow Train Metrics

The researchers evaluated AutoML training performance using the mean Average Precision (mAP) statistic, which encompasses both precision and recall, enabling a comprehensive evaluation of instance segmentation methods. By continuously optimizing dataset performance based on mAP scores, they ensure a stable training model and enhance the AutoML model's ability to detect and segment focal liver lesions. The mAP classification criteria are as follows: mAP = 0 indicates no relevant object retrieval, 0 < mAP < 1 signifies average retrieval, and mAP = 1 denotes excellent ranking performance and retrieval of relevant objects.

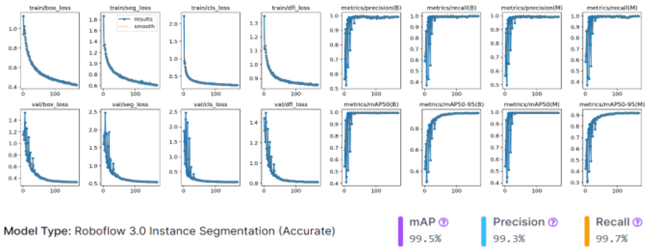


Fig. 20. mAP of Unenhanced Phase

The depicted graph displays the mAP outcome for the unenhanced phase of training in Roboflow. It illustrates a trend where mAP is nearing an optimal value of 1, indicating precise detection of input classes. Additionally, the numerical data confirms a mAP percentage of 99.5%, reinforcing the accuracy of the detection.

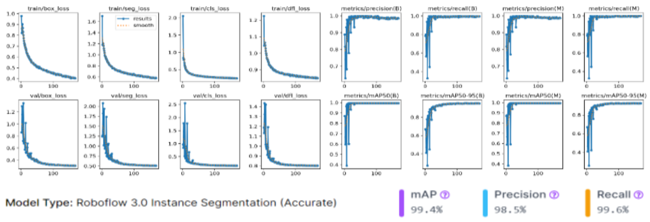


Fig. 21. mAP of Arterial Phase

The graph depicts the mAP outcome for the arterial phase trained using Roboflow. It demonstrates a trend where the line is converging towards 1, suggesting precise detection of classes. This is further validated by the numerical data, which confirms a mAP value of 99.4%.

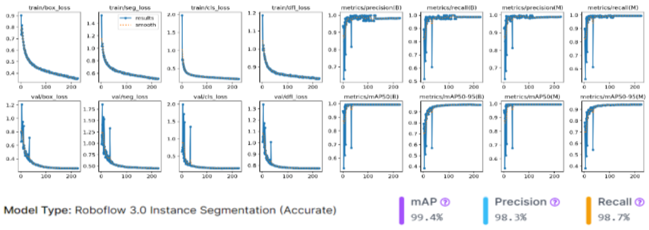


Fig. 22. mAP of Portal Venous Phase

The graph illustrates the Mean Average Precision (mAP) results for the portal venous phase trained with Roboflow. It indicates a trend where the line is approaching to

1, implying accurate detection of classes. This is supported by the numerical data, confirming a mAP value of 99.4%.

TABLE 7. SUMMARY OF MAP PER PHASES

Phases	mAP	Conclusion
Unenhanced	0.995	Excellent
Arterial	0.994	Excellent
Portal Venous	0.994	Excellent

Table 7 summarizes the mAP values obtained from training with AutoML. The values were close to 1, indicating excellent performance in accurately detecting the classes. Three models were created based on the different phases to avoid any confusion during training process.

1. F1 Score

The F1 score is calculated as the harmonic mean of the precision and recall scores. It ranges from 0-100%, and a higher F1 score denotes a better-quality classifier. A score approaching or reaching one by the end of the epoch indicates a highly accurate model.

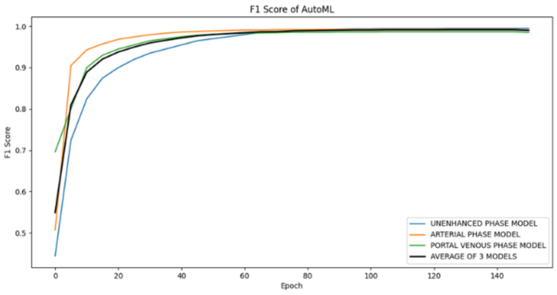


Fig. 23. F1 Score Chart for Combined Datasets Across All Phases

Figure 23 presents the F1 score chart for the combined datasets across all phases using AutoML. The trend lines for each phase—unenhanced, arterial, and portal venous—converge towards one at the final epoch. This indicates a highly accurate model in predicting true positives. Three models were created based on the different phases to avoid any confusion during training process.

2. Confusion Matrix

If the diagonal scores in confusion matrix are around 98 instead of 100, it indicates that there might be a very small number of misclassifications or errors in model's predictions, despite it being highly accurate.

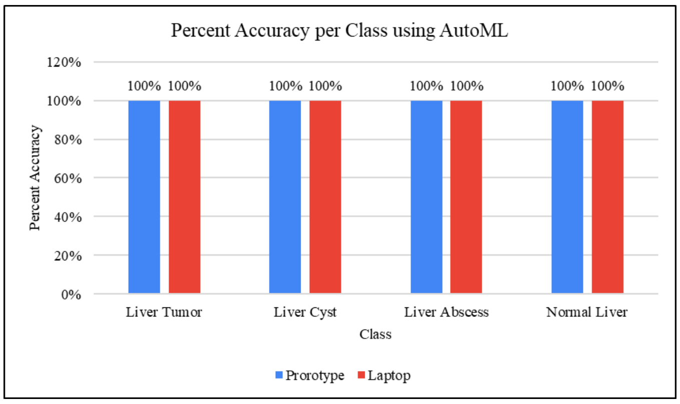


Fig. 24. Confusion Matrix for Unenhanced Phase

Figure 24 shows the confusion matrix for unenhanced phase. The diagonal scores produced a unanimous result of 99.50 that suggested a very small number of misclassifications or errors in model's predictions.

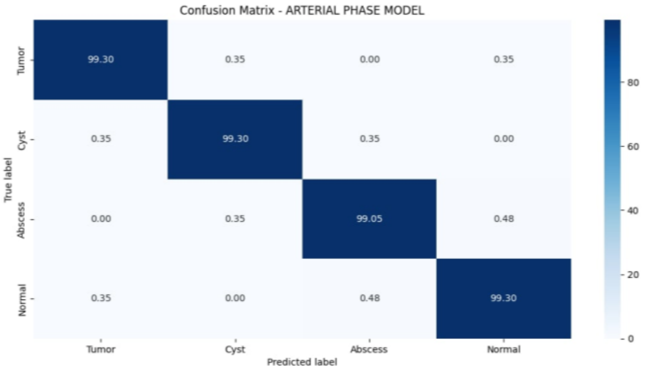


Fig. 25. Confusion Matrix for Arterial Phase

Figure 25 shows the confusion matrix for arterial phase. The diagonal scores produced a result of 99.30 for tumor, cyst and normal while 99.05 in abscess. These scores suggested a very small number of misclassifications or errors in model's predictions.

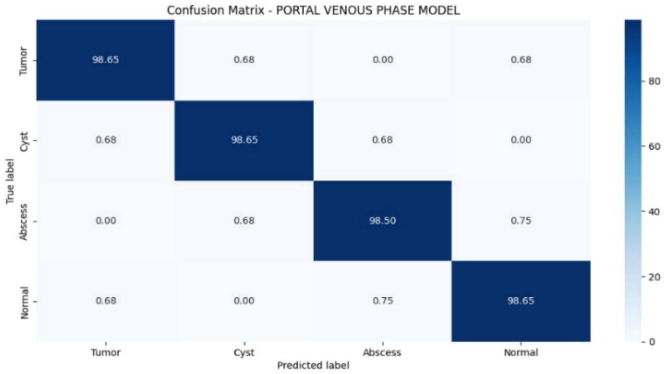


Fig. 26. Confusion Matrix for Portal Venous Phase

Figure 26 shows the confusion matrix for portal venous phase. The diagonal scores produced a result of 98.65 for tumor, cyst and normal while 99.50 in abscess. These scores suggested a very small number of misclassifications or errors in model's predictions.

C. *Results from the Testing Set*

The results of this process were obtained using a testing dataset, which included accuracy tests for both the prototype and the system detection. The accuracy comparison involved assessing both the laptop and prototype devices to determine any significant differences in detecting FLLs. Furthermore, this study compares AutoML, utilized here, with ResUNet, commonly employed in medical imaging for image segmentation.

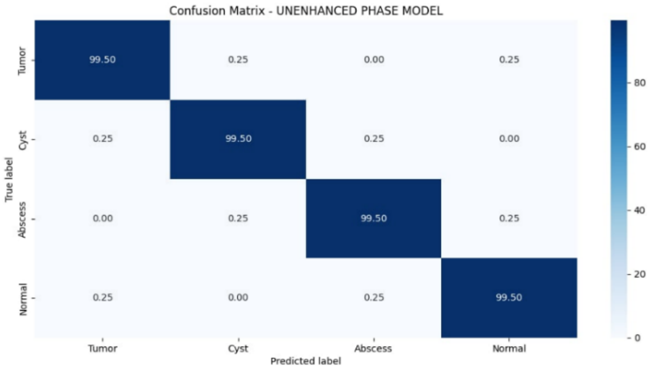


Fig. 27. Accuracy Test Result of Prototype and Laptop in Detecting FLLs using AutoML

Figure 27 shows the percentage accuracy results for each class using a prototype and a laptop that utilize the AutoML algorithm. According to the raw data from the accuracy tests of both the prototype and the laptop using AutoML, both devices achieved a 100% accuracy rate. This indicates that across various classes and three different phases of the testing set, all inputs matched the official CECT scan report.

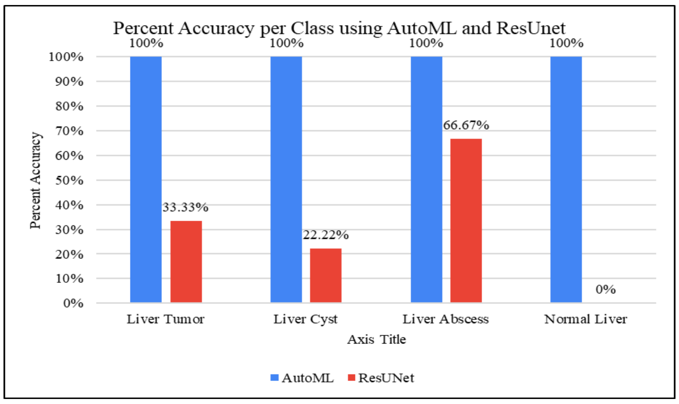


Fig. 28. Accuracy Test Result of AutoML and ResUNet Algorithms in Detecting FLLs

Figure 28 shows the percentage accuracy results for each class using a laptop that utilizes both the AutoML and ResUNet algorithms. Based on the raw data from the accuracy tests of the laptop using AutoML and ResUNet, the AutoML algorithm achieved a 100% accuracy rate in detecting FLLs when compared to the official CECT scan report. In contrast, the ResUNet algorithm exhibited a 30.56% accuracy rate in class detection.

VII. CONCLUSION

The research paper provides a comprehensive exploration of the application of Convolutional Neural Networks (CNNs) for the analysis of contrast-enhanced triphasic CT scans in the context of liver lesion detection. The study's findings reveal that automated machine learning (AutoML), particularly the Roboflow train, demonstrates notable potential in image segmentation and classification tasks associated with focal liver lesions (FLLs). Notably, in determining the accuracy of the AutoML using Roboflow Train, the data achieved a 100% accuracy rate based on the matched diagnosis of the AutoML to the official CECT scan report. This indicates that the AutoML outperforms the ResUNet algorithm with regards to image segmentation and classification of FLLs. The data was gathered through an accuracy test for the prototype in detecting FLLs using AutoML. Through rigorous evaluation metrics such as mean Average Precision (mAP) and confusion matrices, the study showcases the effectiveness of the Roboflow train in accurately detecting and segmenting FLLs, with promising results in achieving high accuracy levels. The comparative analysis between AutoML and the ResUNet model provides valuable insights into the potential of AutoML-based applications for medical image analysis, highlighting its comparable accuracy to traditional classifiers and its potential for future adoption in clinical settings. These findings contribute to the advancement of computer-aided diagnosis in the medical field, emphasizing the significance of accurate detection and classification of liver lesions for improved patient care and management.

VIII. REFERENCES

[1] Mohdsanadzakirizvi Sanad, “Image Classification Using CNN (Convolutional Neural Networks),” Analytics Vidhya, Feb. 18, 2020. https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/.

[2] “What Is a Convolutional Neural Network? | 3 things you need to know,” Mathworks.com, 2024. https://www.mathworks.com/discovery/convolutional-neural-network.html.

[3] Hagar Louye Elghazy and Mohamed Waleed Fakhr, “Dual‐ and triple‐stream RESUNET/UNET architectures for multi‐modal liver segmentation,” Iet Image Processing, vol. 17, no. 4, pp. 1224‐1235, Dec. 2022, doi: https://doi.org/10.1049/ipr2.12708.

[4] S. Madeleine, “Convolutional Neural Networks architectures for classification in medical imaging,” IMAIOS, 2021. https://www.imaios.com/en/resources/blog/classification-of-medical-images-the-most-efficient-cnn-architectures.

[5] J. Zhang, D. Li, L. Wang, and L. Zhang, “Auto Machine Learning for Medical Image Analysis by Unifying the Search on Data Augmentation and Neural Architecture.” Available: https://arxiv.org/pdf/2207.10351.pdf

[6] R. V. Manjunath and Karibasappa Kwadiki, “Automatic liver and tumour segmentation from CT images using Deep learning algorithm,” Results in Control and Optimization, vol. 6, pp. 100087–100087, Mar. 2022, doi: https://doi.org/10.1016/j.rico.2021.100087.

[7] H. Ryu, Seung Yeon Shin, Jae Young Lee, Kyoung Mu Lee, Hyo jin Kang, and J. Yi, “Joint segmentation and classification of hepatic lesions in ultrasound images using deep learning,” European Radiology, vol. 31, no. 11, pp. 8733–8742, Apr. 2021, doi: https://doi.org/10.1007/s00330-021-07850-9.

[8] “Build Vision Models with Roboflow | Roboflow Docs,” Roboflow.com, Mar. 06, 2024. https://docs.roboflow.com/?ref=blog.roboflow.com&fbclid=IwAR04B7rYDkp0K.

[9] D. Jha et al., “ResUNet++: An Advanced Architecture for Medical Image Segmentation,” arXiv.org, 2019. https://arxiv.org/abs/1911.07067.

[10] “XAutoML: A Visual Analytics Tool for Understanding and Validating Automated Machine Learning | ACM Transactions on Interactive Intelligent Systems,” ACM Transactions on Interactive Intelligent Systems, 2023. https://dl.acm.org/doi/10.1145/3625240.

[11] R. V. Manjunath, Anshul Ghanshala, and Karibasappa Kwadiki, “Deep learning algorithm performance evaluation in detection and classification of liver disease using CT images,” *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 2773–2790, May 2023, doi: https://doi.org/10.1007/s11042-023-15627-z.

[12] Muhammad Waheed Sabir *et al.*, “Segmentation of Liver Tumor in CT Scan Using ResU-Net,” *Applied sciences*, vol. 12, no. 17, pp. 8650–8650, Aug. 2022, doi: https://doi.org/10.3390/app12178650.

[13] N Nanda Prakash, V. Rajesh, Syed Inthiyaz, S. D. P, and Sk Hasane Ahammad, “A DenseNet CNN-based liver lesion prediction and classification for future medical diagnosis,” Scientific African, vol. 20, pp. e01629–e01629, Jul. 2023, doi: https://doi.org/10.1016/j.sciaf.2023.e01629.

[14] R. Murugesan and K. Devaki, “Liver Lesion Detection Using Semantic Segmentation and Chaotic Cuckoo Search Algorithm,” *Information Technology and Control*, vol. 52, no. 3, pp. 761–775, Sep. 2023, doi: https://doi.org/10.5755/j01.itc.52.3.34032.

[15] J. Zhou et al., “Automatic Detection and Classification of Focal Liver Lesions Based on Deep Convolutional Neural Networks: A Preliminary Study,” Frontiers in Oncology, vol. 10, Jan. 2021, doi: https://doi.org/10.3389/fonc.2020.581210.

[16] M. P. Hartung, A. Brown, and M. P. Hartung, “Abdominal CT: Phases,” *Life in the Fast Lane • LITFL*, Feb. 02, 2024. https://litfl.com/abdominal-ct-phases/ (accessed Jun. 09, 2024).

[17] “Intersection over Union (IoU): Definition, Calculation, Code,” *V7labs.com*, 2024. https://www.v7labs.com/blog/intersection-over-union-guide (accessed Jun. 09, 2024).

[18] “Mean Average Precision (mAP) Explained: Everything You Need to Know,” V7labs.com, 2024. https://www.v7labs.com/blog/mean-average-precision (accessed Jun. 09, 2024).

[19] “F1 Score in Machine Learning: Intro & Calculation,” V7labs.com, 2024. https://www.v7labs.com/blog/f1-score-guide (accessed Jun. 09, 2024).

[20] “Confusion Matrix: How To Use It & Interpret Results [Examples],” V7labs.com, 2024. https://www.v7labs.com/blog/confusion-matrix-guide (accessed Jun. 09, 2024).

[21] Ka Wing Wan, Chun Hoi Wong, Ho Fung Ip, Fan, D., Pak Leung Yuen, Hoi Ying Fong, & Ying, M. (2021). Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study. Quantitative Imaging in Medicine and Surgery, 11(4), 1381–1393. https://doi.org/10.21037/qims-20-922

[22] “Mann-Whitney U Test in SPSS Statistics | Setup, Procedure & Interpretation | Laerd Statistics,” *Laerd.com*, 2018. https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php (accessed Jun. 09, 2024).

[23] T. W. MacFarland and J. M. Yates, “Mann–Whitney U test,” in *Springer eBooks*, 2016, pp. 103–132. doi: 10.1007/978-3-319-30634-6_4.

[24] “Wilcoxon Test: Definition in Statistics, Types, and Calculation,” Investopedia, 2024. https://www.investopedia.com/terms/w/wilcoxon-test.asp (accessed Jun. 09, 2024).