Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multi-Class Liver Disease Classification Using a Hybrid Deep Learning Framework Based on YOLO and CaiT Transformer Architectures

Kerim Berkay BUÇAN

kerimbucan@ogr.bandirma.edu.tr

Bandırma Onyedi Eylül University

Serhat KILIÇARSLAN

Bandırma Onyedi Eylül University

Aykut DİKER

Bandırma Onyedi Eylül University

Research Article

Keywords: Liver Disease Classification, Hybrid Deep Learning, YOLOv8, CaiT Transformer, Medical Image Analysis

Posted Date: February 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8870050/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multi-Class Liver Disease Classification Using a Hybrid Deep Learning Framework Based on YOLO and CaiT Transformer Architectures

[1]Kerim Berkay BUÇAN, berkaybucan@gmail.com
[2]Serhat KILIÇARSLAN, https://orcid.org/0000-0001-9483-4425
[3]Aykut DİKER, https://orcid.org/0000-0002-1207-8548
[1,2,3]Department of Software Engineering, Faculty of Engineering, Bandırma Onyedi Eylül University, Balıkesir 10010, Türkiye, skilicarslan@bandirma.edu.tr, adiker@bandirma.edu.tr
Correspondence: berkaybucan@gmail.com

Abstract- Doctors must tell liver disease patterns apart fast but reading scans by eye often drifts with opinion and the classes look. This work builds a two part deep-learning tool - it first pulls local shapes through the YOLOv8m backbone then lets a Class Attention Image Transformer (CaiT) see the whole picture. Validation and test data came from Roboflow: 3 976 tagged liver images that show four structural lesions - steatosis, ballooning, inflammation, fibrosis - split 70 %, 20 %, 10 %. The pipeline keeps YOLOv8m as the visual encoder, feeds its feature tensor through an adapter so the Transformer can ingest it, sends the result to a CaiT-XS24-384 block for context vectors plus ends with a dense layer that outputs the class. We measured how well the model worked - computing accuracy, balanced accuracy, precision, recall, the F1-score, Cohen's kappa and the confusion matrix. On the separate test set, the hybrid model reached 95.00 % accuracy, 95.03 % balanced accuracy, 95.25 % precision, 95.03 % recall, 95.00 % F1-score and 93.33 % Cohen's kappa. Those numbers show high agreement between predicted plus true labels and stable separation among disease classes. When we compared methods, combining convolutional locality with Transformer based class-attention delivered solid performance but also balanced semantic detail against classification reliability for liver disease images.

Keywords: *Liver Disease Classification, Hybrid Deep Learning, YOLOv8, CaiT Transformer, Medical Image Analysis.*

Introduction

The liver belongs to the biggest organs inside the human body. It sits in the upper right section of the abdomen, directly under the ribs [1]. With a mass of about 1.5 kilograms, it carries out key tasks that keep the body working [2]. Besides handling nutrients that arrive from food, the liver removes toxic substances and forms some compounds the body must have [3]. The liver acts like a chemical factory for the body - the liver filters blood, stores glycogen so the body has energy, produces bile to break down fats and stores certain vitamins [4]. The liver also keeps steady levels of numerous substances that move through the body. Because it performs so many functions, the liver is essential for human health [2]. The liver carries out essential tasks - its condition matters greatly. But its structure and operation sometimes suffer damage or disruption [5Impairments of this type slowly lower the liver's ability to carry out its tasks and harm the health of the entire body [6]. Liver diseases often bring

tiredness, reduced hunger, weight loss and yellow skin. When the illness worsens, severe complications follow [7]. Liver disease starts when liver cells are injured - months or years later, the shape and internal layout of the liver alter [8]. Extra connective tissue, inflammation and fat buildup are the main structural problems that harm the liver. Each of those changes blocks normal liver activity plus if they persist, the organ eventually stops working. Because every disorder causes distinct damage and ranges from mild to severe, doctors must detect but also classify them with precision. A clear identification of the altered liver structure shows how far the illness has moved forward and allows treatment to start without delay. However, traditional diagnostic methods often fall short in detecting these disorders [9]. Those methods show poor sensitivity when disease is early or mild, depend heavily on the operator, require an invasive biopsy and allow only limited measurement [10]. As technology moves forward, methods that rely on artificial intelligence - especially computer vision and image processing - now stand out as practical substitutes [11]. Computer vision systems examine pictures of liver tissue - they spot small structural changes faster, with fewer mistakes and with steady results. This shortens the time needed to reach a diagnosis plus lowers the risk of human error [12]. This study presents a hybrid system that links the YOLOv8, YOLOv9 and YOLOv11 networks to the CaiT backbone to classify liver diseases from images. The design seeks sharper results because it extracts finer details than the standard YOLO families. Training and testing used 3,976 images taken from Roboflow. The team judged the network with precision, recall, F1-score, balanced accuracy besides Cohen's Kappa.

Related Work

During the past few years, doctors and researchers have started to rely repeatedly on artificial intelligence systems to identify liver diseases. This part of the report looks back at earlier published work plus points out how the new model differs from those earlier efforts.

Gupta et al. (2022) conducted liver disease prediction using various machine learning algorithms (LightGBM, XGBoost, kNN, Naïve Bayes, etc.) on the ILPD dataset obtained from the UCI Machine Learning Repository [13]. The analysts first filled empty cells in the table then deleted rows that sat too far from the rest and finally added synthetic rows so that every category held the same number of cases. A gradient boosting tree model named LightGBM labeled 63 out of every 100 records correctly, the best score obtained. All work used only numbers arranged in rows and columns - no pictures were examined.

Ghazal et al. (2022) also proposed an intelligent model based on an Artificial Neural Network (ANN) using the ILPD dataset [14]. The model reached a maximum accuracy of 88.4 % and scored better than multiple alternative machine learning techniques. As in the earlier project, the work used only structured tables of numbers - it did not analyse any medical images.

Yousefzamani et al. (2023) presented a comprehensive review on ultrasound-based liver disease classification approaches, ranging from conventional CNNs to hybrid CNN–Transformer models [15]. The report stressed that liver fat, fibrous scarring and liver cancer must be found soon after they begin. It listed the weak points of standard tests plus the strong points of diagnosis guided by artificial intelligence.

Yao et al. (2023) proposed the Bio-Inspired Network (BiNet) for diagnosing liver steatosis in ultrasound images [16]. BiNet copied key parts of how the human eye and brain handle sight. It added attention modules that let doctors see which image regions drive each choice. Because of this design, the network classifies fatty liver cases with high accuracy.

Yen et al. (2024) introduced a risk-controlled neural network for fatty liver classification using ultrasound images from the Taiwan Biobank [17]. The model uses conformal prediction methods to ensure reliability and delivers strong predictive performance.

Alshagathrh et al. (2024) developed a hybrid deep learning and machine learning framework for detecting hepatocyte ballooning in liver ultrasound images [18]. The model used a CNN to pull out features and fed them to standard machine learning classifiers. Test results showed it marked most true ballooning cases as positive plus rarely flagged normal tissue by mistake.

Huang et al. (2024) aimed to classify liver diseases using the YOLOv8 architecture on a dataset obtained from Roboflow [19]. The YOLOv8s model reached strong scores - accuracy hit 94 percent and recall reached 96 percent. The project stayed with the standard YOLO layout - no Transformer style blocks were added to learn deeper features.

Sharma et al. (2024) proposed a Contextual Correlation Network (COCOn) model developed for the segmentation of liver diseases [20]. A study used a large dataset from Roboflow to test a model on CT besides MRI images. The model achieved 94.2 percent overall accuracy, a cross entropy loss of 0.17 and F1 scores for individual classes as high as 99 percent. It segmented lesion types like steatosis, inflammation and ballooning with strong results. The architecture included multi scale contextual feature fusion and attention mechanisms. The work addressed only segmentation - no classification step was carried out.

Zeng et al. (2025) proposed a deep learning model for staging liver cirrhosis from multi-sequence MRI images [21]. The system used attention layers that respond to specific sequences - this design surpassed older radiomics approaches when it separated early cirrhosis from advanced cirrhosis.

In comparison to earlier work, the new hybrid model links YOLOv8, YOLOv9 and YOLOv11 to a CaiT Transformer backbone. This union delivers reliable object detection and stronger classification. The assessment went beyond mean average precision and raw accuracy - it also included the F1-score, balanced accuracy besides Cohen's Kappa. Rich visual features from deep layers supported clearer decisions plus the model raised multi class liver disease classification to a markedly higher level.

The rest of this paper is arranged in the following way - Section II describes the materials and methods. It explains the dataset, the model design plus how training was set up. Section III gives the experimental outcomes and compares them with baseline models but also those found in other papers. Section IV examines what the results mean and what they show about the new approach. Section V ends the paper with a brief list of what was added to knowledge as well as what should be studied next. The Figure displays the proposed systems.

Material and Method

A. Proposed Method

This study introduces a combined deep learning system that links the YOLOv8m network with the CaiT Transformer so that it can label multiple kinds of liver disease in one pass. Figure 1 shows how the YOLOv8m backbone first reads the liver image and pulls out convolutional features at different scales [26]. An adapter then reshapes the channels and the spatial size of those feature maps so that the Transformer layer can work with them. The CaiT architecture processes the representations so that the

model learns long range links across the whole image [29]. A fully connected layer then outputs one logit score for each of the four labels - steatosis, ballooning, inflammation and fibrosis.

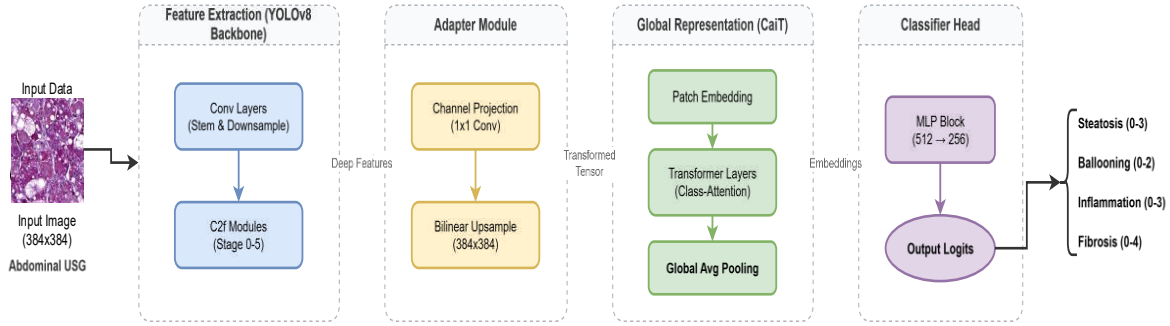


Figure 1. The block diagram of overall proposed systems.

B. Dataset

The image data used in this study consists of a total of 3,976 high-resolution medical images, provided by the Roboflow platform and labeled specifically for liver disease classification [22]. The pictures show four common liver defects - fatty change, swollen cells, infiltrated white cells and scar tissue. Every picture carries a YOLO label that names the defect plus gives a tight rectangle round the affected area. The label lets the network decide what the defect is and where it sits in one pass. Figure 2 presents example images representing the four disease categories included in the dataset.

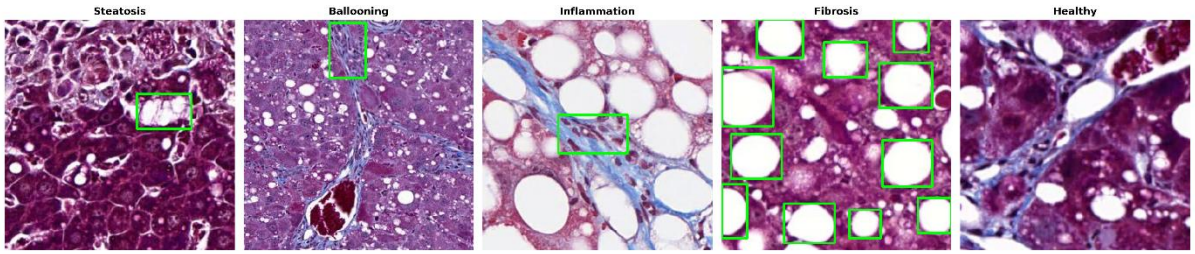


Figure 2. Sample images from the Roboflow dataset representing four major liver disease classes

B.1. Image Characteristics

All images in the dataset are in color format and have been standardized to a resolution of 640×640 pixels [22]. The pictures show clear signs of disease that are visible either to the naked eye or with magnification. We obtained the pictures - scanning liver biopsy samples or - using medical imaging equipment like magnetic resonance scanners, computed tomography scanners or light microscopes [23]. The labeling process was carried out under expert pathological evaluation, and each image corresponds to a single disease class. This structure allows the problem to be formulated as a multi-class classification task [24].

B.2. Class Distribution and Label Balance

The dataset shows a balanced class distribution - each disease category holds roughly the same number of samples - every class contains between 950 and 1,000 images. The balance gives the model a clear advantage while it trains, because it stops the system from learning to ignore the classes that appear least often [24]. This uniformity also guarantees that the performance metrics used for evaluation -

specifically the F1-score and Cohen's Kappa - are interpreted fairly and have statistical significance [25].

B.3. Training, Validation, and Test Split

To stop the model from memorizing the training data, the full image collection was split into three parts. One part, the training set, holds 2,782 pictures - about seven out of every ten. A second part, the validation set, contains 794 pictures - close to two out of every ten. The last 400 pictures, one in ten, form the test set. A fixed random seed controlled the split - the same division appears each time the code runs and every class keeps the same proportion in each part.

The model adjusted its internal weights on the training set - it chose hyperparameters and decided when to halt on the validation set. The test set stayed untouched until the very end - only then did the model face those pictures, giving a clear measure of how well it handles new, unseen images.

C. Model Architecture

This study introduces a combined deep learning system that sorts liver diseases into multiple classes. The design links the convolutional feature extraction part of the YOLOv8m backbone to the context encoding ability of the CaiT framework. The goal is to build a layered learning path that first records small structural details and then adds larger semantic meaning. Every part of the system is set out step by step in the sections that follow.

The proposed architecture consists of five main components:

1. Feature extractor based on the YOLOv8 backbone
2. Adapter block for tensor transformation
3. Representation learner using CaiT Transformer layers
4. Fully connected classifier block
5. End-to-end model inference flow

C.1. YOLOv8 Backbone: Deep Convolutional Feature Extraction

Stage one of the model employs the first six Ultralytics YOLOv8m modules as its convolutional backbone [26,27]. The structure pulls basic visual elements from the input image, like edges, textures and color gradients, to create the first set of feature maps. The modules involved are Focus, C2f, Conv, Bottleneck besides SPPF layers. Each performs stacked convolutional operations to record spatial details at multiple scales [28].

The backbone produces a feature tensor with multiple channels and acts as the basic visual encoder for the layers that follow. Figure 3 shows the structure of the YOLOv8 backbone plus its main convolutional blocks.

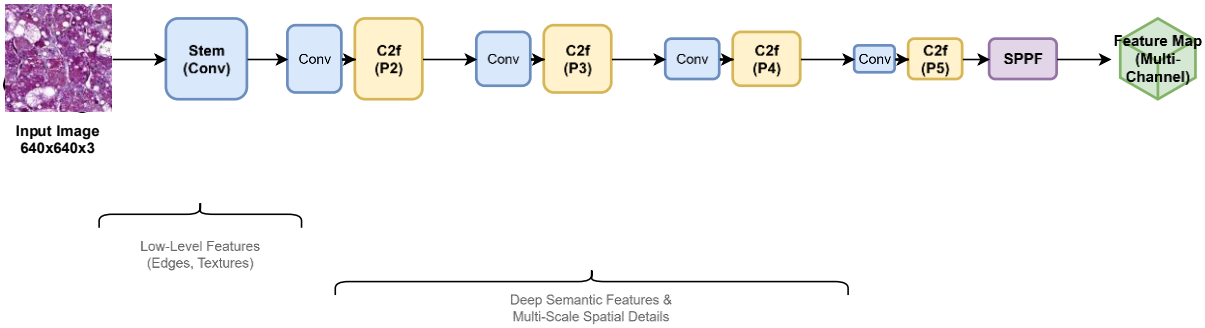


Figure 3. Architectural structure of the YOLOv8 backbone.

C.2. Adapter Block: Tensor Transformation

YOLOv8 produces results that do not match the Transformer design because the number of channels and the image size differ. To solve this, an adapter block sits between the two systems. It changes the shape of the data plus brings the values to a range that CaiT expects. The block carries out three main steps:

- Channel compression - A convolutional projection lowers the channel count of the high dimensional YOLOv8 output to 32.
- Channel rearrangement - The tensor with 32 channels is rearranged into three channels so that the result resembles the RGB image a vision Transformer expects.
- Rescaling - The tensor is resized to the fixed spatial size of 384×384 pixels, which the CaiT model requires [27].

The block also contains batch normalization and a ReLU activation function. They keep training stable plus add non linearity. Figure 4 shows how data moves through the adapter block and how the block is built.

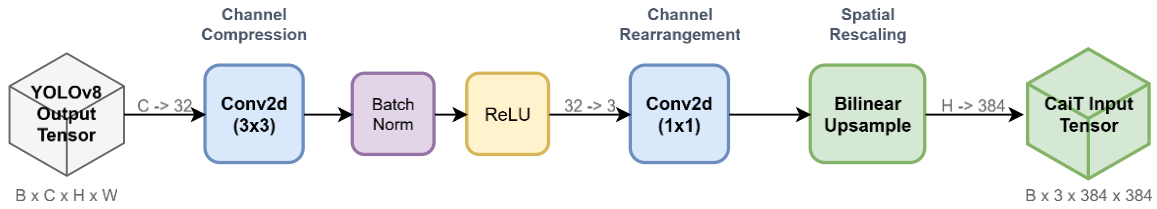


Figure 4. Flow diagram of the adapter block.

C.3. CaiT Transformer Layers: Contextual Representation Learning

The changed visual tensor passes through the CaiT-XS24-384 model that already trained. This model is an improved Vision Transformer. It adds class attention mechanisms [18,19]. The model does not confine itself to local neighbourhood data - it records long range links that stretch across the whole spatial domain [29].

Key characteristics of this module include:

- Global attention applied over all pixel patches
- Enhanced information flow through class-specific tokens

- Improved representation learning for structural abnormalities in pathology

In this work, we removed the classifier head from the CaiT model - setting num_classes to 0. We extract feature embeddings using the forward_features method. Each input image becomes a fixed length vector. Figure 5 shows the architecture and attention mechanism of the CaiT Transformer.

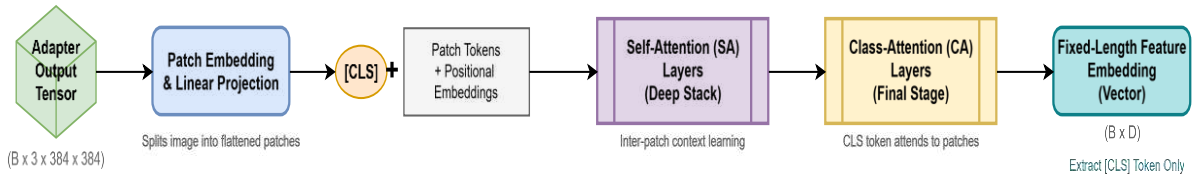


Figure 5. Overview of the CaiT Transformer architecture.

C.4. Classifier Block: Fully Connected Projection

The output vector from the Transformer is forwarded through a series of fully connected layers for classification. The design is as follows:

- Layer normalization for input stability
- Linear projection from embedding dimension to 512 units
- ReLU activation
- Dropout with a rate of 0.3 to prevent overfitting
- Linear projection from 512 to 256 units
- ReLU activation
- Final linear layer projecting to 4 output units, representing softmax class probabilities [30]

This block maps the abstract Transformer representations into discrete class scores.

C.5. End-to-End Inference Pipeline

The overall flow of the model from input to output is described below [31]:

1. Input: RGB image with a resolution of 640×640
2. YOLOv8 backbone generates feature maps with shape (B × C × H × W)
3. Adapter block transforms it into a 3-channel tensor of size 384×384
4. CaiT Transformer produces a contextual class representation vector (B × D)
5. Classifier block outputs a score vector for four classes (B × 4)

The overall inference flow of the proposed model is shown in Figure 6.

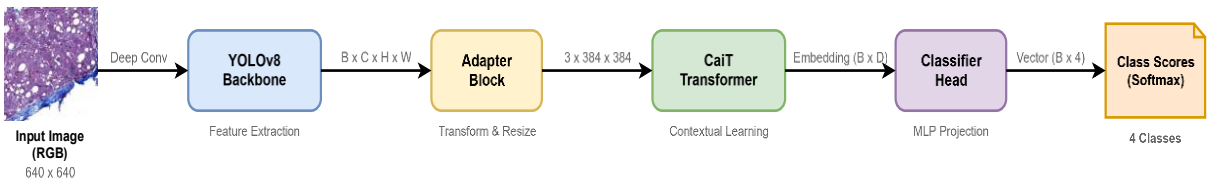


Figure 6. End-to-end flow diagram of the proposed YOLOv8-CaiT hybrid architecture.

D. Training Procedure and Experimental Setup

The following section gives a step-by-step description of how the model was trained. It covers the preparation of the data, the choice of hyperparameters and the plan used for optimization. Every part of the training was built to use the convolutional power of YOLOv8 together with the deep contextual attention supplied by the CaiT Transformer architecture.

D.1. Data Preprocessing and Augmentation Strategy

The Roboflow Liver Disease v2 dataset already holds enough images in every category but we still applied extra data augmentation steps so the model learns broader patterns and becomes less prone to overfitting [33].

Every image was first scaled to the exact shape of 384 by 384 pixels, because the Transformer model only accepts that size. While the network learned, each picture had a fifty percent chance of being mirrored left-to-right so the system saw more variations of the same scene. Brightness and contrast were also shifted at random to mimic camera plus lighting differences found outside the lab. Every red-green-blue value was shifted and scaled so its mean became 0.5 but also its spread became 0.5. After this normalization, the pixels were packed into a tensor, the data structure required by the PyTorch training loop. Figure 7 shows those operations as a block flow diagram.

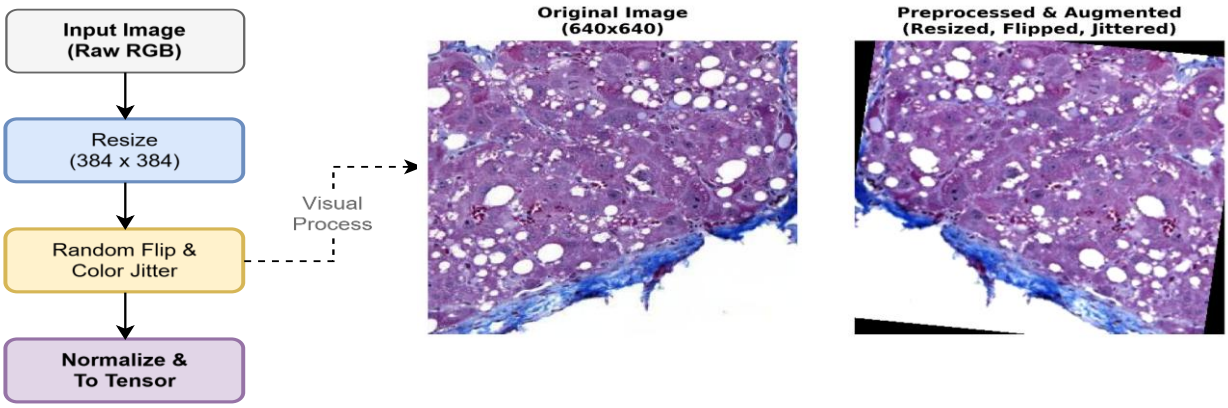


Figure 7. Data Augmentation Flow Diagram

D.2. Training Workflow

The training began when we fed the images into a combined system. YOLOv8 first pulled out simple location cues. An adapter block reshaped those cues so the CaiT Transformer could read them. The Transformer turned the cues into meaning vectors. A classifier turned the vectors into raw scores. We measured how far the scores sat from the true labels with cross entropy loss then sent the loss back through every layer. AdamW shifted the weights. We ran this same sequence for a set count of epochs until the numbers settled. Figure 8 shows the full loop.

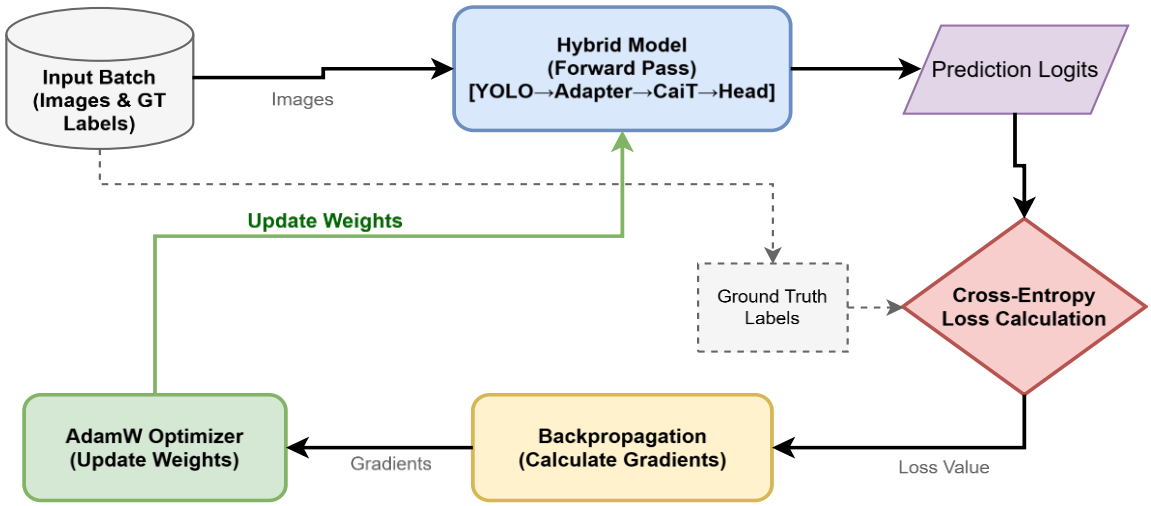


Figure 8. Training Loop Diagram

D.3. Hyperparameters and Training Configuration

The hybrid YOLOv8-CaiT model was trained with a specific set of hyperparameters listed in Table 1. We picked each value after early tests showed it kept loss curves smooth and kept top-1 accuracy high while the network learned.

Table 1. Hyperparameter Configuration

Parameter	Value
Input Image Size	384x384
Number of Epochs	30
Batch Size	16
Learning Rate	1e-4
Optimizer	AdamW
Dropout Rate	0.3
Normalization (Mean/Std)	0.5/0.5

D.4. Performance Evaluation Metrics

We measured the YOLO-CaiT hybrid model with a standard battery of tests. The tests count how often the model predicts the correct class, how it behaves when some classes appear far repeatedly than others and how closely its answers match the true labels. The evaluation measured six standard metrics - Accuracy, Balanced Accuracy, Cohen's Kappa, Precision, Recall besides F1-Score. Each metric serves to test how well the classifier performs and how it handles risk plus together they give a reliable picture of the model's behaviour [32]. A Confusion Matrix was also employed to show the pattern of the model's classification mistakes [34].

D.4.1. Confusion Matrix

Multi-class classification tasks often need more than a single number to show how well the model works for every category. A Confusion Matrix counts the correct and wrong predictions the model gives for each class. The table lists how often each real class was paired with each predicted class. From the counts, one derives precision, recall and other measures [35].

For a binary classification scenario, the Confusion Matrix is structured as shown in **Table 2**.

Table 2. Confusion Matrix Structure

	Predicted: Positive (1)	Predicted: Negative (0)
Actual: Positive (1)	True Positive (TP)	False Positive (FP)
Actual: Negative (0)	False Negative (FN)	True Negative (TN)

- **True Positive (TP)**: The model correctly predicts a positive instance.
- **False Positive (FP)**: A negative instance is incorrectly classified as positive.
- **False Negative (FN)**: A positive instance is incorrectly classified as negative.
- **True Negative (TN)**: The model correctly identifies a negative instance.

For multi-class settings, this matrix is extended such that each row represents an actual class and each column a predicted class, forming an $n \times n$ matrix for n classes [34].

D.4.2 Accuracy

Accuracy is the ratio of correctly predicted instances to the total number of predictions in Equation (1):

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

This metric provides a general overview of model performance across all classes. However, it may be misleading in cases of class imbalance.

D.4.3. Balanced Accuracy

Balanced Accuracy addresses the limitations of standard accuracy by averaging the recall obtained on each class represents in Equation (2):

$$\text{Balanced Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP_i}{TP_i+FN_i}\right) \quad (2)$$

where N is the number of classes. This metric is particularly useful for imbalanced datasets, as it ensures that minority classes contribute equally to the final score [36].

D.4.4. Cohen's Kappa

Cohen's Kappa is a statistical measure of inter-rater agreement that accounts for the possibility of the agreement occurring by chance in Equation (3):

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (3)$$

- PoP_oPo: Observed agreement (i.e., Accuracy)
- PeP_ePe: Expected agreement by chance

The Kappa coefficient ranges from -1 to +1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance, and negative values indicate systematic disagreement [37].

C.4.5. Precision

Precision, also known as Positive Predictive Value, quantifies the proportion of correctly predicted positive observations out of all predicted positives in expression Equation (4):

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

This metric penalizes false positives and is important in applications where over-prediction of a class is costly.

D.4.6. Recall

Recall, also known as Sensitivity or True Positive Rate, measures the proportion of actual positive instances that were correctly identified by the model in Equation (5):

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

It captures the model's ability to detect all relevant instances of a particular class.

D.4.7. F1-Score

The F1-Score is the harmonic mean of Precision and Recall and serves as a balanced metric that is robust to class imbalance as a Equation (5):

$$F1 = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \quad (6)$$

A high F1-Score indicates that the model achieves both high precision and recall, making it suitable for real-world classification problems with uneven class distributions.

Experimental Results

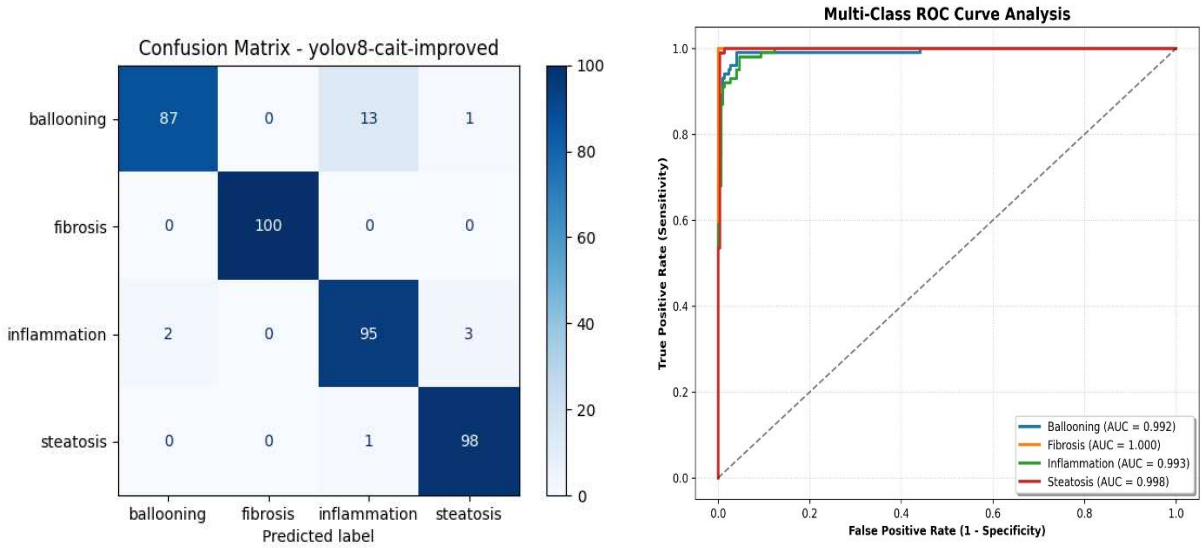
The section shows the test results for the YOLOv8m-CaiT hybrid model. It lists how well the model works when set beside pure transformer networks, the plain YOLO family and published baselines. The reported values cover accuracy, balanced accuracy, Cohen's kappa, precision, recall besides F1-score. A confusion matrix accompanies each run so the reader can see how predictions spread across the classes.

A. Hybrid Model Performance

The performance of the proposed YOLOv8m-CaiT hybrid model is summarized in Table 3, and its classification behavior and ROC curve across classes is visualized in Figure 9.

Table 3. Performance Metrics of the YOLOv8m-CaiT Hybrid Model

Model	Accuracy	Balanced Accuracy	Cohen's Kappa	Precision	Recall	F1-Score
YOLOv8m-CaiT	%95,00	%95,03	%93,33	%95,25	%95,03	%95,00



(a) **(b)**

Figure 9. (a) Confusion Matrix (b) ROC curve of YOLOv8m-CaiT

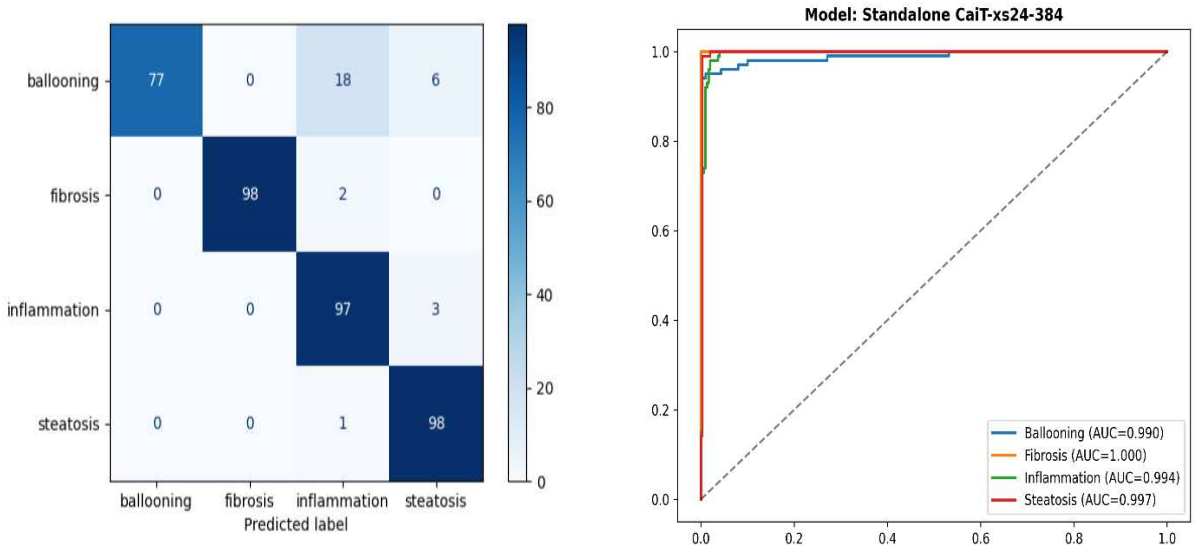
B. Transformer-only vs. Hybrid Comparison

In this section, the standalone CaiT-xs24-384 model is compared with the YOLOv8m-CaiT hybrid approach. Table 4 presents their performance side by side, while Figure 10 illustrates the confusion matrix and ROC curve of the Transformer-only model.

Table 4. Comparison of YOLOv8m-CaiT and CaiT-xs24-384 Models

Model	Accuracy	Balanced Accuracy	Cohen's Kappa	Precision	Recall	F1-Score
YOLOv8m-CaiT	**%95,00**	**%95,03**	**%93,33**	**%95,25**	**%95,03**	**%95,00**

CaiT-xs24-384	%92,50	%92,56	%90,00	%93,45	%92,56	%92,41



(a) **(b)**

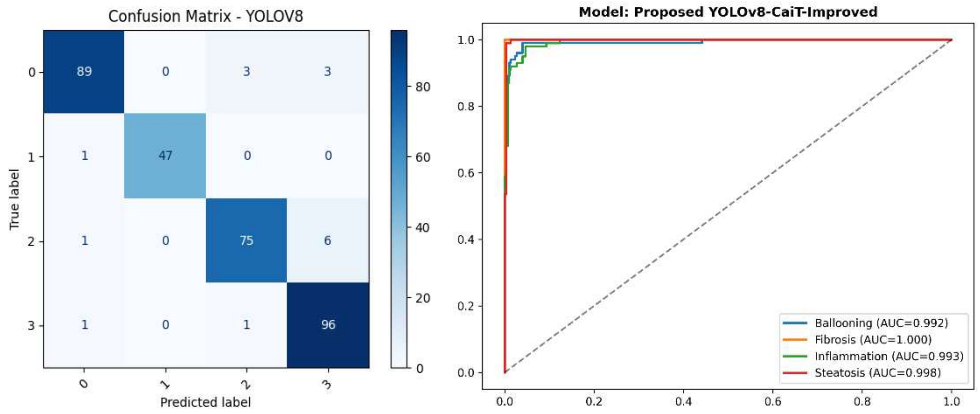
Figure 10. (a) Confusion Matrix (b) ROC Curve of CaiT-xs24-384

C. Comparison with Classical YOLO Models

We tested YOLOv8m-CaiT against the standard YOLOv8, YOLOv9 and YOLOv11 versions. Table 5 lists the scores for each model. Figures 12 - 14 display the confusion matrices besides ROC curves for the same four models.

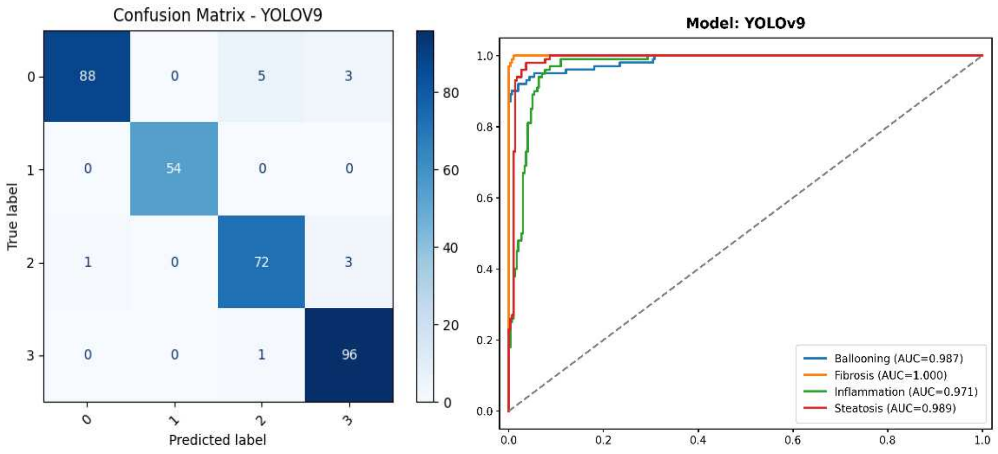
Table 5. Comparison with YOLOv8, YOLOv9, and YOLOv11 Models

Model	Accuracy	Balanced Accuracy	Cohen's Kappa	Precision	Recall	F1-Score	mAP@50	mAP@50-95
YOLOv8m-CaiT	%95,00	%95,03	%93,33	%95,25	%95,03	%95,00	-	-
YoloV8	%95,05	%95,26	%93,25	%95,78	%95,26	%95,47	%54,42	%35,06
YoloV9	%95,98	%96,34	%94,55	%96,33	%96,34	%96,28	%55,09	%34,85
YoloV11	%97,38	%97,54	%96,42	%97,71	%97,54	%97,61	%55,61	%35,56



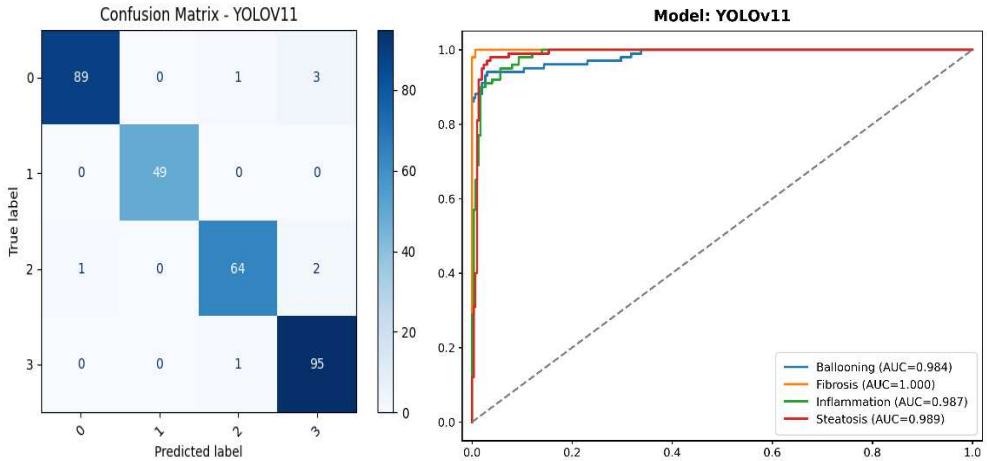
(a) **(b)**

Figure 12. (a) Confusion Matrix (b) ROC curve of YOLOv8



(a) **(b)**

Figure 13. (a) Confusion Matrix (b) ROC curve of YOLOv9



(a) **(b)**

Figure 14. (a) Confusion Matrix (b) ROC curve of YOLOv11

Discussion

We studied how to sort liver diseases into multiple distinct classes. Our solution combines the YOLOv8m backbone with the CaiT Transformer. To measure the new model fairly and to place it beside prior work, we matched its scores against the original YOLO design plus against recent published results. Table 6 lists the performance numbers for each method.

Table 6 lists many ways to spot liver disease - papers move from older machine learning rules to deep learning models that carve out the liver shape. Huang et al. [12] tested two older methods - a Random Forest plus a CNN whose output was sent to an SVM. The first reached 85 % exact matches - the second reached 88 %. The same study compared three approaches - a system that combined chaotic cuckoo search with AlexNet reached 86 % precision. UNet-60, a standard network in medical imaging, raised the score to 91 %. The new YOLOv8m-CaiT model went beyond both, recording 95.25 % precision and 95.00 % F1-score on the test data. The hybrid design extracts features that separate diseased tissue from normal tissue more effectively than earlier algorithms and older deep networks.

Table 6 reports that Huang et al. [12] applied the unmodified YOLOv8s network and achieved 94 % precision and 59 % mAP@50. Our combined network delivered 95.03 % balanced accuracy, because precision plus recall stayed nearer to each other. This result indicates that the Class Attention component within the CaiT block supplies semantic information - the network learns to tell apart tissue alterations like steatosis and ballooning that appear similar.

A review of the modern convolutional networks examined in the study - YOLOv9m and YOLOv11m - shows that YOLOv11m reached the top result: 97.38 % accuracy and 97.61 % F1-score. Standard CNN models of this type learn structural patterns well when the data set is large and balanced. The new YOLOv8m-CaiT hybrid stayed steady across classes, recording a Cohen's Kappa of 93.33 % with only modest parameter tuning.

When the outcomes are set beside published work, the hybrid strategy surpasses traditional machine learning techniques and older deep architectures like AlexNet besides UNet. It also yields clear visual explanations plus strong semantic features, matching the performance of the current best detector, YOLOv11.

Table 6. Comparison with Literature Methods

Model	Accuracy	Balanced Accuracy	Cohen's Kappa	Precision	Recall	F1-Score	mAP@50	mAP@50-95
YoloV11m	%97,38	%97,54	%96,42	%97,71	%97,54	%97,61	%55,61	%35,56

YoloV9m	%95,98	%96,34	%94,55	%96,33	%96,34	%96,28	%55,09	%34,85
YoloV8m	%95,05	%95,26	%93,25	%95,78	%95,26	%95,47	%54,42	%35,06
YOLOv8m-CaiT	**%95,00**	**%95,03**	**%93,33**	**%95,25**	**%95,03**	**%95,00**	**-**	**-**
CaiT-xs24-384	%92,50	%92,56	%90,00	%93,45	%92,56	%92,41	-	-
YOLOv8m (Huang vd.)	-	-	-	%89,00	%96,00	-	%56,00	-
YOLOv8s (Huang vd.)	-	-	-	%94,00	%96,00	-	%59,00	-
UNet-60 (Huang vd.)	-	-	-	%91,00	%94,00	-	%57,00	-
CNN+SVM (Huang vd.)	-	-	-	%88,00	%92,00	-	%54,00	-
RF(Huang vd.)	-	-	-	%85,00	%90,00	-	%53,00	-
Chaotic Cuckoo + AlexNet (Huang vd.)	-	-	-	%86,00	%89,00	-	%52,00	-

Conclusion

The study presents a hybrid deep learning architecture that links the convolutional trunk of YOLOv8m with the contextual attention module found in Class Attention in Image Transformers (CaiT) to classify multiple liver diseases. Tests showed that the combined YOLOv8m - CaiT system reached high accuracy, strong per class agreement and feature maps rich in semantic content, proving that it separates visually liver disease patterns in a dependable way.

When compared with pure transformer models plus earlier YOLO-based networks, the hybrid design merged local spatial feature extraction with global semantic reasoning, giving steadier and more consistent predictions for every class. Although the newer YOLOv11 convolutional network delivered higher raw scores on the same data set, the hybrid approach still offered a competitive but also trustworthy option - it learned each disease category in a balanced manner and retained improved context awareness without adding undue architectural complexity. Such a balance matters in medical imaging, where clarity, stability as well as general ability count as much as top accuracy.

The study shows that the hybrid framework will improve if it receives more advanced convolutional backbones and transformer variants. It also needs tests on multi center plus heterogeneous datasets so that its behaviour in real clinical settings becomes clear. When explainable artificial intelligence techniques and multi modal clinical information join the system, clinicians will interpret the output more easily but also receive stronger decision support. The present work relies on one public dataset - external validation with clinical data from multiple institutions will measure generalizability in future work. The YOLOv8m - CaiT framework adds value to computer aided diagnosis and demonstrates that a union of convolutional as well as transformer architectures handles complex multi class medical image classification, especially in liver disease imaging and related areas.

Funding: The study was not supported by any funding.

Ethical approval: This work contains no data or other information derived from studies.

CRediT authorship contribution statement: Kerim Berka BUÇAN: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Serhat KILIÇARSLAN: Writing – review & editing, Methodology, Investigation. Aykut Diker: Writing – review & editing, Writing – original draft, Resources,

Declaration of competing interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability: Data will be made available on request.

References

[1] Ishibashi, Hiromi, Minoru Nakamura, Atsumasa Komori, Kiyoshi Migita, ve Shinji Shimoda. "Liver Architecture, Cell Function, and Disease". *Seminars in Immunopathology* 31, sy 3 (01 Eylül 2009): 399-409. https://doi.org/10.1007/s00281-009-0155-6.

[2] Yoshizumi, T., G. E. Gondolesi, C. A. Bodian, H. Jeon, M. E. Schwartz, T. M. Fishbein, C. M. Miller, ve S. Emre. "A simple new formula to assess liver weight". *Transplantation Proceedings* 35, sy 4 (01 Haziran 2003): 1415-20. https://doi.org/10.1016/S0041-1345(03)00482-2.

[3]Stange, Jan, Steffen R. Mitzner, Sebastian Klammt, Jens Freytag, Piotr Peszynski, Jan Loock, Heiko Hickstein, vd. "Liver Support by Extracorporeal Blood Purification: A Clinical Observation". *Liver Transplantation* 6, sy 5 (01 Eylül 2000): 603-13. https://doi.org/10.1053/jlts.2000.7576.

[4]Rui, Liangyou. "Energy Metabolism in the Liver". *Comprehensive Physiology* 4, sy 1 (Ocak 2014): 177-97. https://doi.org/10.1002/cphy.c130024.

[5]Devarbhavi, Harshad, Sumeet K. Asrani, Juan Pablo Arab, Yvonne Ayerki Nartey, Elisa Pose, ve Patrick S. Kamath. "Global burden of liver disease: 2023 update". *Journal of Hepatology* 79, sy 2 (01 Ağustos 2023): 516-37. https://doi.org/10.1016/j.jhep.2023.03.017.

[6] Gan, Can, Yuan Yuan, Haiyuan Shen, vd. "Liver Diseases: Epidemiology, Causes, Trends and Predictions". *Signal Transduction and Targeted Therapy* 10, sy 1 (2025): 33. https://doi.org/10.1038/s41392-024-02072-z.

[7] Bergasa, Nora V. "Approach to the Patient with Liver Disease". Içinde *Clinical Cases in Hepatology*, editör Nora V. Bergasa, 5-26. London: Springer, 2022. https://doi.org/10.1007/978-1-4471-4715-2_2.

[8] Dukić, Marija, Tijana Radonjić, Igor Jovanović, Marija Zdravković, Zoran Todorović, Nemanja Kraišnik, Bojana Aranđelović, vd. "Alcohol, Inflammation, and Microbiota in Alcoholic Liver Disease". *International Journal of Molecular Sciences* 24, sy 4 (Ocak 2023): 3735. https://doi.org/10.3390/ijms24043735.

[9]Olveres, Jimena, Germán González, Fabian Torres, José Carlos Moreno-Tagle, Erik Carbajal-Degante, Alejandro Valencia-Rodríguez, Nahum Méndez-Sánchez, ve Boris Escalante-Ramírez. "What is new in computer vision and artificial intelligence in medical image analysis applications".

Quantitative Imaging in Medicine and Surgery 11, sy 8 (Ağustos 2021): 3830-53. https://doi.org/10.21037/qims-20-1151.

[10] Sumida, Yoshio, Atsushi Nakajima, ve Yoshito Itoh. "Limitations of liver biopsy and non-invasive diagnostic tests for the diagnosis of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis". *World Journal of Gastroenterology : WJG* 20, sy 2 (2014): 475-85. https://doi.org/10.3748/wjg.v20.i2.475.

[11]Li, Xin, Lei Zhang, Jingsi Yang, ve Fei Teng. "Role of Artificial Intelligence in Medical Image Analysis: A Review of Current Trends and Future Directions". *Journal of Medical and Biological Engineering* 44, sy 2 (01 Nisan 2024): 231-43. https://doi.org/10.1007/s40846-024-00863-x.

[12] Yin, Chenglong, Huafeng Zhang, Jin Du, Yingling Zhu, Hua Zhu, ve Hongqin Yue. "Artificial Intelligence in Imaging for Liver Disease Diagnosis". *Frontiers in Medicine* 12 (Nisan 2025). https://doi.org/10.3389/fmed.2025.1591523.

[13]Gupta, Ketan, Nasmin Jiwani, Neda Afreen, ve Divyarani D. "Liver Disease Prediction using Machine learning Classification Techniques". İçinde *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, 221-26, 2022. https://doi.org/10.1109/CSNT54456.2022.9787574.

[14]Ghazal, Taher M., Aziz Ur Rehman, Muhammad Saleem, Munir Ahmad, Shabir Ahmad, ve Faisal Mehmood. "Intelligent Model to Predict Early Liver Disease using Machine Learning Technique". İçinde *2022 International Conference on Business Analytics for Technology and Security (ICBATS)*, 1-5, 2022. https://doi.org/10.1109/ICBATS54253.2022.9758929.

[15] Yousefzamani, Midya, ve Farshid Babapour Mofrad. "Deep Learning without Borders: Recent Advances in Ultrasound Image Classification for Liver Diseases Diagnosis". *Expert Review of Medical Devices* 22, sy 8 (2025): 827-43. https://doi.org/10.1080/17434440.2025.2514764.

[16] Yao, Yuan, Zhenguang Zhang, Bo Peng, ve Jin Tang. "Bio-Inspired Network for Diagnosing Liver Steatosis in Ultrasound Images". *Bioengineering (Basel, Switzerland)* 10, sy 7 (2023): 768. https://doi.org/10.3390/bioengineering10070768.

[17] Yen, Tso-Jung, Chih-Ting Yang, Yi-Ju Lee, Chun-houh Chen, ve Hsin-Chou Yang. "Fatty Liver Classification via Risk Controlled Neural Networks Trained on Grouped Ultrasound Image Data". *Scientific Reports* 14, sy 1 (2024): 7345. https://doi.org/10.1038/s41598-024-57386-3.

[18] Alshagathrh, Fahad, Mahmood Alzubaidi, Samuel Gecík, vd. "Hybrid Deep Learning and Machine Learning for Detecting Hepatocyte Ballooning in Liver Ultrasound Images". *Diagnostics (Basel, Switzerland)* 14, sy 23 (2024): 2646. https://doi.org/10.3390/diagnostics14232646.

[19]Huang, Junjie, Caihong Li, Fengjun Yan, ve Yuanchun Guo. "An Improved Liver Disease Detection Based on YOLOv8 Algorithm". *International Journal of Advanced Computer Science and Applications (IJACSA)* 15, sy 7 (31 Haziran 2024). https://doi.org/10.14569/IJACSA.2024.01507114.

[20]Sharma R, Rajesh, Akey Sungheetha, Rajiv Gandhi K, Sasikala Rani, GGS Pradeep, ve Ellappan V. "Segmentation of Medical Liver Image Diseases Using Improved Contextual Convolutional Neural Network (COCOn) Model". İçinde *2024 First International Conference on Innovations in*

Communications, Electrical and Computer Engineering (ICICEC), 1-5, 2024. https://doi.org/10.1109/ICICEC62498.2024.10809006.

[21] Zeng, Jun, Debesh Jha, Ertugrul Aktas, vd. *Liver Cirrhosis Stage Estimation from MRI with Deep Learning*. 2025. https://doi.org/10.48550/arXiv.2502.18225.

[22]Roboflow. "Liver Disease Dataset". https://universe.roboflow.com/roboflow-100/liver-disease.

[23]Van Beers, Bernard E., Jean-Luc Daire, ve Philippe Garteiser. "New imaging techniques for liver diseases". *Journal of Hepatology* 62, sy 3 (01 Mart 2015): 690-700. https://doi.org/10.1016/j.jhep.2014.10.014.

[24]He, Haibo, ve Edwardo A. Garcia. "Learning from Imbalanced Data". *IEEE Transactions on Knowledge and Data Engineering* 21, sy 9 (Eylül 2009): 1263-84. https://doi.org/10.1109/TKDE.2008.239.

[25] Sokolova, Marina, ve Guy Lapalme. "A systematic analysis of performance measures for classification tasks". *Information Processing & Management* 45, sy 4 (2009): 427-37. https://doi.org/10.1016/j.ipm.2009.03.002.

[26]Ultralytics. "YOLOv8". https://docs.ultralytics.com/models/yolov8.

[27]Kumar, Deepak, ve Vinay Kukreja. "CaiT-YOLOv9: Hybrid Transformer Model for Wheat Leaf Fungal Head Prediction and Diseases Classification". *International Journal of Information Technology*, 13 February 2025. https://doi.org/10.1007/s41870-025-02443-x.

[28] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, ve Jian Sun. "Deep Residual Learning for Image Recognition". *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Haziran 2016, 770-78. https://doi.org/10.1109/CVPR.2016.90.

[29] Touvron, Hugo, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, ve Hervé Jégou. "Going deeper with Image Transformers". 01 Ekim 2021, 32-42. https://doi.org/10.1109/ICCV48922.2021.00010.

[30] Hinton, Geoffrey, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Yoesoep Rachmad, ve Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research* 15 (Haziran 2014): 1929-58.

[31] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, vd. "A Survey on Deep Learning in Medical Image Analysis". *Medical Image Analysis* 42 (Aralık 2017): 60-88. https://doi.org/10.1016/j.media.2017.07.005.

[32] Zhang, Yonglai, Wenai Song, Shuai Li, Lizhen Fu, ve Shixin Li. "Risk Detection of Stroke Using a Feature Selection and Classification Method". *IEEE Access* 6 (2018): 31899-907. https://doi.org/10.1109/ACCESS.2018.2833442.

[33]Ying, Xue. "An Overview of Overfitting and Its Solutions". *Journal of Physics: Conference Series* 1168, sy 2 (Şubat 2019): 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.

[34]Swaminathan, Sathyanarayanan, ve B Roopashri Tantri. "Confusion Matrix-Based Performance Evaluation Metrics". *African Journal of Biomedical Research* 27 (30 Kasım 2024): 4023-31. https://doi.org/10.53555/AJBR.v27i4S.4345.

[35] Powers, David M. W. “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation”. arXiv:2010.16061. Preprint, arXiv, 11 Ekim 2020. https://doi.org/10.48550/arXiv.2010.16061.

[36]Velez, Digna R., Bill C. White, Alison A. Motsinger, William S. Bush, Marylyn D. Ritchie, Scott M. Williams, ve Jason H. Moore. “A Balanced Accuracy Function for Epistasis Modeling in Imbalanced Datasets Using Multifactor Dimensionality Reduction”. *Genetic Epidemiology* 31, sy 4 (2007): 306-15. https://doi.org/10.1002/gepi.20211.

[37]Sasikala, B S, Vinai George Biju, ve C. M. Prashanth. “Kappa and accuracy evaluations of machine learning classifiers”. İçinde *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 20-23, 2017. https://doi.org/10.1109/RTEICT.2017.8256551.