Université de Strasbourg

École Doctorale : **Mathématiques, Sciences de l'Information et de l'Ingénieur (ED269)**

Unité de Recherche : **ICube – UMR 7357**

THÈSE présenté par

Güinther SAIBRO

Soutenue le : **26 Février 2025**

Pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline/Spécialité : **Informatique**

Classification automatique des vidéos échografiques abdominales : de l'annotation subjective à l'inférence en temps réel des vidéos brutes

THÈSE dirigée par :

DIANA Michele	Professeur, Hôpitaux universitaires de Genève(HUG)
COLLINS Toby	Directeur de Recherche, IRCAD France

RAPPORTEURS :

BEGHDADI Azeddine	Professeur, Université Sorbonne Paris Nord
BERNARD Olivier	Professeur, Université de Lyon

AUTRES MEMBRES DU JURY :

ESSERT Caroline	Professeur, Université de Strasbourg
VALERIO Massimo	Professeur, Hôpitaux universitaires de Genève (HUG)

IRCAD

Surgical Data Science Team

Hôpitaux Universitaires
1, place de l’Hôpital
67091 Strasbourg Cedex, France

sds.ircad@ircad.fr
https://www.ircad.fr

Acknowledgments

To the esteemed members of my thesis jury :

- My deepest gratitude goes to **Toby Collins** and **Michele Diana**, my thesis advisors, whose unwavering support, exceptional guidance, and profound expertise have been invaluable throughout my doctoral journey. Their encouragement and insightful advice have shaped not only this work but also my growth as a researcher.
- I am sincerely thankful to **M. Azeddine Beghdadi** and **M. Olivier Bernard**, distinguished experts in the field, for agreeing to evaluate my thesis with such enthusiasm. Their constructive feedback and thoughtful perspectives are deeply appreciated.
- A special acknowledgment to **Mme. Caroline Essert**, whose interest in my research dates back to her participation in my thesis advisory committee. It is a great honor to have her as part of my jury, and I am profoundly grateful for her thoughtful evaluation and support.
- Finally, my heartfelt thanks to **M. Valerio Massimo**, whose participation in the jury enriches this evaluation with a vital medical perspective that underscores the interdisciplinary nature of this work.

I would like to express my sincere gratitude to the four key institutions whose support has been instrumental to the success of my work :

- **IRCAD France**, for welcoming me into the Surgical Data Science team and providing the equipment, resources, and unwavering support essential to this research.
- **IRCAD Africa**, which has served as both the core inspiration and primary driving force behind my research project, as well as its founding supporter.
- The **Région Grand-Est de France**, for generously funding my research throughout this thesis.
- The **Université de Strasbourg** and the **École Doctorale MSI (ED 269)**, for their warm welcome and invaluable administrative support throughout my doctoral journey.

To **Benoit Sauer**, whose expertise and provision of data were fundamental to the success of this work, I owe my deepest thanks. I am also profoundly grateful to

Yvonne Keeza, **Grace Ufitinema**, and **Florien Ujemurwego** from IRCAD Africa for their invaluable expertise in ultrasonography and their meticulous efforts in labeling our datasets, which significantly enriched this research.

A special thank you to all the members of the **France-Rwanda Surgical Data Science Team**, who have been a constant source of support and camaraderie throughout these years. In particular, I would like to thank **Alexandre Hostettler** for his unwavering assistance and kindness in addressing every need I encountered during my thesis. I also extend my gratitude to **Jacques Marescaux** and the entire team at **IRCAD France**. While not directly involved in my research, their efforts and dedication made it possible for this work to thrive within the institute's inspiring environment.

Lastly, I wish to thank my family and friends, whose unwavering love and support have given meaning to this journey. Without them, none of this would have been possible.

Abstract (English)

B-mode ultrasound (US) is a preferred imaging modality for screening early-stage abdominal pathologies due to its cost-effectiveness and non-invasive nature. Many deep learning-based methods for abdominal US computer-aided diagnosis (CAD) have been proposed to reduce the elevated level of expertise needed for these tasks. While these methods show promising performance, they are typically trained on manually curated datasets consisting of high-quality, expert-selected images, which are accepted as ground truth without further validation. However, these methods have not been proven effective on raw (untrimmed) US video data, which presents greater challenges due to a larger proportion of low-quality images.

This work addresses the challenge of training deep learning models for use in US-based CAD of liver and kidney pathologies, covering the entire process from annotation to real-time inference in untrimmed videos. In our first contribution chapter, we compare human visual annotations with histological examinations and develop a method to improve the accuracy of visual annotations by 10% (F1-Score) using a combination of Learning to Rank with pairwise comparative annotations. In the second chapter, we focus on the problem of ultrasound diagnostics with untrimmed video data. We propose a novel solution to train video transformer models with the guidance of an external Relevant Frame Assessor (FRA), which automatically scores high-relevance frames according to contents and image quality. In the final contribution chapter, we replace the need for external guidance with a novel network architecture that learns relevance scores end-to-end solely based on video-level labels. We achieved a 0.90 ROC-AUC for diagnosing liver pathologies on untrimmed videos using only video-level diagnostic labels. Additionally, decision explainability is provided by identifying the most contributing frames used in the diagnosis, which can facilitate autonomous report generation.

Résumé (Français)

L'échographie b-mode (US) est une modalité d'imagerie privilégiée pour le dépistage des pathologies abdominales à un stade précoce en raison de son rapport coût/efficacité et de son caractère non-invasif. De nombreuses méthodes basées sur l'apprentissage profond pour le diagnostic assisté par ordinateur (CAD) des échographies abdominales ont été proposées afin de réduire le niveau élevé d'expertise nécessaire pour réaliser ces tâches. Bien que ces méthodes montrent des performances prometteuses, elles sont généralement entraînées sur des ensembles de données soigneusement sélectionnées, composées d'images de haute qualité choisies par des experts, qui sont acceptées comme vérité fondamentale sans validation supplémentaire. Cependant, ces méthodes n'ont pas démontré leur efficacité sur les données vidéo brutes (non-découpées) d'échographie, qui présentent des défis plus importants en raison d'une proportion plus élevée d'images de faible qualité.

Ce travail aborde le défi de l'entraînement de modèles d'apprentissage profond pou l'automation du diagnostic par échographique des pathologies du foie et du rein, couvrant l'ensemble du processus, de l'annotation à l'inférence en temps réel sur des vidéos non-découpées. Dans notre premier chapitre de contribution, nous comparons les annotations classiques réalisées par des annotateurs aux examens histologiques et développons une méthode permettant d'améliorer la précision des annotations visuelles de 10% (F1-Score) en utilisant une combinaison de l'apprentissage au classement (Learning to Rank) avec des annotations comparatives par paires. Dans le deuxième chapitre, nous nous concentrons sur le problème du diagnostic échographique à partir de données vidéo non-découpées. Nous proposons une solution innovante pour entraîner des modèles du type Transformer vidéo avec l'aide d'un évaluateur de qualité des images (Relevant Frame Assessor, FRA), qui attribue automatiquement des scores aux images de grande pertinence en fonction de leur contenu et de leur qualité. Dans le chapitre final, nous remplaçons le besoin d'un agent externe pour l'évaluation de la qualité des images par une nouvelle architecture de réseau de neurones capable d'apprendre des scores de pertinence de bout en bout en utilisant uniquement d'étiquettes au niveau de la vidéo. Nous avons atteint un ROC-AUC de 0.90 pour le diagnostic des maladies du foie sur des vidéos non non-découpées en utilisant uniquement des étiquettes diagnostiques au niveau de la vidéo. De plus, l'identification automatique des images les plus contributives pour le diagnostic permet la génération autonome de rapports.

Table des matières

Acronyms

ADDLE	*Auto-Decoded Deep Latent Embedding*
AI	*Artificial Intelligence*
AUC	*Area Under the Curve*
AUROC	*Area Under the ROC Curve*
BCE	*Binary Cross-Entropy*
BCEWithLogitsLoss	*Binary Cross-Entropy with Logits Loss*
BIQES	*Blind Image Quality Evaluation System*
CAD	*Computer-Assisted Diagnosis*
CAP	*Controlled Attenuation Parameter*
CEUS	*Contrast-Enhanced Ultrasound*
CNN	*Convolutional Neural Network*
CVL	*Comparative Visual Labeling*
DR-Clips	*Diagnostically Relevant Clips*
EHR	*Electronic Health Record*
eGFR	*Estimated Glomerular Filtration Rate*
FCN	*Fully Convolutional Network*
FOV	*Field of View*
FPS	*Frames Per Second*
FRA	*Frame Relevance Assessor*
GAN	*Generative Adversarial Network*
GLCM	*Gray-Level Co-occurrence Matrix*
GNN	*Graph Neural Network*
HCC	*Hepatocellular Carcinoma*
HSROC	*Hierarchical Summary Receiver Operating Characteristic*
IoU	*Intersection over Union*
IRCAD	*Institut de Recherche contre les Cancers de l´Appareil Digestif*
LTR	*Learning To Rank*
LSTM	*Long Short-Term Memory*
MAP	*Mean Average Precision*
MCD	*Monte Carlo Dropout*
MRI	*Magnetic Resonance Imaging*

MRI-PDFF	*Magnetic Resonance Imaging Proton Density Fat Fraction*
MViTv2	*Multiscale Vision Transformer Version 2*
NAFLD	*Non-Alcoholic Fatty Liver Disease*
NASH	*Non-Alcoholic Steatohepatitis*
NDCG	*Normalized Discounted Cumulative Gain*
NIQE	*Natural Image Quality Evaluator*
PCA	*Principal Component Analysis*
PIQUE	*Perceptual Image Quality Evaluator*
RF	*Radiofrequency*
ROC-AUC	*Receiver Operating Characteristic - Area Under the Curve*
ROI	*Region of Interest*
SVM	*Support Vector Machine*
SWE	*Shear Wave Elastography*
Swin Transformer	*Shifted Window Transformer*
SVL	*Single-Image Visual Labeling*
U-Net	*A Neural Network for Image Segmentation*
UVT	*Ultrasound Vision Transformer*
VideoMAE V2	*Video Masked Autoencoder Version 2*
ViT	*Vision Transformer*
WNNM	*Weighted Nuclear Norm Minimization*
WTAL	*Weakly-Supervised Temporal Action Localization*
X3D	*eXploring Xtra-large Depth*
US	*Ultrasound*

Liste des tableaux

Table des figures

1. Introduction

Context and problem motivation

Medical imaging has revolutionized healthcare by enabling practitioners to visualize internal structures and diagnose conditions that were previously undetectable. While modalities like Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) offer precise diagnostic capabilities, their high costs and radiation exposure make them impractical for large-scale screening.

In contrast, b-mode ultrasound (US) stands out as a more practical and advantageous alternative for the early detection and screening of various abdominal pathologies. Its unique benefits make it particularly well-suited for large-scale use, offering significant advantages over other imaging modalities. These benefits include :

1. **Real-Time Imaging :** US provides real-time visualization of internal structures, allowing assessment during the examination. Unlike CT or MRI, it does not require lengthy setup or acquisition times.
2. **Non-invasiveness :** b-mode US is a non-invasive modality, and does not require contrast agents or ionizing radiation, ensuring a safer experience for patients.
3. **Cost-Effectiveness :** Compared to other imaging modalities such as CT and MRI, US is more affordable and widely accessible. Additionally, the growing availability of portable ultrasound devices in the market can further reduces costs and expands access to care.
4. **High Sensitivity for Certain Pathologies :** US is highly effective in detecting certain pathologies, particularly those involving soft tissues and fluid collections, making it a valuable tool for targeted diagnostic tasks.

While the benefits of US screening are evident, several **challenges** still limit its widespread usability, including :

1. **Low Image Quality :** US images generally have reduced quality compared to other imaging modalities, with significant noise and artifacts such as shadowing and motion blur.

2. **Operator Dependent :** The quality of image acquisition heavily depends on the operator's skill in probe positioning, angle adjustment, and parameter settings.
3. **Subjectivity in Diagnosis :** The diagnosis of US data is inherently subjective and relies on the expertise of the practitioner. Even experienced radiologists can disagree on a diagnosis
4. **Diversity in Human Anatomy :** Variations in internal organ structures among individuals can complicate image acquisition and diagnosis. Additionally, certain pathologies may further challenge the process, making both tasks even more difficult.

All these challenges highlight the critical reliance on highly skilled professionals for both the acquisition and interpretation of US images. This dependency, aggravated by the global shortage of experienced radiologists, significantly restricts access to screening and early-stage pathology detection, particularly in low-resource settings where the need for affordable and accessible healthcare solutions is most needed.

A promising solution to this challenge is the integration of Computer-Aided Diagnosis (CAD) methods into ultrasound systems. CAD-assisted screening can enhance diagnostic accuracy and reduce reliance on expert radiologists, making it particularly useful for enabling less experienced healthcare professionals to identify suspicious cases that require further evaluation by an expert radiologist. This approach helps alleviate bottlenecks in the clinical workflow by improving the initial screening process and optimizing the use of radiologists' expertise. Consequently improving access to early detection and treatment of pathologies on a global scale.

This work addresses these challenges by proposing a CAD-based system to assist healthcare professionals in **diagnosing liver and kidney pathologies using b-mode US acquisitions**. In doing so, the system allows practitioners to benefit from all the advantages of B-mode ultrasound while minimizing its drawbacks.

Given the video nature of US acquisitions, such a system should support operators in two key aspects : identifying diagnostically relevant frames and automatically indicating suspicious findings. The first aspect involves helping practitioners recognize frames that are suitable for diagnostic purposes, while the second focuses on providing automated diagnostic assistance.

Limitations in the State-of-the-art

While CAD methods for assisting in the screening of abdominal pathologies have been proposed since the 2010s [2] and have demonstrated promising performance, these approaches are typically developed under highly controlled conditions. They rely on datasets that are manually curated by experts, with most focusing on single-image analysis. As a result, these datasets fail to represent the true distribution of raw, untrimmed ultrasound videos, which often include numerous non-diagnostic frames captured during the search for diagnostically relevant ones. Consequently, while these single-frame methods have shown impressive accuracy, their usability is limited (particularly for inexperienced operators) since the operator is still required to manually identify diagnostically relevant frames, a task that is itself one of the key challenges in ultrasound diagnostics.

In this context, video classification models present a clear and effective solution to the problem. These models are trained to classify video inputs, represented as ordered collections of image frames, using only video-level annotations. By relying on pathology reports as ground truth, video classification eliminates the need for manual image annotation, significantly reducing effort and avoiding subjective decisions about the identification of diagnostically relevant frames. Moreover, it can profit from multi-frame context during training and inference, enhancing performance, while ensuring alignment with the data distribution encountered during inference.

Most video classification models have been demonstrated primarily on short, time-cropped video clips, commonly referred to as trimmed videos. These clips are typically spatiotemporally centered around the pathological findings of interest. In contrast, real-world untrimmed videos present a more complex challenge, as the findings may appear at any point in time and are often not centered within the frame. An alternative to address this challenge involves in adopting external guidance agents, such as object detectors or segmentation networks, to extract trimmed clips from untrimmed videos. However, these methods requires substantial extra supervision, making them less practical for widespread use.

Additionally, the common reliance on radiologists' visual labeling as an annotation method compromises the reliability of ultrasound datasets. This is due to the inherent subjectivity of ultrasound interpretation, which introduces label uncertainty, potentially impacting the performance of CAD models trained on this type of data. Although alternatives such as histopathology analysis or MRI-based labeling could offer more objective annotations, their high costs and limited accessibility make them impractical for large-scale dataset acquisition.

Objectives and Contributions

Objectives

Our research aims to address these challenge by developing **deep learning models for US-based CAD targeting liver and kidney pathologies in untrimmed b-mode ultrasound videos**. Our objective is to propose novel methods capable of training and performing real-time inference with minimal supervision and annotation effort. We aim to design generalizable approaches that can be extended to other pathologies with ease, while providing insights into the feasibility of diagnostics based on the acquired data.

Additionally, we propose an approach to improve the reliability of visual annotations performed by annotators, which are the most commonly used method for obtaining labeled datasets. Given the inherent challenges in interpreting b-mode ultrasound data, these annotations are prone to errors that can negatively impact the training of deep learning models trained on such data. Addressing this issue is critical to ensuring the robustness and accuracy of CAD systems.

Contributions

Our objectives are translated into our three methodological contributions :

1. **CVL+RankNet : A New Approach to Label Images for Computer-Assisted Diagnosis :** In our first contribution, We propose an annotation method based on Comparative Visual Labeling (CVL) combined with a Learning-to-Rank framework to improve annotation reliability, called CVL+RankNet. We demonstrate that our proposed method, improve annotation accuracy by 10% when compared to standard Single-image Visual Labeling (SVL) approaches, which underestimates pathological levels by 20% when compared to the actual ground truth from histopathological labels.
2. **DR-Clips : A novel frame-guidance approach for computer-assisted diagnosis with untrimmed ultrasound video :** In our second contribution, we highlight the significant limitations of single-image-based methods in diagnosing liver and kidney pathologies from untrimmed ultrasound videos. To address these challenges, we introduce DR-Clips, a novel solution for assessing ultrasound pathologies in untrimmed videos using video-level annotations and an external Frame Relevance Assessor (FRA) to guide the video diagostic neural network. DR-Clips uses the FRA network to identify and sample diagnostically relevant clips from untrimmed ultrasound videos, employing these clips as a data augmentation tool during training and as guidance during inference. This innovative approach enables video classification models to be trained di-

rectly on untrimmed data, achieving results comparable with models trained on manually curated single-image datasets.

3. **KeyFrameDiagFormer : A Weakly-Supervised Transformer Model for Keyframe Localization and Diagnosis in Untrimmed Ultrasound Videos** In our final contribution, we advance our previous work by eliminating the reliance on an external agent to guide the video diagnostic model. Our proposed architecture is inspired by weakly-supervised action localization, enabling the localization of diagnostically relevant frames (analogous to actions) in time using only video-level labels. The model incorporates several key features : a memory-bank frame encoder to handle large video sequences, a local self-attention block for identifying organ-specific diagnostically relevant frames and distinguishing them from background frames, and a hierarchical classification head with organ-specific self-attention for pathology diagnosis based solely on relevant organ images. This approach demonstrates a strong ability to localize diagnostically relevant keyframes while also indicating when insufficient information is available to make a diagnosis.

Structure of the Thesis

This thesis is structured as follows :

— **Chapter 1 — Introduction :** This chapter presents the motivation, objectives, and key concepts underlying our work.

— **Chapter 2 — State of the Art :** A comprehensive review of the literature on single-image and video-based CAD methods using machine learning. Additionally, we discuss deep learning approaches that can serve as external guidance and quality control mechanisms for image and video models.

— **Chapter 3 — CVL+RankNet : Comparative Visual Labeling :** This chapter introduces our annotation method based on comparative visual labeling, as described in the Contribution 1.

— **Chapter 4 — DR-Clips : Guided Video Diagnosis :** This chapter described our DR-Clip guided untrimmed video diagnostic model, as described in the Contribution 2.

— **Chapter 5 — KeyFrameDiagFormer : Unguided Video Diagnosis :** This chapter described our unguided untrimmed video diagnostic model, as described in the Contribution 3.

— **Chapter 6 — Conclusion :** This chapter summarizes our contributions, discusses the implications of our findings, and suggests directions for future research.

2. Literature Review

This literature review explores various approaches to developing CAD systems for ultrasound, ranging from single-image methods to video-based techniques, while also discussing the integration of auxiliary guidance systems to improve diagnostic accuracy.

1. **Single-Image-Based Method :** These techniques utilize machine learning methods to analyze individual ultrasound frames. While effective in controlled settings, they often fail to capture the full diagnostic potential of an ultrasound exam, which typically requires contextual and temporal information across multiple planes. We also discuss about single-image selection and image quality control methods, which are used as external agents in combination with diagnosis methods.
2. **Video-Based Methods :** Video-based CAD models address the limitations of single-image approaches by incorporating temporal and spatial context from ultrasound video sequences. These methods process consecutive frames to improve diagnostic accuracy and reliability, using deep learning architectures to model temporal dependencies and spatial relationships. They also reduce annotation efforts by allowing video-level labeling rather than frame-by-frame annotation.
3. **Research Gaps :** Despite advancements, several challenges persist :
   - Human annotation subjectivity reduces data reliability.
   - The processing of untrimmed ultrasound videos remains a significant challenge and is relatively underexplored in the existing literature.
   - Current CAD systems lack the automation and usability needed for seamless integration into diverse clinical workflows.

2.1 Single-Image Ultrasound Computer-Assisted Diagnosis

These methods use manually curated single-image datasets, which are formed from snapshots taken from the clinician during US procedures or manually curating ultrasound videos acquired retrospectively.

2.1.1 Methods using Image Classification and Regression

For the diagnosis of ultrasound images, one of the most straightforward approaches is using single-image classification and regression models. These models take an ultrasound image as input and output either a continuous value score, indicating the severity of the condition, or a categorical class, representing the stage of the condition.

Historically, these methods have utilized handcrafted features based on texture, color, and shape, which are then applied to train classical bayesian models [228]. However, since the development of modern convolutional neural networks (CNNs) with the introduction of AlexNet [137], most single-image approaches have shifted toward CNN-based models. A diverse range of CNN architectures is available, such as Inception-ResNet-v2 [96], GoogLeNet [234], AlexNet [137], ResNet-101 [97], MobileNet v2 [211], Xception [49], and VGG-19 [224]. Despite differences in their architectures, these models share a fundamental approach : they use convolutional layers combined with pooling operations to progressively reduce the dimensionality of the input image. This compression simplifies the data into a manageable form, which is then passed into fully connected (dense) layers. The output from these dense layers is then designed to suit the specific task, aligning with the structure of the training labels and the training loss function, as detailed in Table 2.1.

Table 2.1 – Configuration Details for CNN Model Architectures Based on Task Type. While other configurations of neurons in the last layer, activation functions, and loss functions can be used, these are the most common.

Problem	Neurons	Activation	Loss Function	torch.nn function
Binary Classification	1	Sigmoid	Binary Cross Entropy Loss	BCELoss()
Multiclass Classification	class number	Softmax	Cross Entropy Loss	CrossEntropyLoss()
Regression	1	None	L1/L2 Loss (or others)	L1Loss(), torch.nn.MSELoss()

Notably in Table 2.1, binary classification and regression models may have more than a single neuron in the output layer when handling multi-label problems where multiple outputs are expected. In this case, an individual loss is computed for each output and them combined. This approach can also be extended to multiclass classification by applying the loss individually to each set of outputs, although this is less commonly encountered.

While CNN models continue to evolve, the emergence of Vision Transformers (ViTs) [64] in the 2020s has positioned them as potential successors to CNNs. While ViT models consistently dominate image classification leaderboards nowadays [9, 180], it remains unclear whether ViTs consistently outperform CNNs in ultrasound diagnosis. Evidence suggests that ViTs may indeed surpass CNNs in performance, but this advantage appears significant only when large datasets are available. Under low-data conditions, common in the medical field, CNNs often demonstrate superior performance due to their inductive bias, making them more efficient with limited training data [14, 248]. This explains why most single-image classification and regression methods remain CNN-based, as CNNs have consistently shown reliable performance, leading to less interest in exploring transformer models.

Given that ultrasound acquisitions typically produce videos or a set of snapshot images as output, this data first need to be curated to be analyzed on a single-image basis. This step is typically performed manually and can incorporate anatomical and medical prior knowledge to aid in solving the problem, although this comes at the cost of increased data curation effort. Below, we list groups of methods identified in our literature review, categorized by the level of curation and the nature of data available for single-image neural networks :

- **Region of Interest (ROI)-Based Diagnosis :** This approach focuses on specific regions within the organ, such as the liver parenchyma, to simplify the localization of relevant structures. By targeting these specific areas, the methods aim to enhance diagnostic accuracy while reducing computational complexity.
- **Whole Image Diagnosis :** this method utilizes entire ultrasound images for analysis. While it captures more contextual information about the organ and surrounding tissues, it adds complexity in localizing the regions of interest within the image.
- **Multiview Diagnosis :** these methods incorporate multiple ultrasound views of the organ to enhance diagnostic performance. By using images from different planes, they provide a more comprehensive understanding of the organ's anatomy by incorporating complementary information, making it easier to achieve reliable results.
- **Additional Modalities Diagnosis :** some approaches integrate other imaging modalities or data types, such as radiofrequency (RF) signals or quantitative ultrasound (QUS), to improve diagnostic accuracy. By combining different types of data, these methods can capture additional tissue characteristics not visible in b-mode images.

Below, we present the most relevant works within these categories, organized by the specific pathologies being diagnosed.

2.1.1.1 ROI-Based Diagnosis

Region of Interest (ROI)-based methods are among the most widely used approaches for pathology detection in ultrasound and continue to be a focus in recent research. The strength of these methods lies in simplifying the machine learning task by narrowing down the area of analysis, thereby reducing the effort needed to identify highly discriminative regions.

Although various approaches and protocols are used, the general process for ROI-based diagnosis includes the following steps (illustrated in Figure 2.1) :

1. **ROI Extraction :** In this initial step, one or more ROIs (also referred to as patches) are selected to delimit the organ or structure of interest, typically through manual or semi-automatic means. These ROIs may be defined using bounding boxes or segmentation masks.
2. **Feature Extraction from Patches :** Next, mathematical descriptors are computed from each patch to form feature vectors, which capture aspects such as color, shape, and texture. Common feature extraction techniques include CNNs, GLCM, wavelet transforms, and Gabor filters.
3. **Patch Processing or Feature Fusion :** At this stage, two main approaches are possible : each patch can be processed independently, generating a prediction for each, or the feature vectors from all patches can be concatenated or fused to form a unified input for further processing.
4. **Final Image Classification :** In the final step, predictions from the previous step are combined to produce an overall classification for the entire image. This may involve majority voting among individual patch predictions or applying machine learning to the fused features, resulting in a final image class or score.

Below, we describe key methods for autonomous diagnosis of various organs in ultrasound images. Although ROI-based methods date back to 1996 [121], we focus on approaches from 2017 onward, which build on earlier advancements.

ROI-Based Diagnosis : Liver

Beginning with the research developed in [159], the authors propose a method for the automatic diagnosis of liver cirrhosis using the features extracted from the liver capsules in ultrasound images. The liver capsule, also known as Glisson's capsule, refers to the boundary separating the liver parenchyma from surrounding structures. In healthy subjects, it appears thin and uniform ; however, in patients with cirrhosis, it can become thickened and uneven due to fibrotic changes. To

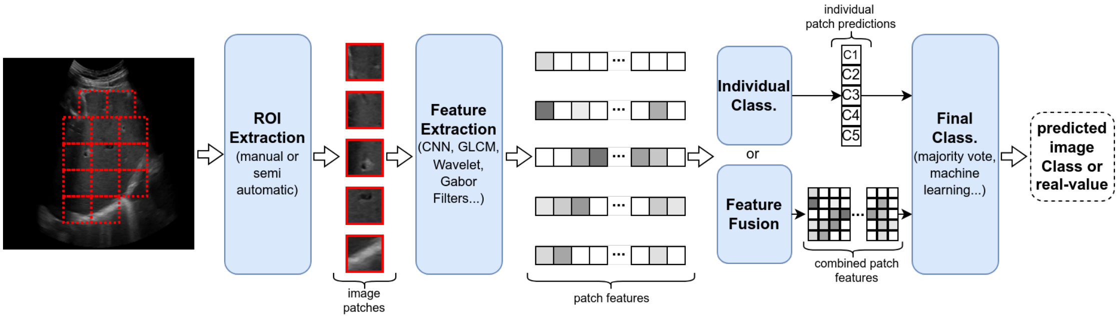

ROI Extraction
(manual or semi automatic)
image patches
Feature Extraction
(CNN, GLCM, Wavelet, Gabor Filters...)
patch features
individual patch predictions
C1
C2
C3
C4
C5
Individual Class.
or
Feature Fusion
combined patch features
Final Class.
(majority vote, machine learning...)
predicted image Class or real-value

FIGURE 2.1 – Overview of a Region of Interest (ROI)-based approach for pathology detection in ultrasound images. The process begins with ROI extraction, where specific areas of interest are defined, typically using bounding boxes or segmentation masks. Features are then extracted from each ROI, capturing relevant patterns such as texture and shape. In the next step, these features are either individually processed or fused into a unified feature set. Finally, the image classification stage combines the individual or fused predictions to produce a final pathology classification, which may be in the form of a categorical label or a continuous score.

detect the liver capsule, they first apply feature descriptor operations around a 30×30 pixel sliding window across the ultrasound image and classify these windows using depth-2 decision trees. They then filter out false positives by retaining only the ROIs that contribute to forming a liver capsule-like shape, effectively isolating the capsule for analysis.

The detected liver capsules are used to define a ROI containing the entire capsule boundary within the image, which is used for the classification of liver cirrhosis. They employ a CNN to compute features from the extracted ROI, followed by a SVM classifier to achieve the final classification scores. Using this method, they achieved an AUC of 0.951, indicating high accuracy in diagnosing cirrhosis.

A well-know work using ROI-based approach for the classification of liver steatosis is [20]. In their work, the authors propose a custom 22-layer CNN to classify liver images as either normal or steatotic. In their work, they manually define a ROI comprising all the liver parenchyma pixels, removing all the rest from the image. Their dataset consists of images from the right lobe of the liver, which were manually curated from a cohort of 63 patients (36 with steatosis and 27 diagnosed as normal). Ground-truth labels were obtained from biopsy reports. The model was evaluated using 10-fold cross-validation, achieved a perfect AUC score of 1.0, underscoring the effectiveness and promise of region-based diagnostic methods as early as 2018.

Following their work, in [285] the authors developed a custom CNN trained on liver patches to classify varying degrees of steatosis. For each of the four classes representing the severity stages, they collected 500 patches for training and evaluation, although the specific labeling protocol was not detailed. The model achieved a classification accuracy of 90%. Similarly, [56] extracted visual features from manually defined ROIs within the liver using specialized software [215, 232] and applied classical machine learning techniques for steatosis classification in pediatric patients. Labeling was performed by a board-certified pediatric radiologist, and the method yielded AUC scores exceeding 0.933.

In 2020, the authors [44] applied a VGG-16 network [224] to diagnose hepatic steatosis by analyzing manually selected patches of liver parenchyma. The model achieved AUC scores of 0.71, 0.75, and 0.88 for distinguishing mild, moderate, and severe steatosis, respectively. Additionally, they incorporated Shannon entropy-based imaging to quantify microstructural variations in the tissue from ultrasound backscatter signals, which enhanced the diagnostic performance. With this method, the model's AUC scores improved to 0.68, 0.85, and 0.9 for mild, moderate, and severe steatosis classifications, respectively.

In their 2022 work [76], the authors combine ROI-based feature extraction with genetic algorithms to enhance steatosis classification in ultrasound images. They employ a genetic algorithm [268] to identify optimal ROIs by adjusting the position, size, and classification thresholds specific to each ROI. At each iteration, GLCM [206] and statistical features are extracted and used to train independent Extreme Learning Machine (ELM) classifiers [32]. The genetic algorithm's fitness function minimizes the number of incorrectly classified ROIs, while the classifier's objective function maximizes correct classifications through majority voting across ROIs. Their dataset consists of 300 ultrasound images from different patients, with diagnostic labels and initial ROI positions for the genetic algorithm manually annotated by three experienced radiologists. Achieving 95.71% accuracy, this approach demonstrates the potential of dynamically optimized ROI placement, inspiring future advancements in adaptive ROI-based diagnostic models.

Lastly, various other studies adopt similar approaches in one or more aspects for the classification of liver pathologies [2, 201, 184, 6, 203, 139, 225, 1, 208]. Interested readers are encouraged to refer to these publications for further details.

ROI-Based Diagnosis : Kidney

Starting with the method proposed by [138] in 2019, the authors trained regression and classification neural networks to diagnose patients with CKD. They developed a model to estimate glomerular filtration rate (eGFR), a key indicator of CKD. The eGFR is a clinical metric calculated using formulas that measure kidney function by quantifying the amount of waste products in the blood. Additionally, eGFR can be used to generate binary labels for CKD classification, with values less than 60 mL/min/1.73m^2 indicating the presence of the disease. The authors collected a dataset of 37697 ultrasound images and manually selected those containing the whole kidney with kidney length annotations, resulting in a final set of 4505 high-quality kidney ultrasound images. These images were then cropped to obtain kidney ROIs for training the neural networks. eGFR values were obtained within four weeks before or after ultrasound acquisition.

They used a ResNet-101 backbone [97], freezing the initial layers and modifying the final layer to a single neuron with linear activation for estimating continuous-valued eGFR scores. Additionally, they applied an eXtreme Gradient Boosting (XGBoost) model [47] using features extracted from the backbone to train a binary classifier for CKD. This approach achieved a Pearson correlation coefficient of 0.741 between predicted and actual eGFR values, with a Mean Absolute Error (MAE) of 17.605 mL/min/1.73m^2. The binary classification yielded an AUC score of 0.904, demonstrating the method's precision.

In [294], the authors developed a deep learning approach to detect congenital abnormalities of the kidney and urinary tract (CAKUT) in children. The input ultrasound images are first segmented using a graph cuts method [295], a feature-based segmentation technique optimized for kidney segmentation in ultrasound images by utilizing color, shape, and texture features extracted with Gabor filters [165, 58].Although this approach avoids manual ROI selection, it relies on a highly specialized hyper-parametrization that requires meticulous and labor-intensive feature engineering. Their kidney-ROI classification method combines features extracted from CNNs [137] with conventional imaging features such as geometrical features and Histogram of Oriented Gradients (HOG) [55]. These combined features are then used to train a Support Vector Machine (SVM) classifier to distinguish between pathological and healthy kidneys. The study cohort included 50 healthy subjects and 50 patients with CAKUT, with labels obtained from medical records. The proposed method achieved AUC scores of 0.92 for the left kidney and 0.88 for the right kidney, showcasing its effectiveness.

Similarly, [42] used ROIs defined by manually segmented kidney masks (whole kidney, kidney parenchyma, and central sinus) to extract GLCM features [93] specific to each mask structure. These features were combined with morphological and statistical characteristics features to create the input vector for an SVM model. However, using a dataset of 798 ultrasound images, the model achieved a modest accuracy of 80%.

In [128], the authors propose a method for CKD diagnosis using features extracted from specific kidney ROIs. They define three distinct ROIs within the kidney : the Renal Cortex, the Cortex-Medulla boundary, and the Renal Medulla. From each ROI, 19 GLCM features [93] are extracted, resulting in a total of 57 kidney features per image. These features are forwarded into a 10-layer dense CNN with a softmax output layer to classify images into three categories : normal kidney, mild CKD, and severe CKD. Ground-truth labels were based on eGFR measurements, and using a dataset of 741 ultrasound images, the model achieved an accuracy of 95.4%

Lastly, in [3], a similar method was developed, focusing on a single kidney ROI and extracting 14 GLCM features [93]. Using a dataset of 700 ultrasound images with eGFR-based ground-truth labels, this approach achieved an accuracy of 97.56%.

2.1.1.2 Whole Image for Diagnosis

Whole-image methods offer a straightforward approach for automatic abnormality detection in ultrasound images. By using entire ultrasound images as input and leveraging standard neural network architectures, these methods are easy to implement, significantly reducing development time. However, this reduction in engineering workload comes at the cost of requiring larger training datasets compared to ROI-based methods. This is because the neural network must learn not only to identify relevant regions but also to ignore irrelevant areas within the image, making effective training more data-intensive.

The general process for Whole-image methods includes the following steps (illustrated in Figure 2.2) :

1. **Backbone Feature Extraction :** A neural network backbone is used to extract features from the entire image. The most commonly used architectures for this purpose are CNNs and Transformers. Given the high complexity of these models, which typically have millions of parameters, pretrained backbones on large datasets such as ImageNet [59] are often used to enhance performance. In many cases, the initial backbone layers are frozen (i.e., kept unchanged during training), allowing only the last layers to be trained.
2. **Classification/Regression Head :** After feature extraction, a classification or regression head transforms the high-dimensional feature map into either categorical or real-valued outputs depending on the training task. Popular choices for this stage include dense (fully connected) layers, as well as SVM or logistic regression, depending on the task. In some Transformers models they can use one of the input tokens for the classification, called the "[CLS]" token (classification token).

Below, we present the most relevant works within these categories, organized by the specific pathologies being diagnosed.

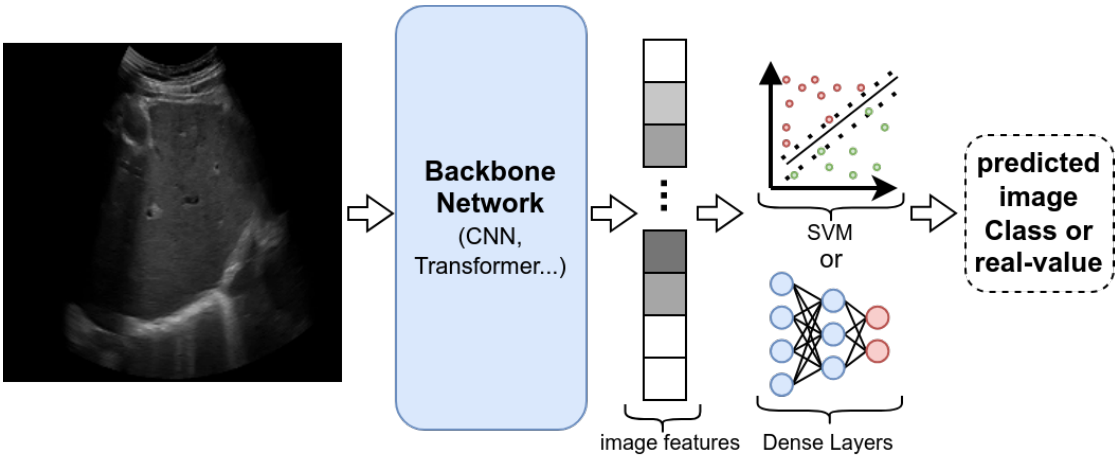


FIGURE 2.2 – Workflow for whole-image methods in ultrasound abnormality detection. These methods use a backbone network (e.g., CNN or Transformer) to extract global image features, followed by a classification or regression head that outputs a diagnosis, such as a categorical label or real value.

Whole Image for Diagnosis : Liver

A landmark 2018 contribution in this category for the assessment of steatosis through single-image analysis is Byra's study [29], notable for introducing a publicly accessible steatosis dataset. The authors utilized an Inception-ResNet-v2 model [96], pre-trained on the ImageNet dataset [59], to extract relevant image features from ultrasound scans. These features were then used to train a SVM to classify liver ultrasound images as either healthy or pathological.

The dataset consists of images from 55 patients, all captured in the liver-kidney plane using standard ultrasound machines with medium-quality resolution. Ground-truth labels were obtained via histopathological analysis, where steatosis was defined as the percentage of hepatocytes exhibiting fatty infiltration. Despite the relatively small dataset size, the study reported near-optimal performance, achieving an impressive AUC score of 0.977, highlighting the efficacy of their approach. This dataset has since become a valuable resource, used widely by other researchers, including ourselves..

Building on this foundation, another notable study from 2018 [194] applied a similar approach using a VGG-16 model [224] for steatosis prediction. The model was trained on a dataset comprising 81 normal liver images and 76 images from patients with steatosis, acquired at multiple planes. These images were labeled by two experienced radiographers, and the trained model achieved an accuracy of 90.6%.

In 2021, Zamanian et al. [278] advanced Byra's methodology by combining multiple neural networks for feature extraction. Each image was processed by four different pre-trained models : Inception-ResNet-v2 [96], GoogLeNet [234], AlexNet [137], and ResNet-101 [97]. The features from each model's output were aggregated and normalized before being used to train an SVM for steatosis classification. Using the same dataset as Byra's study, they significantly improved the AUC score from 0.977 to 0.999, demonstrating the effectiveness of combining multiple feature extractors.

Similarly, [50] explored the application of several single-image neural networks to classify the severity of steatosis (normal, mild, moderate and severe) using b-mode ultrasound images. Their dataset consisted of 21855 images from 2080 patients, with ground-truth diagnoses confirmed by at least two gastroenterologists. The authors evaluated various CNN architectures, including VGG-19 [224], ResNet-50 v2 [96], MobileNet v2 [211], Xception [49], and Inception v2 [112]. ResNet-50 v2 emerged as the top-performing model, surpassing the others and achieving AUC scores of 0.974 (mild steatosis vs. others), 0.971 (moderate steatosis vs. others), 0.981 (severe steatosis vs. others), 0.985 (any severity vs. normal), and 0.996 (moderate-to-severe vs. normal-to-mild).

Building on the success of previous ROI-based diagnosis methods, the authors in [200] propose a three-step approach to classify hepatic steatosis in ultrasound images by automatically isolating the liver-kidney (sagittal) ROI. To automate the selection of the ROI, they employ a DeepLabv3+ semantic segmentation network [45] to isolate and crop the liver and kidney regions. The model, trained on 2650 manually annotated segmentation masks, achieves precise segmentation within the sagittal plane.

After segmentation, two Inception V3-based neural networks [235] are used to further refine the dataset. The first network classifies ultrasound images into parasagittal or non-sagittal planes, while the second detects a ring-shaped contour surrounding the kidney cortex (a defining feature of parasagittal images). Using these two networks, the authors ensure that only high-quality images from the liver-kidney plane are selected, what also filters erroneous segmentation masks from the previous step. The result is a curated dataset of ultrasound images that contain only the liver and kidney regions.

This dataset is then used to train their final model, nammed SteatosisNet, their final classification network. SteatosisNet, a modified version of Inception V3, classify steatosis severity into four levels : normal, mild, moderate, and severe. The method flowchart is presented in Figure 2.3. The authors report a perfect classification accuracy (100%), sensitivity, and specificity. However, despite this impressive performance, the approach has several limitations, such as the need for extensive manual annotations and its reliance on consistently capturing the liver-kidney plane during ultrasound exams. Additionally, the method's heavy reliance on specialized design choices tailored for steatosis classification reduces its generalizability to other liver diseases or different medical imaging tasks. This specificity, while beneficial for steatosis diagnosis, limits its broader applicability without significant modifications.

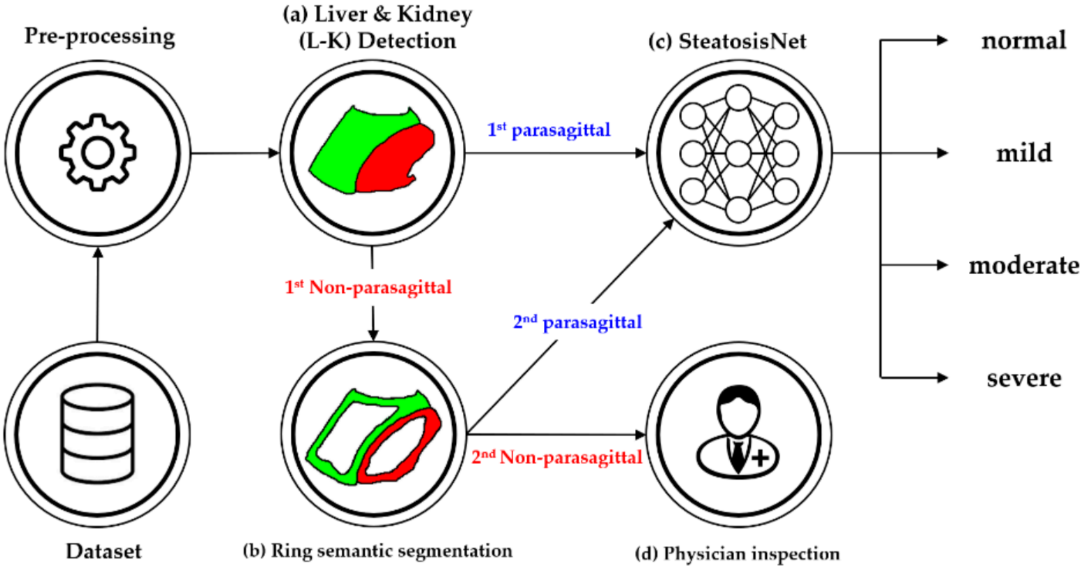


FIGURE 2.3 – Simplified flowchart of the proposed method : (a) Liver and kidney (L-K) detection yielding 1st parasagittal and non-parasagittal images. (b) Ring detection to double-check the 1st nonparasagittal image, producing 2nd parasagittal and non-parasagittal images. (c) Grading of the 1st and 2nd parasagittal images using SteatosisNet according to the level of steatosis. Figure and description from the authors [200]

In [120], the authors investigate the use of multiple image classification networks to classify the severity of liver fibrosis using ultrasound images. They evaluated the performance of VGG-16 [224], ResNet-50 [96], DenseNet-121 [103], EfficientNet-B0 [236], and ViT [64]. The training dataset included 766 patients, with each image manually labeled by an expert radiologist into one of five classes : no fibrosis (F0), portal fibrosis (F1), periportal fibrosis (F2), septal fibrosis (F3), and cirrhosis (F4). ResNet-50 achieved the highest accuracy, with 85.92% for five-level classification (F0-F4) and 87.92% for three-level classification (combining F1, F2, and F3 into one class). These results demonstrate that ResNet-50 remains competitive with more modern architectures like ViT and EfficientNet. An ablation study further showed that models trained on images from specific ultrasound machines developed biases toward those machines, impacting generalizability.

In [66], the authors propose an ensemble neural network approach for classifying two types of liver lesions in US images : hemangioma and hepatocellular carcinoma (HCC). The training dataset includes 350 US images from 59 patients,

Vision Transformer (ViT)

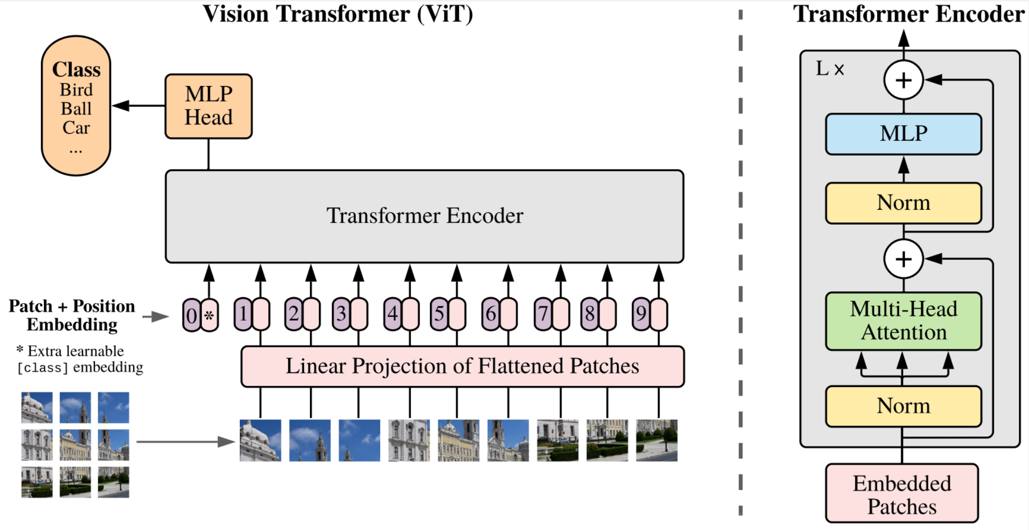


FIGURE 2.4 – ViT Model Overview

The ViT model [64] transforms an input image into a collection of tokens, analogous to text inputs in language models. The image is divided into N non-overlapping patches of size $P \times P$, which are flattened ($\mathbf{p}_i$) and mapped to an embedding dimension d using a linear layer :

$$\mathbf{E}_i = \mathbf{p}_i \mathbf{W}_E + \mathbf{b}_E, \quad \mathbf{E} \in \mathbb{R}^{N \times d} \tag{2.1}$$

Positional embeddings are then added based on patch location :

$$\mathbf{X} = \mathbf{E} + \mathbf{P}_{\text{pos}}, \quad \mathbf{X} \in \mathbb{R}^{N \times d} \tag{2.2}$$

The attention mechanism follows the same principle as in text processing. The input $\mathbf{X}$ is projected using learnable matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ for queries, keys, and values :

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}_V \tag{2.3}$$

These projections compute the Attention Weights (Att) :

$$\text{Att} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_{\text{head}}}}\right) \tag{2.4}$$

The weights are used to combine patch values :

$$\mathbf{z} = \text{Att} \cdot \mathbf{V} \tag{2.5}$$

For Multi-Head Attention, the input $\mathbf{X}$ is split into n heads processed in parallel. The outputs are concatenated and projected by a linear layer $\mathbf{W}_o$ to match the original dimension d.

$$(\mathbf{X}) = \text{Concat}(\text{z}_1, \dots, \text{z}_n)\mathbf{W}_O \tag{2.6}$$

After multiple transformer blocks, the final features are used for vision tasks. Transformers excel by combining information from different image regions, which is critical for tasks requiring global context [248, 281, 166].

consisting of 202 images of HCC and 148 of hemangioma. Multiple CNN-based architectures were evaluated, including VGG-16/19 [224], DenseNet [103], Inception [235], InceptionResNet, ResNet [96], and EfficientNet B0-B7 [236], with the entire image used for classification without cropping around the lesion. The three best-performing networks (DenseNet201, DenseNet169, and ResNet152V2) were then combined in an ensemble model by averaging their outputs. This ensemble model achieved an AUC score of 0.944, demonstrating that effective lesion classification is achievable without the need for lesion cropping.

Lastly, various other studies uses Whole Images for the classification of liver pathologies [87, 222, 52, 33, 40, 170]. Interested readers are encouraged to refer to these publications for further details.

Whole Image for Diagnosis : Kidney

In [229], the authors propose an ensemble neural network model to classify kidney pathologies and differentiate them from healthy kidneys. To ensure to forward only high-quality images to the deep learning model, they use the PIQUE score [253] to select only images having a noise score below 50, indicating low noise levels. While the PIQUE score doesn't guarantee anatomical feature presence, it does ensure minimal image degradation. This method was use to obtain a dataset of 4940 b-mode kidney ultrasound images from retrospective data, being 10% reserved for evaluation and 90% for training. Labels were assigned through visual inspection by expert radiologists, categorizing each image as either a normal kidney, renal cyst, renal stone, or renal tumor ; which are presented in the Figure 2.5.

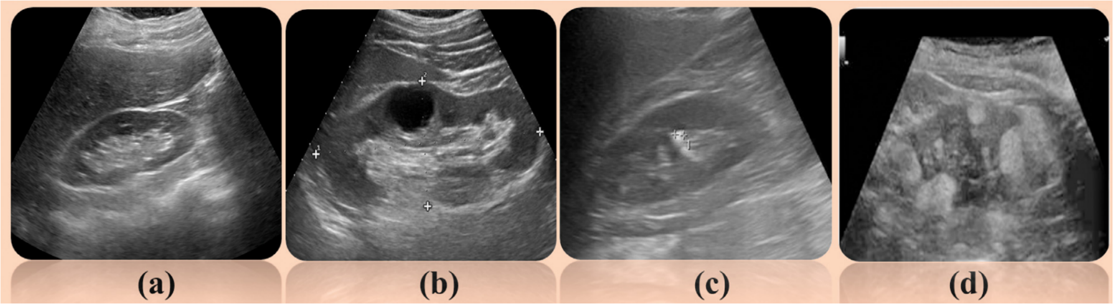


Figure 2.5 – Examples of kidney abnormalities as categorized by [229] : (a) Normal kidney, (b) Kidney with cysts, (c) Kidney with stones, (d) Kidney with tumor. Figure extracted from their article.

Their strategy involves an ensemble approach utilizing pre-trained CNN models : ResNet-101 [97], MobileNet v2 [211], and ShuffleNet [289]. Each CNN extracts features that are used to train independent SVM classifiers. The final classification is determined by majority voting across the three CNN-based SVM outputs. This method achieved an accuracy of 95.58%, highlighting the effectiveness of their approach.

2.1.1.3 Multiview Diagnosis

Single-view automatic diagnosis methods in ultrasound (US) have shown promising results ; however, they may underperform if the selected image lacks distinctive or sufficient diagnostic features. To address this limitation, many researchers have introduced approaches that use multiple pre-defined ultrasound views (or planes) as input to machine learning models. By combining information from different views, these methods enhance diagnostic accuracy by analyzing distinct structures and examining a larger portion of the organ. This approach is an intermediate step toward fully volumetric (3D or video) ultrasound data, although it typically relies on manually selected planes.

The overall process for Multiview Diagnosis has many parallels with that of ROI-Based Diagnosis, where pre-defined ROIs correspond to pre-defined views. The steps are as follows (illustrated in Figure 2.6) :

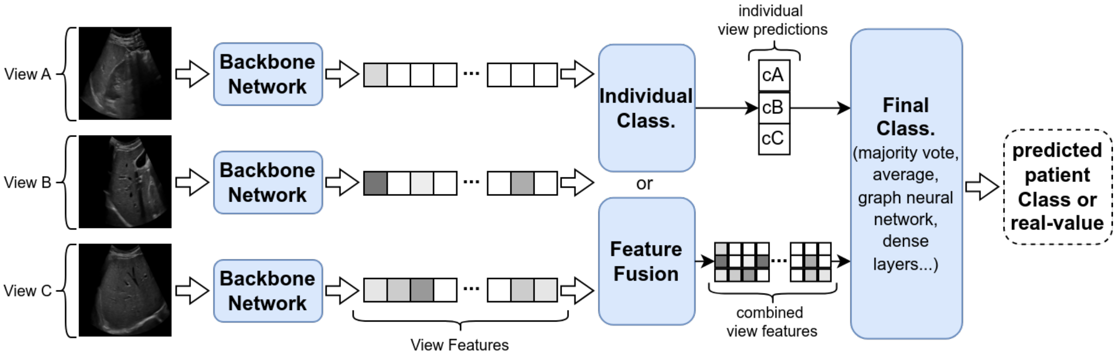

View A
View B
View C
Backbone Network
Backbone Network
Backbone Network
View Features
Individual Class.
or
Feature Fusion
individual view predictions
cA
cB
cC
combined view features
Final Class. (majority vote, average, graph neural network, dense layers...)
predicted patient Class or real-value

FIGURE 2.6 – Multiview diagnosis approach for ultrasound, where multiple pre-defined views are processed by backbone networks to extract features. These features are either classified individually or combined through methods like averaging or neural layers to produce a final diagnosis, leveraging complementary information across views.

1. **Feature Extraction from Views :** In this initial step, each view is processed by a backbone network model, typically a CNN, to extract relevant features. Each view model can either have independent weights or share weights across all views.
2. **View Processing or Feature Fusion :** Next, features from each view are either passed independently to a classification/regression head (end-to-end training), concatenated, or transformed through methods like Graph Neural Networks for enhanced information sharing.
3. **Final Diagnosis :** In this final step, view-specific predictions can be combined through majority voting or averaging, or the concatenated/transformed

features can be fed into a final classification/regression head to generate the final diagnosis.

Below, we present the most relevant works within these categories, organized by the specific pathologies being diagnosed.

Multiview Diagnosis : Liver

In the study by Colantonio et al. [51], a custom multiview CNN is introduced, designed to predict continuous-valued steatosis scores using two input branches. These branches correspond to two different scan views : the oblique subcostal view (AR) and the supine/left lateral view (HR). Ground-truth labels for model training were obtained using Hydrogen Magnetic Resonance Spectroscopy (H-MRS), a highly accurate method that measures the ratio of fat to water by detecting hydrogen atoms in liver tissues.

It is important to note that H-MRS and Magnetic Resonance Imaging Proton Density Fat Fraction (MRI-PDFF) are distinct techniques. While MRI-PDFF quantifies the proton density of fat relative to water across the entire liver, providing a whole-liver fat percentage, H-MRS offers localized measurements by analyzing the chemical composition of fat and water in specific regions. Both methods are highly accurate, but H-MRS provides more detailed spectroscopic analysis. In a comparative study [62], the authors evaluated both techniques using histopathological results as the ground truth, demonstrating that H-MRS slightly outperformed MRI-PDFF in terms of sensitivity (92.6% vs. 89%), specificity (95.7% vs. 88%), and Pearson correlation (0.68 vs. 0.63).

The method proposed by the authors achieved promising results, with a root mean square error (RMSE) of 1.11 and a standard deviation of 0.77, underscoring the utility of multiview inputs in improving hepatic fat quantification accuracy.

Similarly, Byra et al. [28] expanded upon their previous research [29] by implementing a multiview approach for hepatic steatosis classification. In this study, they analyzed four distinct liver views : three in the transversal plane (hepatic veins, portal vein, and right posterior portal vein) and one in the sagittal plane (liver-kidney interface). A sample set of images for each view, shown in Figure 2.7, is drawn from their study. The dataset consists of 135 patients diagnosed with NAFLD, with ground-truth labels derived from MRI-PDFF, different from their previous work they used histopathology results as ground-truth.

The authors designed regression and classification view-specific models and an ensemble model for the combined views, utilizing the same Inception-ResNet-v2 architecture [96] used in their previous work. To fuse the information from different views, the authors averaged the outputs from each view-specific model to gene-

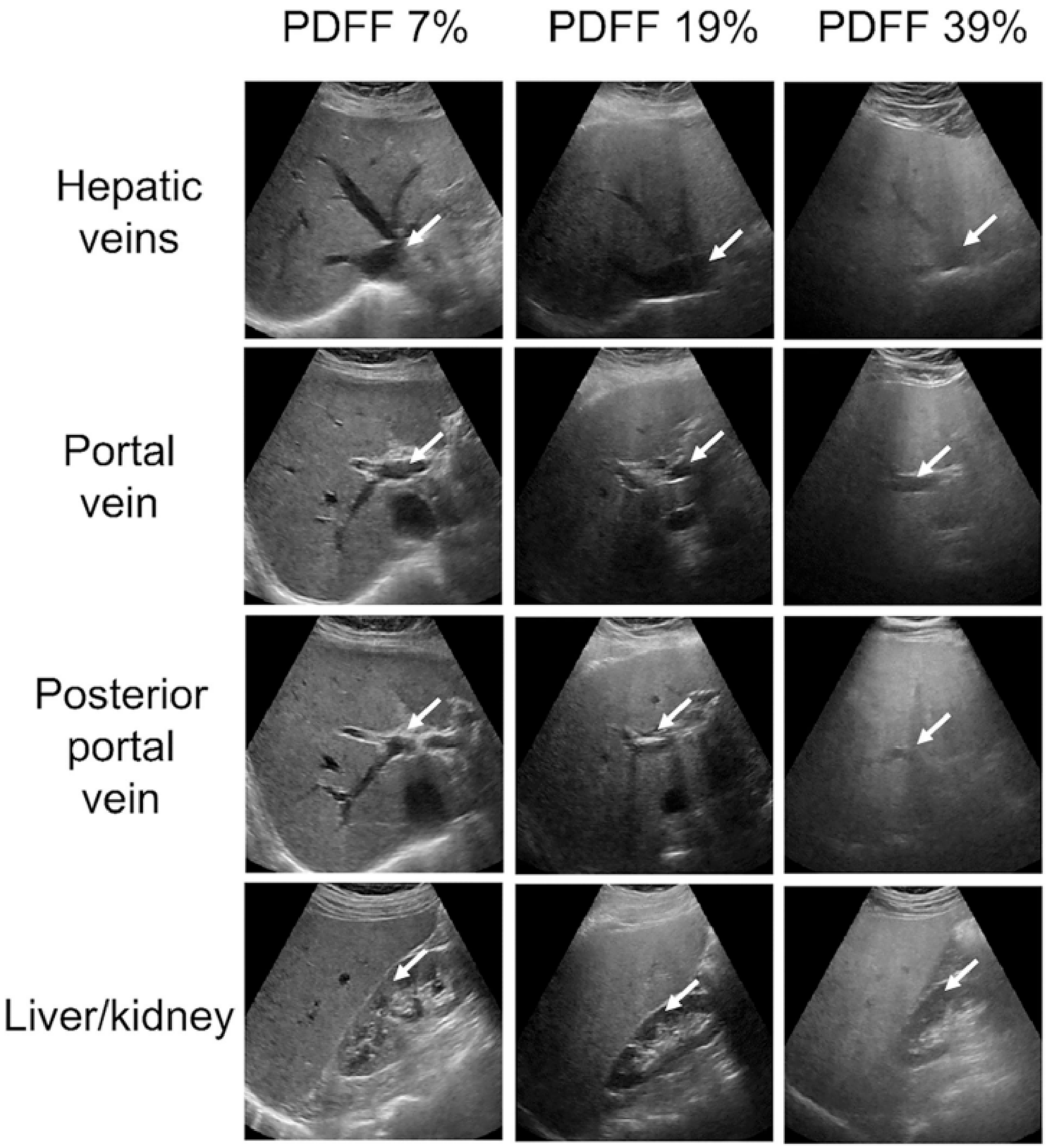


Figure 2.7 – Ultrasound images of livers with different PDFF values (7%, 19%, and 39%), corresponding to increasing fat accumulation. The blurring of veins and the liver/kidney interface becomes more pronounced as fat accumulation increases. White arrows indicate blood vessels and the kidney region. Image and description adapted from [28].

rate a multiview prediction. The best performing single-view model was the right posterior portal vein, achieving an AUC score of 0.90 and a Pearson correlation coefficient of 0.78. The multiview ensemble method offered a marginal improvement, raising the AUC to 0.91 and the Pearson correlation to 0.81, indicating the value of combining multiple views for enhanced steatosis classification accuracy.

Following a similar approach, Li et al. [144] present a deep learning framework designed for multiview liver steatosis prediction. The authors retrospectively collected ultrasound images from various anatomical regions of the liver, focusing on the following key views : left liver lobe (longitudinal and transverse views), right liver lobe (right lobe intercostal view), liver-kidney contrast (lower right lobe intercostal view), and subcostal view (with kidney and hepatic veins). This comprehensive image collection enabled the development of a robust training set for their predictive model. They employed a ResNet18 [97] architecture, configured for single-image multiclass classification, to categorize liver steatosis into four degrees of severity. For view-specific predictions, the authors averaged the prediction scores across all images associated with each view. To generate a multiview prediction, they further averaged the scores across the individual views.

The training dataset consists of ultrasound images from 2899 patients, with visual diagnoses for steatosis used as the training labels. For testing, the authors employed histopathology-proven diagnoses to validate their model. The AUC scores for classifying mild, moderate, and severe steatosis were 0.85, 0.91, and 0.93, respectively. When compared to FibroScan, the proposed solution either outperformed or matched the performance of FibroScan's Controlled Attenuation Parameter (CAP) scores.

In their latest work [94], Li et al. revisit a limitation from their earlier research [144], where they relied exclusively on histology-proven diagnoses as the evaluation ground truth. This strategy, however, may introduce selection bias since biopsies are recommended primarily for patients with suspicious conditions. To address this issue, they leveraged the correlation between liver steatosis and body weight to introduce a new cohort of patients who underwent significant weight changes. Their findings revealed a clear correlation between weight gain and AI predicted steatosis values, with a coefficient of determination (R^2) of 0.62. By using body weight, considering it a strong indicator of steatosis, the authors were able to validate their approach more broadly and demonstrate its applicability to the general population.

Many other works assessing liver lesions with a multiview paradigm exists in b-mode [130] and others with Contrast-Enhanced Ultrasound (CEUS) data [283, 282, 91, 72, 254]. Interested readers are encouraged to refer to these publications for further details.

Multiview Diagnosis : Kidney

Building on the work of [294] in classifying CAKUT in children, [274] proposes a multiview deep learning approach using Multiple Instance Learning (MIL). In this framework, multiple images from the kidney of the same patient are analyzed, with ground truth labeled as normal only if all images are normal, and as pathological if at least one image exhibits pathological signs. The authors first extract image features using a custom 3-layered CNN model and then construct a Graph Neural Network (GNN) by representing each image's features as a node. The graph edge weights are assigned based on the Euclidean distance between nodes (image features), creating an undirected graph optimized with the GNN framework [132]. An attention layer is then applied to the resulting node features to assign higher weights to relevant nodes. Additionally, a single-image classification loss is computed from features obtained from images with reliable individual ground truth labels, including all normal images and manually selected pathological ones. The entire network was trained on a dataset of 105 infants with CAKUT and 120 children with mild hydronephrosis as a control group. The model achieved a sensitivity of 85.82% and a specificity of 83.81%, which corresponding to an approximate AUC score of 0.85, using the single operating point formula.

In their subsequent work [273], the authors propose a simplified version of their approach, replacing the custom CNN and GNN with pretrained CNNs and predefined kidney views. Two VGG-16 models [224] were trained independently : one for sagittal view images and another for transverse view images. For diagnosis, a CAKUT prediction is made if at least one view is classified as pathological ; for a control diagnosis, both views must be classified as normal. This study used a dataset of 86 infants with CAKUT and 71 children with mild hydronephrosis as a control group. By applying this multiview approach, they significantly improved the AUC score from 0.815 to 0.961 compared to the single-view method.

Yin's progression in this work illustrates how leveraging robust, pretrained off-the-shelf CNN models and refining the problem with additional information (such as specific views) can dramatically enhance the accuracy of deep learning models while reducing engineering complexity.

2.1.1.4 Additional Modalities for Diagnosis

Multiview methods bring the innovation of processing images from multiple views through dedicated branches, combining view-specific features in later stages of the network. While this approach enhances deep learning models with complementary visual data, it remains limited to the information available in b-mode ultrasound images. To expand beyond this, researchers have integrated additional modalities like elastography, FibroScan, Doppler, and patient clinical data aiming to incorporate unique insights beyond standard ultrasound. For example, Doppler is particularly useful for assessing vascular characteristics in liver lesions to help identify malignancy, while FibroScan and elastography excels for evaluating fibrosis and cirrhosis. By incorporating these diverse modalities, models benefit from richer multimodal data, what can lead to more precise and comprehensive diagnoses.

The general process for multimodal diagnosis methods is similar to that of multiview diagnosis, with the key distinction being the use of distinct neural network architectures specific to the data type of each input branch. The steps are as follows (illustrated in Figure 2.8) :

1. **Feature Extraction from Modalities :** In this step, each input modality is processed by an independent branch, where a dedicated backbone network extracts relevant features to a common space. Depending on the input data type (image, text, 1D signal, etc.), preprocessing and network architectures may vary across branches.
2. **Modality Feature Fusion :** The extracted features from each branch are combined to create a multimodal feature array, providing a highly informative feature representation for the next stage.
3. **Final Diagnosis :** The fused features are then forwarded into a final network followed by a classification or regression head to generate the final diagnosis.

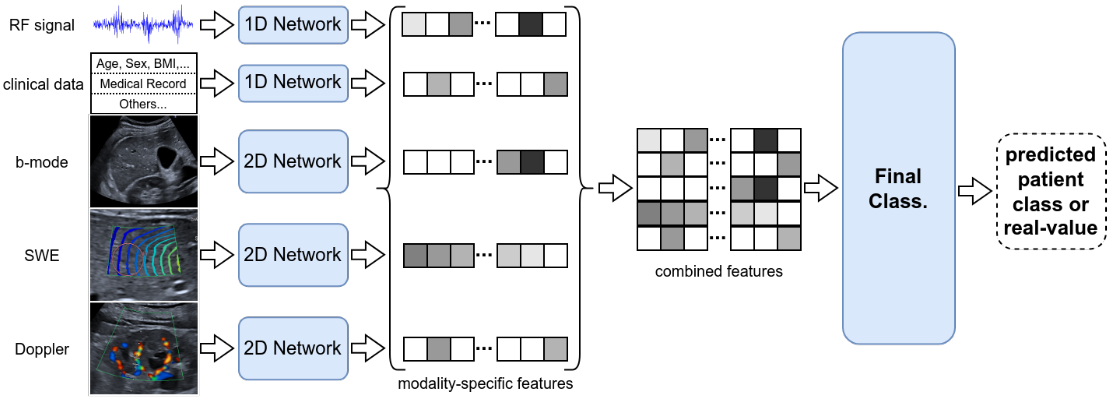


FIGURE 2.8 – Overview of multimodal diagnosis methods incorporating additional data types such as clinical records, elastography, and Doppler signals, processed through dedicated network branches to improve diagnostic accuracy.

Below, we present the most relevant works within these categories, organized by the specific pathologies being diagnosed.

Additional Modalities for Diagnosis : Liver

Starting by [90], the authors developed a one-dimensional convolutional neural network (1D CNN) to predict hepatic steatosis using ultrasound radiofrequency (RF) signals as input. They collected data from a cohort of 204 patients, comprising 140 diagnosed with NAFLD and 64 control participants without liver disease. The labels for training were derived from MRI-PDFF measurements.

To acquire the set of 1D RF signals from the liver, they first defined a ROI using b-mode ultrasound images as a reference. This ROI was intended to include as much as possible of the liver parenchyma below the liver capsule as possible while avoiding tissues outside the liver. They then acquired RF scan lines within this ROI ; each patient had 10 ultrasound frames, each containing 256 RF scan lines, resulting in a total of 2560 samples per patient. They trained two custom 1D CNN models : one for binary classification between pathological and normal liver tissue, and another for continuous-valued fat fraction estimation. The models achieved an AUC score of 0.98 for binary classification and a Pearson correlation coefficient of 0.85 for fat fraction estimation.

In [271], the authors used transfer learning to stage liver fibrosis utilizing two different ultrasound modalities : b-mode images and elastogram images from two-dimensional shear wave elastography. They studied a cohort of 466 patients, with histological examination of liver tissue serving as the ground truth for fibrosis staging. Images were acquired from the right liver lobe. On the elastogram images, they defined a ROI on the liver parenchyma used for stiffness measurement, avoiding lesions and the liver border. The same ROI was automatically applied to the corresponding b-mode images to ensure both modalities analyzed the exact same liver region.

They employed a pretrained Inception-V3 network [235], fine-tuning the higher-level layers to adapt to their specific task (independent models having b-mode and stiffness imaging as input for categorical fibrosis staging). To combine both modalities, they extracted 2048-dimensional feature vectors from each modality using the corresponding fine-tuned CNNs. These feature vectors were concatenated to form a 4096-dimensional feature vector, which was then input into three fully connected layers acting as the final classifier. Using this method, they achieved AUC scores consistently above 0.93 for predicting different stages of liver fibrosis.

In [60], the authors propose a machine learning framework for the classification of three liver pathologies : steatosis, inflammation, and fibrosis. The framework involves first delineating a ROI within the liver and then extracting quantitative ultrasound (QUS), elastography, and RF data from that. From the selected ROI, 11 features are extracted : Point Shear Wave Elasticity, Normalized Mean Intensity

Mean, Normalized Mean Intensity IQR, Scatterer Clustering Parameter Mean, Scatterer Clustering Parameter IQR, Coherent-to-Diffuse Signal Ratio Mean, Coherent-to-Diffuse Signal Ratio IQR, Diffuse-to-Total Signal Power Ratio Mean, Diffuse-to-Total Signal Power Ratio IQR, Total Attenuation Coefficient Slope, and Local Attenuation Coefficient Slope.

The extracted features were then used to train a Random Forest model, with histopathology-proven diagnoses serving as the ground truth. Using this approach, the authors achieved AUC scores of 0.90 for steatosis, 0.75 for inflammation, and 0.77 for fibrosis, significantly outperforming elastography alone in all classification tasks.

Following a similar approach, in [210], the authors evaluated different CNN architectures (1D, 2D, and 3D CNNs) to process raw RF ultrasound data in combination with B-mode images. The RF data allowed for the computation of additional ultrasound features, such as spectral and phase representations. Their findings revealed that models using RF data achieved significantly higher accuracy than the ones using b-mode images alone, with AUC scores of 0.994 and 0.938, respectively. The study included 31 patients, with ground-truth labels derived from magnetic resonance imaging proton-density fat fraction (MRI-PDFF). The deep learning models also surpassed the accuracy of radiologists' annotations when compared against MRI-PDFF values, achieving an accuracy of 0.989 versus 0.914.

Finally in [115], the authors predict real-valued liver fat fraction percentages using multiple US image modalities, with MRI PDFF serving as the reference fat percentage labels. Their network is designed as a multi-branch CNN, where each branch is dedicated to processing a different modality, as shown in Figure 2.9. The first branch handles b-mode ultrasound images, similar to previously discussed CNNs. The second branch processes tissue attenuation imaging (TAI), a pixel-by-pixel map of tissue attenuation properties [150, 171]. The third branch uses tissue scatter-distribution imaging (TSI), which reflects the arrangement and concentration of scatterers within the tissue [171, 99]. The results were evaluated on 173 patients, 126 of whom had steatosis. The model demonstrated strong performance, with a Pearson correlation of 0.86 ($p < 0.001$) when compared to MRI PDFF, and an AUC of 0.97 in distinguishing steatosis from healthy patients. However, the main limitations of the proposed method are the long acquisition times required for TAI and TSI images, as well as the high cost of MRI PDFF for dataset annotation.

Additional Modalities for Diagnosis : Kidney

In [221], the authors designed a method to predict Acute Kidney Injury (AKI) using a combination of ultrasound imaging and patient data (including demo-

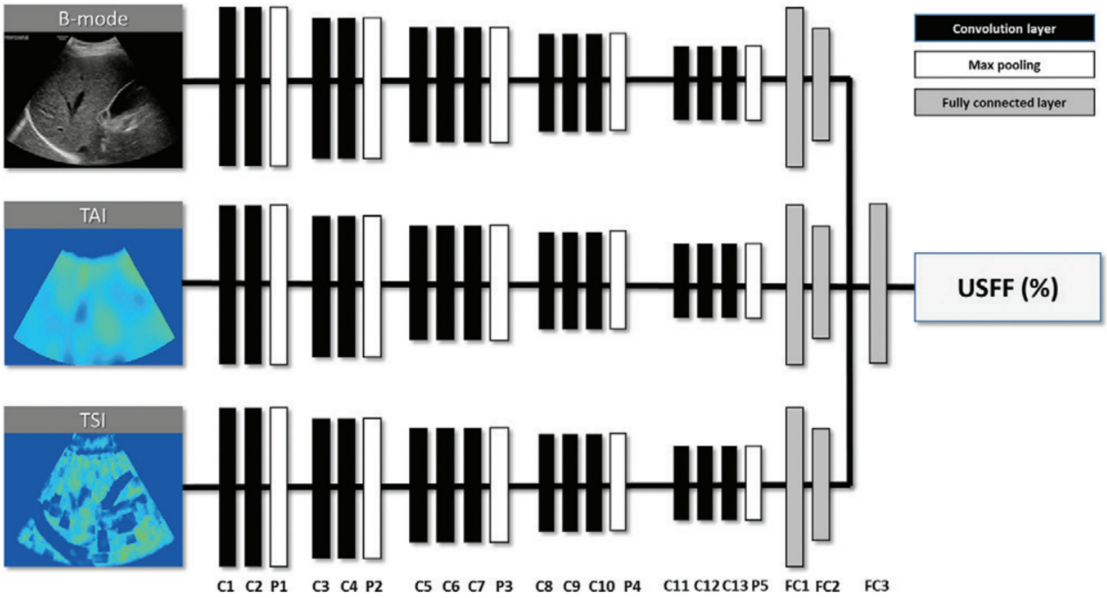

B-mode
TAI
TSI
Convolution layer
Max pooling
Fully connected layer
USFF (%)
C1 C2 P1 C3 C4 P2 C5 C6 C7 P3 C8 C9 C10 P4 C11 C12 C13 P5 FC1 FC2 FC3

Figure 2.9 – The schematic illustrates the development of a two-dimensional convolutional neural network algorithm for estimating hepatic fat fraction. Input data include B-mode images, tissue attenuation imaging (TAI) maps, and tissue scatter-distribution imaging (TSI) maps, all derived from radiofrequency data analysis. These three input datasets (B-mode image, TAI map, and TSI map) generate a single output : the deep learning–estimated US fat fraction (USFF). The network consists of convolutional layers (C), pooling layers (P), and fully connected layers (FC). Figure and description from the authors [115]

graphics, vital signs, and key laboratory results). The proposed hybrid model has two input branches : one processes numeric patient data, while the other processes b-mode US imaging data. The numeric data is first processed through 1D convolutions and then reshaped to match the dimensions required for a ResNet-34 model [96]. The image branch uses a ResNet-50 model [96]. The features from both branches are concatenated, processed through an additional CNN block, followed by two fully connected layers, and finally a softmax layer, yielding a binary output.

Given that their dataset contains both paired and unpaired data (numeric and image data from different patients), the model is trained with each branch separately when only unpaired data is available. This approach allowed them to expand their dataset, incorporating 612 cases with paired data and 2532 additional cases containing numeric data only. The proposed method achieved an AUC score of 0.95, marking a significant advancement in AKI diagnosis accuracy.

To integrate multiple sources of input information, the authors of [300] proposed a framework for diagnosing CKD, with a particular focus on kidney fibrosis. Their machine learning model combines features from several imaging modalities.

From Shear Wave Elastography (SWE), mean and median elasticity values were measured at multiple positions ; from B-mode ultrasound, kidney length, width, and thickness were measured ; from color Doppler flow imaging, blood flow parameters and resistance indices were obtained ; and finally, clinical variables such as age, sex, body mass index (BMI), and medical history were included. All features were concatenated, and a Random Forest algorithm was used for feature selection and dimensionality reduction [227]. An SVM model was then trained on data from 117 patients, with ground-truth labels obtained via kidney biopsy. The model achieved an overall AUC score of 0.83 for predicting kidney fibrosis, performing less effectively in early stages (AUC 0.64) compared to moderate to severe fibrosis (AUC 0.94).

Using an innovative approach, [86] utilizes cardiac ultrasound imaging to predict CKD, supporting the hypothesis that changes observable in four-chamber heart ultrasound images can be used for CKD staging. They begin by extracting local texture and contour features using steerable filters (similar to Gabor Filters [190]) at different orientations, which are then averaged to create a feature representation image. Next, within a predefined grid of non-overlapping blocks, they compute multiple entropy measures [78, 219, 198, 19] for each block, generating a local feature for each. To incorporate spatial information for the final classification, they apply supervised neighborhood preserving embedding (SNPE) [98, 85], resulting in a new low-dimensional feature representation. Finally, a classical SVM model is used to estimate CKD stages. The study was conducted on a cohort of 220 patients, including 120 with CKD, with ground-truth labels derived from eGFR values. This method achieved 97.27% accuracy in staging CKD (distinguishing between healthy and CKD stages 3, 4, and 5), demonstrating the effectiveness of echocardiographic images in diagnosing kidney diseases. However, despite the high accuracy, the method appears sensitive to hyperparameter tuning and dataset curation, which may limit its applicability in real-world scenarios where data variability is significant.

2.1.2 Image selection and quality control in ultrasound diagnosis

In previous sections, we discussed single-image machine learning methods that perform classification and regression tasks to achieve automatic diagnosis in ultrasound imaging. These approaches have demonstrated impressive performance, often reaching AUC values above 0.95 across all categories (ROI-Based Diagnosis, Whole Image Analysis, Multiview, and Multimodal Diagnosis).

However, the vast majority of these methods are trained and evaluated on manually curated ultrasound datasets. This curated data can vary in form : selected images from one or more pre-defined standard planes, images with minimal noise and artifacts, or images specifically cropped to focus on annotated bounding boxes or segmentation masks around organs and their internal structures.

To automate the image selection and quality control process, deep learning models can be developed to ensure that input images meet the necessary diagnostic standards. In this context, two key factors influencing the suitability of ultrasound images are identified : **recognizability of anatomical structures** and **levels of noise and artifacts**.

1. **recognizability of anatomical structures :** this factors refers to the clear presence of the organ and relevant anatomical structures for diagnosis. It can be assessed through classification tasks (e.g., standard plane classification) or localization tasks (e.g., object detection or semantic segmentation), which can also be used to define ROIs and measure organ dimensions.
2. **levels of noise and artifacts :** This factor addresses the presence of noise, shadows, and other ultrasound artifacts in the image. While the organ may still be identifiable in degraded images, excessive artifacts can compromise diagnostic accuracy.

To evaluate these factors, various deep learning strategies have been proposed. We have grouped these strategies into the following categories :

1. **Object Detection :** This approach involves detecting and localizing the structures of interest by predicting bounding boxes around them. Object detection can be used to define ROIs for cropping or further analysis and to measure organ dimensions, such as the length of the principal axis.
2. **Semantic Segmentation :** In this method, each pixel in the image is assigned a label indicating whether it belongs to the structure of interest. Semantic segmentation enables precise isolation of anatomical structures from the background and can be utilized in methods requiring detailed structural delineation. It also allows for the measurement of organ dimensions and morphological features.
3. **Standard Plane Recognition :** This strategy focuses on classifying images according to predefined standard planes that are considered optimal for diagnosis. Standard plane recognition can be formulated as a categorical classification task or as predicting continuous scores representing the degree of conformity to the standard plane criteria.
4. **Image Quality Assessment :** This involves evaluating and quantifying the level of noise and artifacts present in ultrasound images. Methods in this category may output a real-valued quality score, perform binary classification to indicate whether the image meets quality standards, or generate segmentation masks that highlight areas affected by noise or artifacts.

Specifically, segmentation and object detection techniques enable the isolation of the organ's ROI from the image, so simplifying the diagnostic task, as demonstrated in Section 2.1.1.1. Similarly, standard plane recognition has significant potential to reduce variance in whole image diagnosis (Section 2.1.1.2) and is essential for the automation of multiview diagnostic (Section 2.1.1.3).

In the following sections, we present a selection of methods for each category. Since object detection and semantic segmentation are not the primary focus of this thesis, we will highlight only the works most pertinent to our research. For readers interested in a more extensive examination of these methodologies, we provide references to additional literature.

2.1.2.1 Methods using Object Detection

Object detection aims to identify and localize specific structures within an image by outputting bounding boxes around them.In the context of this work, object detection networks take a 2D ultrasound image as input and output a set of bounding boxes, each labeled with a predicted class

Training these networks involves supervised learning on manually annotated images where each object's bounding box and class label are provided. The loss function typically combines two main terms : a localization term and a classification term. The localization term, often using L1 or IoU loss, evaluates the accuracy of the predicted bounding box coordinates. While the classification term uses cross-entropy loss, as defined in the previous section.

In a 2021 study by [243], the authors trained a RetinaNet model [153] with a ResNet-50 backbone [96] to detect and categorize liver focal lesions (LFFs) in ultrasound images. This multi-class detection and classification problem included the following lesion types : Hepatocellular Carcinoma (HCC), cysts, hemangiomas, focal fatty sparing (FFS), and focal fatty infiltration (FFI). They extracted a total of 65,510 images from a cohort of 4808 patients, including 22472 images with LFF's positions and diagnostic annotated by radiologis with support of other clinical data. The RetinaNet was initially pre-trained on the Microsoft Common Objects in Context (MS-COCO) [154] dataset and then fine-tuned on the ultrasound dataset, using 72% of the data for training.

The model achieved decent localization metrics, with a detection rate of 87% and an IoU of 0.788, and performed well in classification, with 95.4% accuracy, 83.9% sensitivity, and 97.1% specificity. The error analysis revealed that the model's sensitivity was notably lower for small lesions (50% for lesions under 2 cm) but improved significantly for larger lesions (92.3% for those over 3 cm). Sensitivity was also reduced in patients with cirrhosis (79.5% vs. 89.7% in non-cirrhotic patients).

In [53] published in 2022, the authors compared a mixed CNN/transformer-based object detector (DETR [36]) with Faster R-CNN [197] for the detection and categorization of LFFs. From a cohort of 1026 patients, they curated a dataset of 2551 B-mode ultrasound images with visible lesions. Experienced radiologists annotated each image with bounding boxes for the following structures : liver parenchyma (classified as either healthy or with LFFs), benign LFFs (cysts, angioma, focal nodular hyperplasia, or adenoma), and malignant LFFs (metastasis and HCC). During annotation, the radiologists had access to additional imaging modalities such as contrast-enhanced ultrasound, CT, and MRI to ensure accurate labeling. Examples of these bounding boxes are illustrated in Figure 2.10.

Both models (DETR and Faster R-CNN) were trained using standard data augmentation techniques at the image and bounding box levels. Evaluation was conducted in four categories : Lesion Detection within the Liver Parenchyma, Lesion Localization, Lesion Characterization (Benign vs. Malignant), and Subcharacterization into Specific Lesion Types. For lesion detection and localization, both networks performed similarly, achieving an accuracy range of 93-96%, comparable to that of expert radiologists. However, in the lesion characterization and subcharacterization tasks, DETR excelled with 81% and 76% accuracy, respectively, surpassing both Faster R-CNN (76% and 72%) and the expert radiologists (59-61% and 50-52%). For lesion localization, there was no statistically significant difference between the performance of the networks and radiologists, whereas in lesion characterization, DETR was significantly superior to the experts ($p < 0.05$), though not significantly better than Faster R-CNN ($p = 0.18$). Although not statistically significant, these findings suggest the potential of transformer-based models in enhancing ultrasound CAD.

In [124], the authors propose a combined detection and segmentation approach for automatically locating and measuring renal cysts in ultrasound images. The dataset consists of 2,664 B-mode ultrasound images from 1,444 patients, all featuring renal cysts measurable by radiologists. Annotations involve placing two landmarks (points) from which cyst length is calculated as the distance between them. The authors first fine-tuned a YOLOv5 model [119] using bounding boxes created from the cyst landmarks. These output masks were then used to crop cyst regions of interest (ROIs), which served as inputs for a segmentation network designed to pinpoint cyst landmarks. The segmentation model, UNet++ [298] with a DenseNet121 backbone [103], was trained using saliency maps of the landmark positions as ground truth. During inference, the network generated saliency maps, and the lowest saliency coordinates were identified as landmark points following post-processing adjustments. The system was evaluated against two experienced sonographers, achieving 85% precision (compared to 86% and 83%), 86% recall (compared to 87% and 84%), a position error of 3.22 mm (vs. 2.56 mm and 2.34

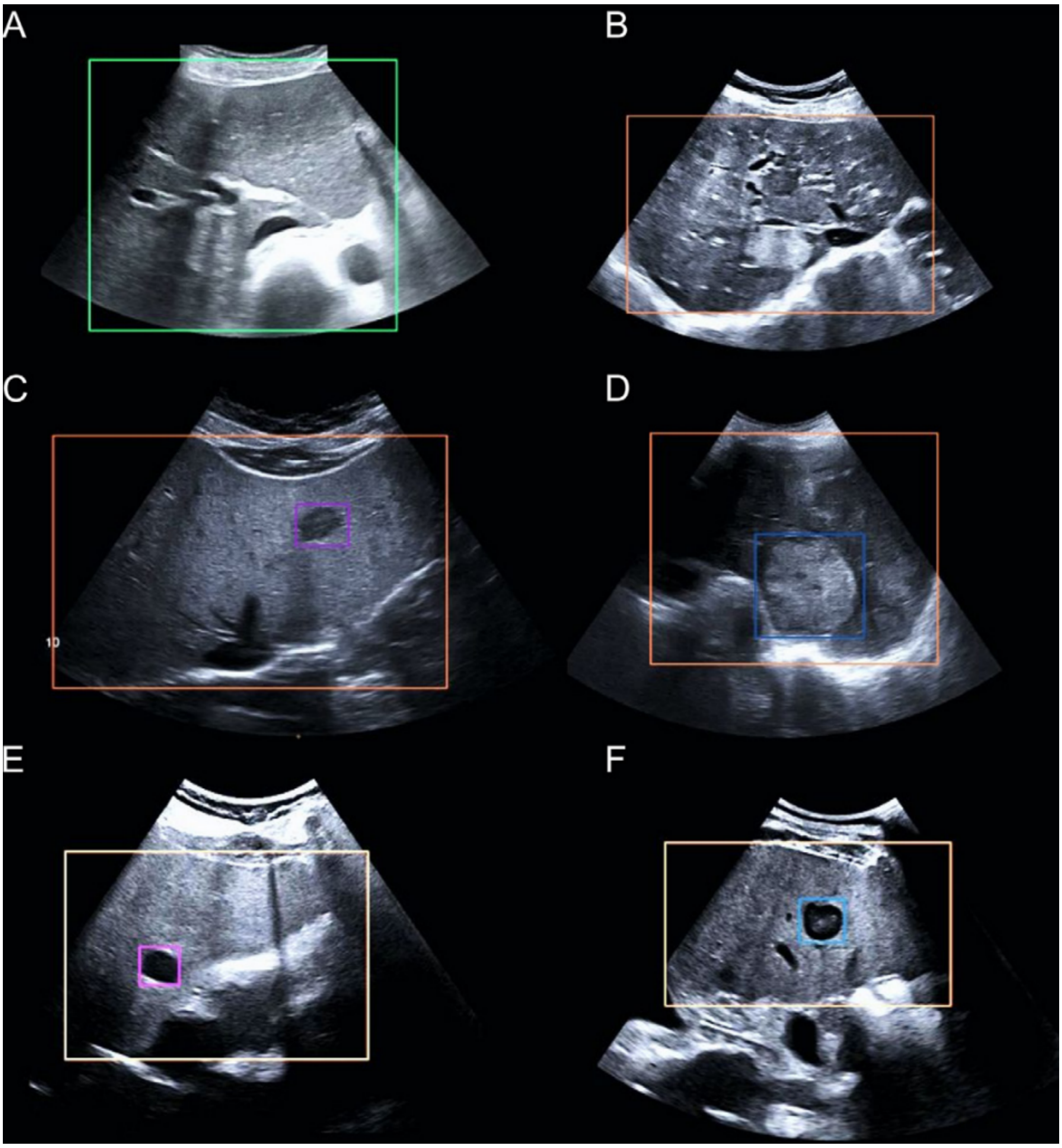


FIGURE 2.10 – (A) A liver without lesions (green box) and (B) a liver with lesions (orange box). (C) A benign lesion (focal nodular hyperplasia [small purple box]) and (D) a malignant lesion (hepatocellular carcinoma [small blue box]). (C, D) show that benign and malignant lesions exhibit differences in texture and size. (E) A benign lesion (cyst [purple box]) with a circular shape and dark pixel intensities, and (F) a malignant lesion (metastasis [blue box]) with similar characteristics. These images highlight the challenges of distinguishing malignant from benign lesions. Figure adapted from [53].

mm), and a diameter length error of 1.09 mm (vs. 1.21 mm and 0.95 mm). Overall, the system demonstrated competitive performance against human experts, highlighting its potential for clinical application.

2.1.2.2 Methods using Semantic Segmentation

Semantic segmentation aims to identify and localize specific structures within an image at the pixel level, assigning a categorical label to each pixel. In our context, neural networks performing semantic segmentation receive a 2D ultrasound image as input and output a categorical mask, where the value of each pixel in this mask indicates the predicted class (e.g., organ, structure).

The annotation effort required for training these networks is more intensive than for previous methods, as it requires precise pixel-level annotations in the form of segmentation masks. The choice of loss function can vary significantly but typically includes pixel-wise classification losses (e.g., Cross-Entropy Loss), shape similarity losses (e.g., Dice or IoU), or a combination of both.

Starting by the diagnosis of **kidney diseases**, to which the useful of semantic segmentation for CAD has already been demonstrated in the ROI-Based Section (2.1.1.1). It is also also supported in [37] back in 2016, where the authors demonstrated that accurate kidney segmentation in b-mode ultrasound images can be used for the diagnosis of kidney pathologies. In this early study, the authors extracted shape descriptors from manually segmented kidneys and trained an SVM model, achieving an impressive AUC score of 0.98. Although reliant on manual segmentation, this work provided early evidence that kidney segmentation could be effectively leveraged for accurate diagnostic purposes.

Also, many methods try to solve the segmentation problem using the approach of region descriptors and machine learning. In [218], the authors segment renal ultrasound images using K-means clustering and extract GLCM texture features, which are then used with a meta-heuristic SVM classifier to diagnose renal calculi with 98.8% accuracy.

A similar approach is used in [177], where the authors adopt a two-step approach : first, they classify kidney images into normal kidney, kidney stones, and kidney tumors, followed by segmentation of abnormal images. They extract 22 GLCM features [93] from the input images and select the most discriminative ones using the Crow Search Optimization Algorithm (CSOA) [12]. These features are then used to train a 3-layer dense neural network. For segmentation, they apply a multi-kernel k-means algorithm, which combines a quadratic loss with the linear loss used in classical k-means, applied solely to pixel intensity with manual identification of pathological clusters. The authors report a classification accuracy of 93.45% and a segmentation accuracy of 99.61%. Although these high values indicate strong performance, they may be due to extensive manual curation, optimal parameter tuning, and manual cluster selection, potentially limiting the method's scalability.

Today, end-to-end Deep Learning approaches are the preferred method for automatic segmentation of ultrasound images. Most of these networks used or are inspired by U-Net, which is well known model first introduced in 2015 [205]. These architectures capture meaningful latent image features by downsampling the input through an encoder block, which are then upsampled in a decoder block. Skip connections link corresponding encoder and decoder layers to help produce the final segmentation mask. The Figure 2.11 illustrate this process.

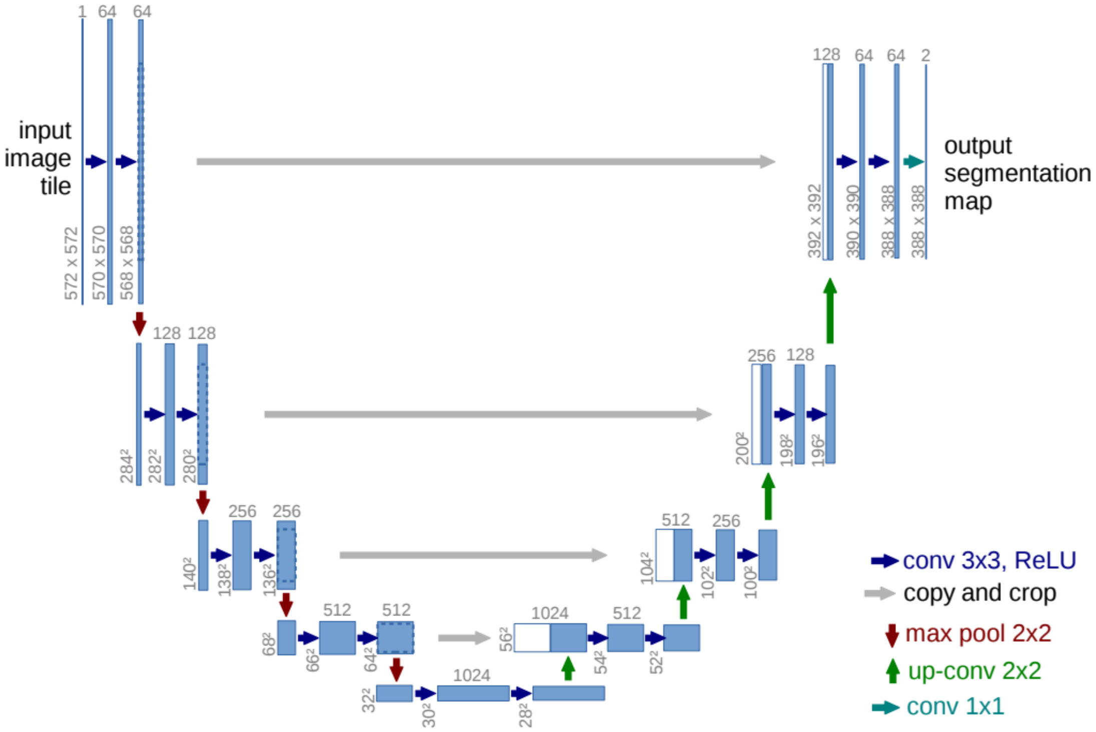


FIGURE 2.11 – U-Net architecture for image segmentation, featuring an encoder-decoder structure with skip connections. Figure extracted from [205].

In [13],using a standard U-Net model the authors generate kidney segmentation masks, which are then used to crop and isolate the kidney from the rest of the image. The cropped kidney image is then passed through a VGG-19 model [224] to extract discriminative features, which are subsequently used to train an XGBoost model [47] for CKD diagnosis. The segmentation network was trained on 500 images, while the diagnostic model was trained on 5122 images from 352 patients. With a segmentation IoU score of 91%, the model achieved a diagnostic accuracy of 89%.

Building on the U-Net model in [156], the authors employ an attention-based

U-Net [205] to segment the kidney for the diagnosis of hydronephrosis. The segmentation network identifies both the kidney and the dilated pelvicalyceal system containing fluid, allowing computation of the fluid-to-kidney area ratio. This ratio serves as a biomarker for hydronephrosis, achieving an approximate AUC score of 0.91 using the single operating point formula.

In [107], the authors also introduce an Attention U-Net segmentation network for detecting pleural effusion in ultrasound images. Pleural effusion, the abnormal buildup of fluid in the pleural cavity surrounding the lungs, often indicates underlying conditions that require prompt treatment. The dataset comprises 800 images with pleural effusion and 640 without, all manually annotated by two experienced radiologists. These images were used to train a fully supervised U-Net [205] enhanced with Attention Gates (AGs) [179]. The model achieved AUC scores between 0.95 and 1.0 for pleural effusion detection and Dice coefficients ranging from 0.83 to 0.90 for the segmentation task.

Finally in [43], the authors introduce a Multi-branch Aware Network (MBANet) for kidney segmentation, designed to process images across multiple resolution branches. This architecture shares several principles with DeepLabv3 [45], where the network process images at different scales ; the readers may refer to the original article for in-depth architectural details. A innovative image quality based staged pre-training strategy is used to initialize the network weights, by exploiting a dataset of 450 images manually annotated into classes by clarity and kidney integrity : ideal (T1), good (T2), and normal (T3). The network is then pre-trained progressively, starting training with the dataset T1, then using T1-trained model weights to initialize training on T2, and so forth, concluding with fine-tuning on the full dataset (T1+T2+T3). This approach enhances MBANet's IoU scores from 91.51% to 92.38%, outperforming all baseline methods.

While the staged pre-training strategy does not achieve statistical significance, it effectively addresses the challenges of variable image quality and visibility that arise when ultrasound images are acquired at scale by radiologists of differing experience levels. This method exemplifies a proactive solution to standardize results across variable data quality, suggesting potential improvements with automated quality classification models to further enhance efficiency and consistency in clinical applications.

Many studies also employ semantic liver segmentation, primarily focusing on localization to support subsequent diagnostic tasks. An example of this is the work from [200] explained in the previous section, as well as many ROI-based methods in Section 2.1.1.1.

Starting by [168] published in 2018, where the authors trained a model based on FCNN [164] for the segmentation of liver Blood vessels and liver focal lesions, in b-mode and contrast-enhanced ultrasound. Annotations were performed manually by expert radiologists, achieving a total of 350 images for the vessel segmentation task nd 152 images for the lesion segmentation task. Their method achieved 0.83 and 0.87 IoU scores over the vessel and lesion tasks, surpassing U-net which is the second best-performing model with 0.81 and 0.84 IoU.

Starting by [168], published in 2018, the authors trained a model based on Fully Convolutional Neural Networks (FCNN) [164] to segment liver blood vessels and liver focal lesions in b-mode and contrast-enhanced ultrasound images, respectively. Expert radiologists manually annotated the dataset, resulting in 350 images for vessel segmentation and 152 images for lesion segmentation. This dataset is split for training and evaluation, achieving IoU scores of 0.83 for vessel segmentation and 0.87 for lesion segmentation, outperforming U-Net, which achieved IoU scores of 0.81 and 0.84 for the respective tasks.

In 2023, the authors of [8] introduced a liver segmentation model called Dense-PSP-UNet, which combines a Dense-UNet backbone [279] with a Pyramid Scene Parsing (PSP) module [293]. The Dense-UNet architecture optimizes model size and inference speed, while the PSP module enhances performance by providing both local and global contextual features. Trained on a cohort of healthy patients, each contributing 300 frames annotated by three radiologists, the proposed model achieved a Dice coefficient of 0.913, outperforming the baseline U-Net [205], which achieved 0.889. Even more notably, the model achieved a real-time inference speed of 37 FPS, being suitable for integration in clinical applications.

In 2024, the authors of [265] introduced a lightweight liver segmentation network called Boundary-Aware Convolutional Attention Network (BACANet), built on a ResNet10t backbone [262] for feature extraction with explicit liver boundary supervision. The model includes several novel modules : the Selective Large Kernel Convolution Module (designed to capture boundary features), the Enhanced Attention Gate (an attention mechanism to reduce the semantic gap between the encoder and decoder), and the Multi-scale Dilated Convolutional Attention Module (which provides global context using multi-scale dilated convolutions in the encoder). Comparing their model with Dense-PSP-UNet on the same dataset, BACANet achieved a Dice score of 0.921, an improvement over the previous score of 0.913.

2.1.2.3 Standard Plane Recognition

Standard Plane Recognition involves classifying 2D ultrasound images into one or more predefined ultrasound planes (views) or identifying them as background. This task is typically performed using multiclass classification neural networks, trained in the same manner as outlined previously in Section2.1.1.

Similar to object detection and semantic segmentation models, these networks can assist in dataset curation and simplify the learning process by filtering or categorizing images. Below, we discuss some key studies in this area that are relevant to our research.

Starting by [212], where the authors develop a multiclass classification neural network to distinguish among abdominal organs (Kidney, Liver, Pancreas, Spleen, Urinary Bladder), a model that can aid in image quality assessment by confirming that the current image corresponds to the target organ for diagnosis. They train and compare several single-image classification neural networks [137, 224, 234, 235, 96], using a dataset of 1906 organ-specific images. Despite achieving an impressive AUC score above 0.99, the model was trained and evaluated on manually selected images containing only the organs of interest (no null class). Consequently, the model's performance on real b-mode ultrasound videos remains uncertain due to the likely presence of numerous of low quality frames, which could impact its reliability in practical settings.

Subsequently in [284], the authors proposed an attention-based neural network to automatically classify standard liver ultrasound planes. They hypothesize that Transformer-based networks may be particularly effective for this task, as identifying standard planes relies on global context rather than localized regions. Ground truth labels were manually annotated by a radiologist with over 10 years of experience. The list of standard planes is provided below and illustrated in Figure 2.12.

a - **LSFLS** : Left liver and stomach fundus longitudinal section

b - **LAALS** : Left liver–abdominal aorta longitudinal section

c - **SVCLS** : Subhepatic vena cava longitudinal section

d - **HTSPH** : High transverse section at the level of the second porta hepatis

e - **MTFPH** : Median transverse section at the level of the first porta hepatis

f - **LTSH** : Low transverse section at the level of the hepatopancreas

g - **HOTSP** : High oblique transverse section of the second porta hepatis

h - **MOTFP** : Median oblique transverse section of the first porta hepatis

i - **LOTGK** : Low oblique cross section at the level of gallbladder and kidney

j - **LAFH** : Long-axis view of the first hepatic portal vein

k - **LGLS** : Liver and gallbladder longitudinal section

l - **LKLS** : Liver and kidney longitudinal section

m - **67OLS** : Sixth and seventh intercostal oblique longitudinal section

To preprocess the images, the authors cropped each image to isolate the field of view (FOV) and remove extraneous content such as device model details. Their

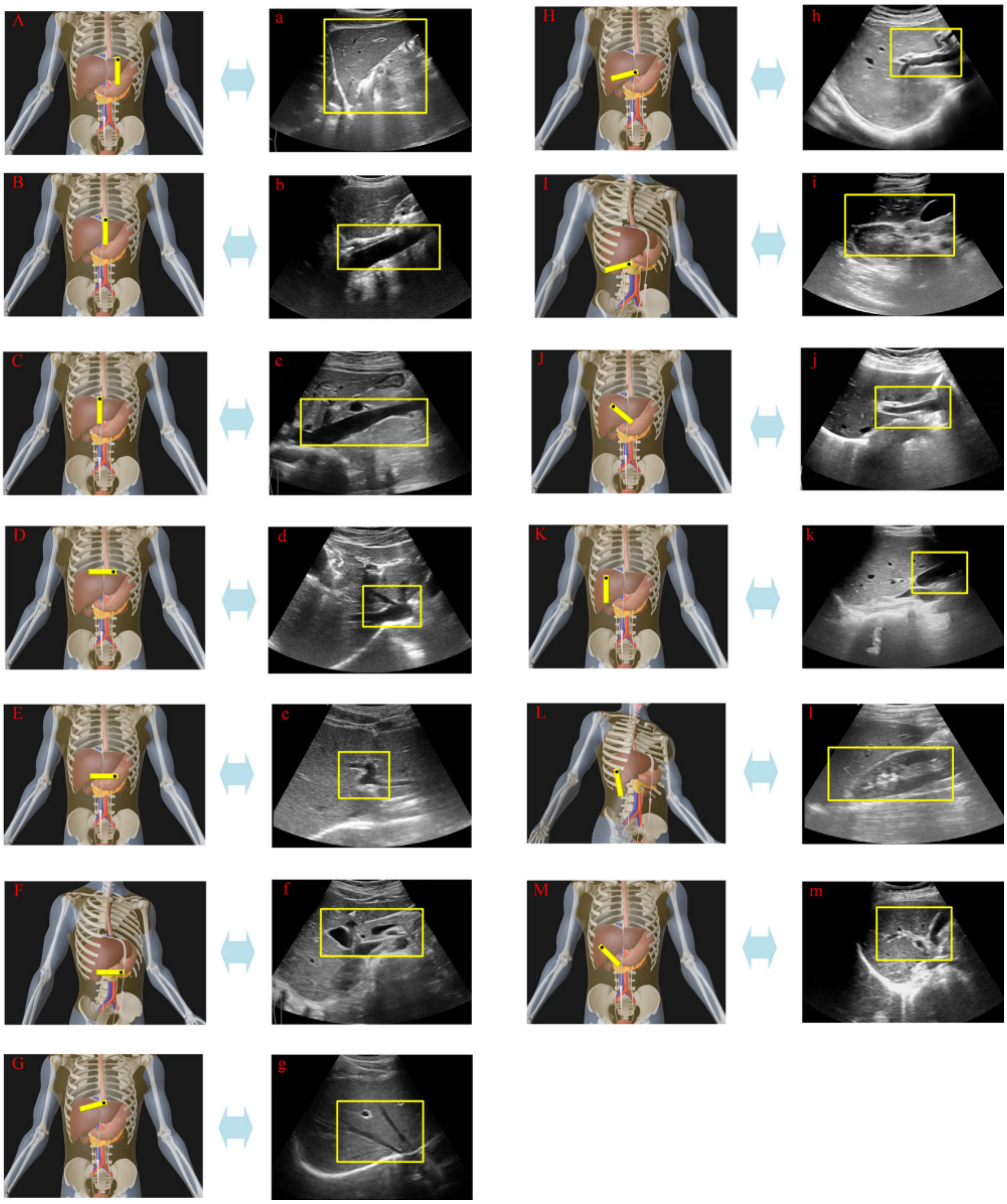


Figure 2.12 – Standard liver ultrasound planes illustrated in clinical sweep order (A-M, a-m), adapted from [284]. Identified sections include : (a) LSFLS, (b) LAALS, (c) SVCLS, (d) HTSPH, (e) MTFPH, (f) LTSH, (g) HOTSP, (h) MOTFP, (i) LOTGK, (j) LAFH, (k) LGLS, (l) LKLS, (m) 67OLS. Each plane highlights specific anatomical features, marked by yellow rectangles (not used by the neural network).

proposed Ultra-Attention network, similar to Vision Transformer (ViT) [64], introduces several innovations to optimize training. First, they add a class vector ([CLS]) to each Transformer block, rather than only in the first block, as is standard in ViT. Additionally, Focal Loss [153] is applied to address class imbalance, with the use of customized layer freezing, transfer learning techniques, and a hyperbolic tangent activation before the softmax layer.

The Ultra-Attention model achieved a classification accuracy of 93.2%, outperforming other networks in their comparison, including AlexNet [137], GoogLeNet [234], ResNet variants [96], MobileNet-v2 [211], DenseNet-121 [103]. It also outperformed DeepCNN [266], another work performing liver standard plane classification based on the VGG-16 network [224]. While the proposed architecture demonstrated superior performance, it is highly similar to ViT [64], which was not compared with.

2.1.2.4 Image Quality Assessment

Ultrasound acquisitions are highly operator-dependent, with variables like probe position, angle, contact pressure, and imaging settings greatly impacting image quality. As a result, images acquired from the same patient and probe can vary significantly in terms of organ visibility, the presence of visual artifacts, and image noise.

To address this issue, many studies have proposed the development of automatic image quality assessment methods. These methods typically estimate a real-valued score from 2D images, reflecting the overall quality or organ visibility (absence of noise and artifacts) of the image. Some approaches also frame this as a categorical classification problem, where quality intervals are represented by classes.

Classically the methods of ultrasound quality estimation can be grouped in two main groups : **Full Reference Image Quality Assesment (FR-IQA)** and **No-Reference/Blind Image Quality Assesment (NR-IQA)**.

1. **Full-Reference Image Quality Assessment (FR-IQA) :** These methods use a set of high-quality reference images for comparison. Image quality is assessed by comparing query images against these high-quality references.
2. **No-Reference/Blind Image Quality Assessment (NR-IQA) :** These methods do not rely on reference images for quality assessment. Instead, they typically apply custom model designs to evaluate quality based on features extracted from the image itself.

We also consider methods that use human-annotated quality labels for supervised training as part of the FR-IQA group, where annotations serve as the quality references. Complementary, unsupervised approaches for quality assessment are included within the NR-IQA group.

Full Reference Image Quality Assesment (FR-IQA)

In [226] the authors first propose an IQA network to estimate the quality of breast ultrasound images. In their work they train a fully-supervised BCNN (Bilinear CNN) [155] which outputs real-valued quality score. They train a fully supervised Bilinear CNN (BCNN) [155] to output a real-valued quality score. Training labels are derived from visual assessments by multiple radiologists, with each image assigned a categorical score ranging from 1 (low quality) to 5 (high quality). These scores are then averaged across radiologists to produce the real-valued quality score used as the training target. The trained network achieved 0.842 correlation with the groubd-truth labels.

In a subsequent publication [259], the authors propose a global-local IQA approach, where the quality of healthy images is assessed using their previous global IQA network from [226], while the IQA score for breast images containing lesions is estimated by a novel branch. This new branch first applies a U-Net [205] network to obtain the lesion ROI, from which local lesion features are extracted. These features are then used to predict a local IQA score, which serves as the IQA for pathological images. Figure 2.13 illustrate their method. This dedicated branch improved the performance for pathological cases, raising the Pearson correlation from 0.6606 to 0.8412.

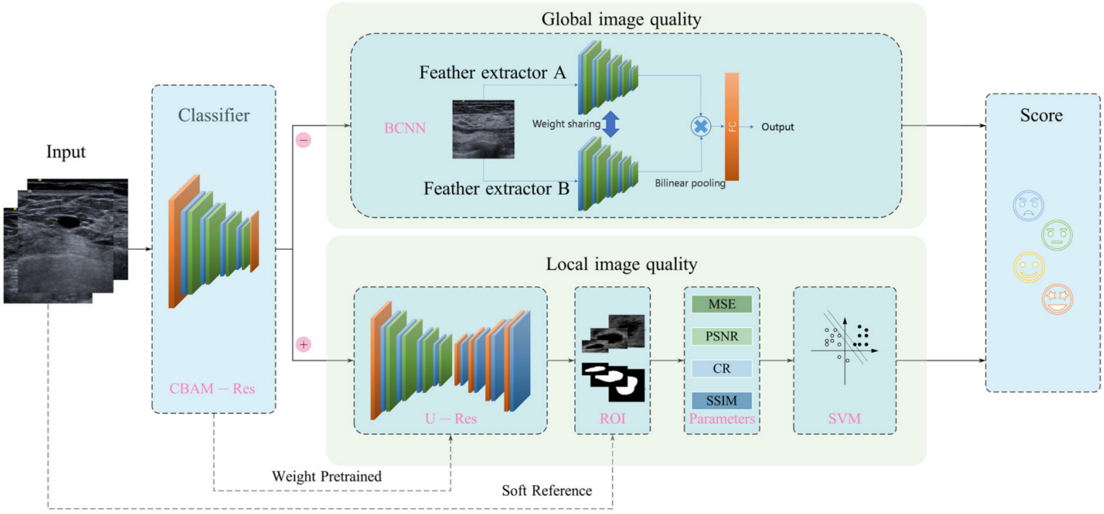


FIGURE 2.13 – Global–local integrated breast ultrasound IQA framework from [259].

In [31], the authors developed a prostate IQA model for ultrasound. In this work, they created a one-class regressor using DenseNet [103] with a Gaussian Process [125] at the output, trained exclusively on images labeled as high quality by experienced radiologists. During inference, the Gaussian Process indicates whether the query image falls within the distribution of high-quality images, achieving 94% accuracy compared to human experts.

Several other notable studies using supervised data for IQA in ultrasound exist in the literature [288, 46, 89]. In [288], the authors trained a supervised regression CNN to predict IQA scores. Conversely, [46] utilized the ResNet model [96] to train a multiclass classification network, with each class representing a quality score. Lastly, [89] addressed fetal plane IQA by training a regression network on pseudo-labels, where these labels are computed based on the distance to manually annotated high-quality frames, facilitating the development of quality assessment for fetal imaging.

No-Reference Blind Quality Assessment (NR-IQA)

Traditionally, these methods evaluate the level of noise in an input image through feature extract . For instance, a ROI-based approach cited in this work ([229]) employed the Perceptual Image Quality Evaluator (PIQUE) metric [253], which uses predefined models to assess noise, blurring, and distortions. Other more commonly used metrics include the Natural Image Quality Evaluator (NIQE) [169] and the Blind Image Quality Evaluation System (BIQES) [209]. NIQE evaluates image quality by calculating the distance between features extracted from the query image and those from distortion-free images, while BIQES measures dissimilarities between the query image and its low-resolution versions. A 2016 study on corrupted ultrasound images [181] suggests that among these methods, NIQE is the most suitable for assessing medical data.

Another approach for measuring image quality without references is to perform image denoising or artifact removal [129, 30, 126, 109], and them measuring the dissimilarities between the original image and the enhanced one. In [129] the authors use a non-local means (NLM) noise-reduction approach [23] to reduce image noise, which basically average correlated pixels locally.

In [30, 126, 109], the authors use Generative Adversarial Networks (GANs) to reduce image noise. These methods involve a Generator Network that enhances the quality of noisy input images, aiming to deceive a Discriminator Network into classifying the generated images as realistic high-quality . A key challenge in these approaches lies in obtaining paired datasets of the same image in both low and high quality—for effective training. In [30] the authors utilize the Pix2Pix network [113] and generate high-quality images for training by denoising low-quality images using the Weighted Nuclear Norm Minimization (WNNM) algorithm [84]. While this approach replicates the performance of WNNM, the main improvement lies in the real-time inference, otherwise impossible with WNNM.

To address the lack of paired data, some authors use CycleGAN networks [299], which are trained on unpaired datasets using a cycle-consistency loss. This loss ensures that transforming a low-quality image into a high-quality one and then back to low-quality, enabling training without paired datasets. In [126] the authors built low-quality and high-qualiyt dataset by controlling acquisition parameters and filters used, which are used to train a CycleGan Network that effectively improve image quality. Similarly, in [109], the authors proposed using a CycleGAN for acoustic shadowing removal, with unpaired datasets obtained through manual selection of images with and without shadowing. Results from this study are illustrated in Figure 2.14.

While GAN-based methods for improving image quality carry the risk of compromising diagnostic reliability by introducing or masking medical findings, they can still be reliably used for image quality assessment by comparing the input and the generated high-quality image.

2.1.3 Annotation Subjectivity

All the works discussed previously rely on ground-truth data, which can be obtained from various sources. While histopathology analysis and clinical measurements such as eGFR represent the gold standard for diagnosing many pathologies, they are not always readily available. This limitation can hinder the development of reliable CAD systems.

In such cases, researchers often depend on manual visual annotations provided by experienced radiologists. Despite adherence to clinical guidelines, annotations are inherently subjective, varying with radiologist expertise and sensitivity. Radiologists often vary in sensitivity to certain pathologies, complicating the standardization of annotation protocols without extensive cross-validation. This issue is particularly pronounced in ultrasound examinations, where image quality is highly operator-dependent, and identifying organs and internal structures can be challenging. For example, this variability contributes to a 20% underestimation rate for steatosis cases during ultrasound screenings [127, 142], a bias that is perpetuated to deep learning models trained on these annotations.

Some studies have addressed the issue of annotation subjectivity in ultrasound [143, 34, 114, 244]. A common strategy among these works [34, 114, 244] involves a two-step approach. The first step detects noisy label candidates, often through out-of-distribution analysis. The second step addresses these noisy labels,

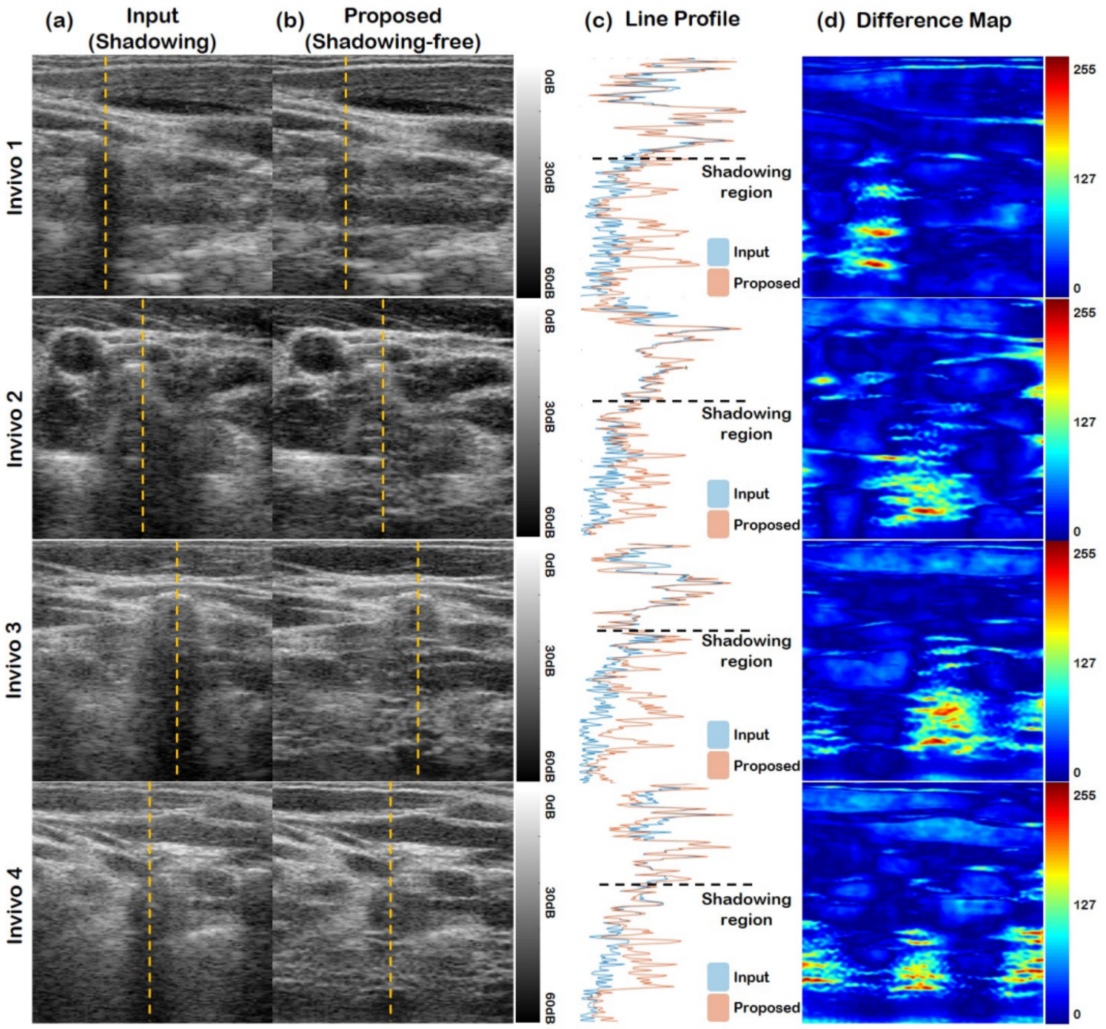


FIGURE 2.14 – Experimental results : (a) input image that has shadowing artifact, (b) result of the proposed method, (c) line profile of shadowing line, which is marked as a yellow line on the input and proposed image, (d) difference map between input and output. All B-mode images are visualized at 60 dB dynamic range and the difference map is shown in pseudo colors normalized from 0 to 255. Figure and description extracted from [109].

either by correcting them or by employing robust loss functions and architecture modifications to mitigate their impact.

Adopting a different approach, in [143] the authors propose a steatosis classification neural network featuring a dedicated annotator embedding branch, called Auto-Decoded Deep Latent Embedding (ADDLE). This ADDLE block takes annotator identifiers as input and learns an annotator-specific latent space representation. This representation is then integrated into intermediate layers of the ResNet-18 [96] model, acting as a learned prompt to enhance ultrasound image classification. A key advantage of this approach is that it does not require multiple annotations for the same image, making it suitable for mixing datasets from diverse centers and annotators. The model was trained with data from 65 different annotators, with the ability to simulate or average the embeddings during inference. The inclusion of the ADDLE block improved steatosis classification accuracy by 10.5%. An overview of the proposed framework is presented in Figure 2.15.

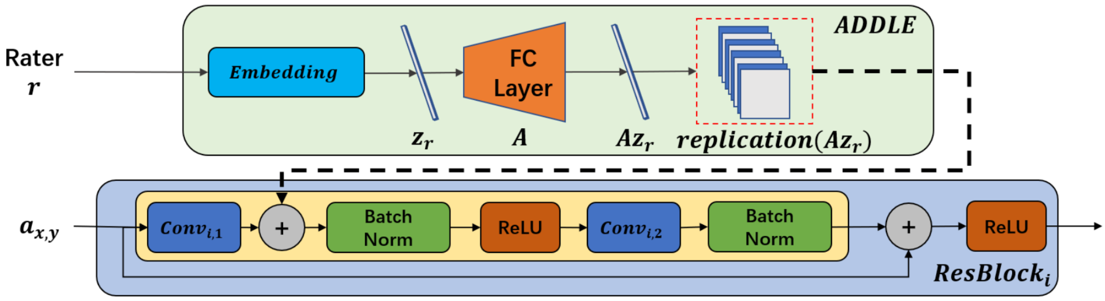


FIGURE 2.15 – Overview of the proposed ADDLE-based framework for steatosis classification, which incorporates annotator-specific latent embeddings into the ResNet-18 architecture to improve accuracy. Figure extracted from [143].

2.2 Video-Based Ultrasound Computer-Assisted Diagnosis

In the previous section, we extensively discussed single-image-based methods. While these approaches have demonstrated strong performance in recent years, their results are often reliant on curated datasets or controlled conditions. Furthermore, single-image diagnosis typically relies on only one or a few ultrasound planes, which may not fully capture the comprehensive nature of a complete ultrasound screening exam.

In this section, we introduce CAD models that utilize video-based neural networks. These models process spatial-temporal volumes (a sequence of consecutive

2D ultrasound frames) and generate a diagnosis, either as a real-valued score (video regression neural networks) or a categorical class (video classification neural networks).

These models also provide significant advantages in terms of label reliability and the effort required to annotate datasets. Since annotators have access to spatial-temporal context during the annotation process, they are likely to feel more confident in their decisions [65]. Additionally, the workload is reduced as each video requires only a single label, rather than annotating individual frames.

Unlike the more established single-image methods, which leverage mature and widely adopted off-the-shelf architectures like ResNet [96], VGG [224], or Inception [112], video-based deep learning lacks a clear consensus on optimal architectures for ultrasound analysis. These architectures differ significantly not only in layer parameters and design choices but also in how they process spatial-temporal information. Frames can be handled independently, as a 3D volume, or through strategies designed to capture temporal dynamics, such as motion tracking.

Based on our research, we categorize video classification methods according to how their neural network architectures process spatial-temporal frames and how temporal information is modeled. The following groups emerged from our analysis :

1. **Independent Singe-Image Network with Late Fusion** : These methods rely on single-image networks (as discussed in previous chapters) trained independently on individual frames. Video-level diagnosis is obtained by aggregating the outputs (features or predictions) from all frames using techniques like averaging, max pooling, or more advanced fusion strategies.
2. **Single-Image Feature Extraction with Temporal Modeling :** These methods process frames individually using a dedicated neural network (typically a CNN) to extract frame-level features, which are then passed to a temporal modeling module for video-level diagnosis. The system is trained end-to-end. Temporal modeling approaches include :
   a) **Recurrent Neural Networks (RNN) :** Frame features are input to recurrent layers (e.g., LSTM) that model temporal dependencies by combining the current frame with video history.
   b) **Attention Mechanisms :** Uses attention layers to focus on key frames or regions within the sequence that are most relevant for diagnosis.
   c) **Self-Attention (Transformer Blocks) :** Encodes each frame feature as a token and applies Transformer blocks to capture relationships between frames for a comprehensive diagnosis.

   d) **Graph Neural Networks (GNN) :** Treats each frame as a node in a graph, optimizing node and edge interactions to understand complementary information across frames.

3. **Two-Stream Networks :** These architectures use two parallel streams to process spatial and temporal information separately. One stream handles RGB frames, while the other processes motion (e.g., optical flow).
4. **3D Neural Networks :** These networks treat videos as 3D volumes, processing spatial and temporal dimensions simultaneously using 3D convolutions or specialized Transformer architectures.

Another important aspect is the incorporation of external information during training and inference. Standard video classification models are typically trained end-to-end, relying solely on video-level annotations or annotations from a minimal number of frames used in the loss function. Conversely, video classification models can also be designed to integrate guidance from external agents, such as networks for image selection and quality control described in Section 2.1.2. These approaches are categorized as **End-to-End Video-Based Diagnosis Models** and **Guided Video-Based Diagnosis Models**, respectively.

Given the limited number of studies in this category focused on diagnosing abdominal pathologies, we also include ultrasound-based research addressing pathologies in other regions of the body, such as cardiac and lung diseases. Finally, we do not explicitly describe models utilizing the Independent Single-Image Network with Late Fusion approach, as it is not a true video classification model and can be derived by adapting several works already discussed in the previous section.

2.2.1 End-to-End Video-Based Diagnosis Models

2.2.1.1 Spatiotemporal Processing

Recurrent Networks

A widely established approach for modeling temporal dependencies in sequences is the use of Recurrent Neural Networks (RNNs). These networks maintain and update an internal memory state, allowing them to influence how new input information is processed, enabling the generation of time-dependent predictions. Among these, Long Short-Term Memory (LSTM) [100] networks are the most well-known, which use capturing long-range dependencies to improve training stability.

features extracted from individual frames by a Frame Encoder serve as inputs to LSTM blocks. These blocks are specifically designed to retain and update an internal state that captures relevant information from previous frames, enabling time-aware predictions for each step in the sequence. To generate a final video-level diagnosis, aggregation techniques such as averaging, max pooling, or fully connected layers are often applied. This process is illustrated in Figure 2.16.

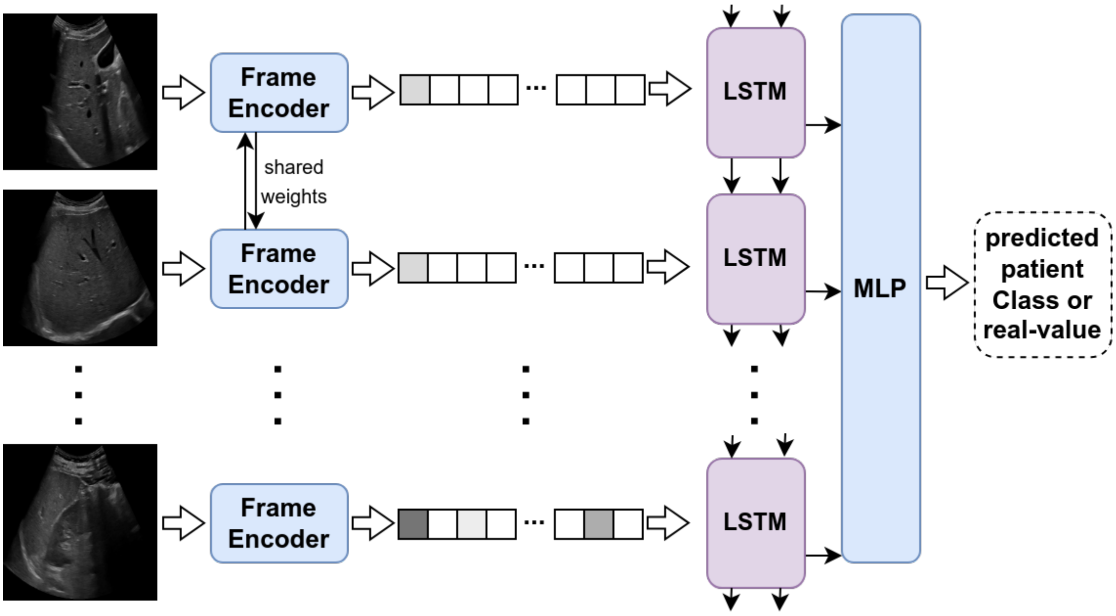


FIGURE 2.16 – Overview of LSTM-based video processing, where frame features are extracted by a Frame Encoder and processed through LSTM blocks for time-aware predictions.

The LSTM blocks are designed to maintain an internal state, denoted as C_t, which is used in combination with the input x_t to produced an output h_t. The architectures choices can vary significantly from one study to another, but overall architecture is presented in the Figure 2.17.

It contains a forget gate layer, which analyzes x_t and h_{t-1} to determine which parts of the previous state C_{t-1} should be retained and which should be discarded for the current state.

$$f_t = \sigma\left(\mathbf{W}_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2.7}$$

Similarly, the input gate layer decides which new information from the input x_t (Equation 2.2.1.1), while a **tanh** layer generates new candidate values to integrate into the current state (Equation 2.2.1.1).

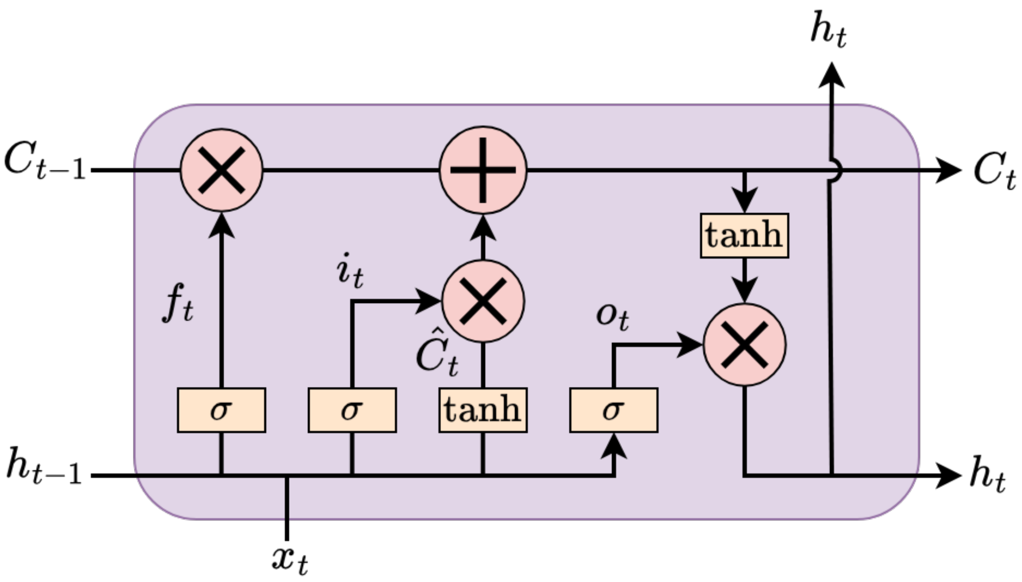


FIGURE 2.17 – Detailed architecture of an LSTM block.

$$i_t = \sigma\left(\mathbf{W}_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2.8}$$

$$\hat{C}_t = \tanh\left(\mathbf{W}_C \cdot [h_{t-1}, x_t] + b_C\right) \tag{2.9}$$

The current state is created taking in account the previous state C_{t-1} and the current candidates $\hat{C}_t$.

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \tag{2.10}$$

Finally, the output of the current stage, h_t, is determined by a layer that computes o_t to filtered by the current state C_t, as defined in Equations 2.2.1.1 and 2.2.1.1.

$$o_t = \sigma\left(\mathbf{W}_o \cdot [h_{t-1}, x_t] + b_o\right) \tag{2.11}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{2.12}$$

The learnable weights in the LSTM include $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_C$, and $\mathbf{W}_o$. The video-level output is computed from the sequence $\{h_0, h_1, \cdots, h_n\}$, as previously discussed. The entire system (Frame Encoder, LSTM blocks and aggregation layer) is trained in an end-to-end manner.

Some works have used LSTM models in order to model temporal dependencing in the automatic diagnosis of ultrasound video data for lung pathologies [15, 17, 220, 57], CEUS lesions [214, 296], fetal pathology assessment [192] and heart diseases [110].

In [57], the authors proposed a CNN+LSTM architecture to predict the severity of COVID-19 across four severity scores ranging from 0 to 3. They utilized a DenseNet [103], pre-trained as an autoencoder for denoising, in combination with a network employing Separable Convolutions [49] as the Frame Encoder, followed by LSTM blocks. The approach in other works is similar, with [17] using the VGG-19 model [224] as the Frame Encoder, [15] employing Inception V3 [235], and [220] also utilizing VGG-19 [224]. In all these studies, the output of the final LSTM unit is passed through a softmax layer for video classification.

In [296], the authors propose a similar approach, utilizing ResNet-18 [96] as the Frame Encoder. Rather than processing the entire video input, they sample frames at equal time intervals, filtering out those with low quality due to respiratory motion. The features extracted by the LSTM block are then used to predict the malignancy of liver lesions in CEUS.

Attention-based Networks

Before the development of Self-Attention and Transformer blocks introduced in *Attention is All You Need* [251], attention mechanisms were already applied to the classification of video data. Unlike the Self-Attention mechanism, which computes attention weights for all possible combinations of input frame features, the classical attention mechanism computes attention weights using a single query vector and the input features.

In this context, given an input video containing N frames, from which frame features are extracted independently from each other using a Frame Encoder $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$. So, a query vector $\mathbf{q}$ can be either be learned ($\mathbf{W}_q$) or provided by a custom method.

$$\mathbf{q} = \alpha\left(\mathbf{W}_q \cdot \mathbf{x} + \mathbf{b}_q\right) \tag{2.13}$$

where α is a non-linear function. This query vector is then used to compute attention scores for each frame feature, one way to do so is with the dot product :

$$e_t = \mathbf{q} \cdot \mathbf{x}_t \tag{2.14}$$

These scores are normalized using the softmax function and used to combine features from all frames to produce a video-level representation $\mathbf{x}_{\text{video}}$, which can then be used for downstream tasks, such as video classification or action recognition.

$$\mathbf{x}_{\text{video}} = \sum_{t}^{N} e_t \mathbf{x}_t \tag{2.15}$$

This approach is applied to ultrasound imaging for detecting Atrial Septal Defects in children, as described in [160]. The proposed model utilizes a ResNet18 [97] as a frame feature extractor. The extracted features are then processed through two separate branches : one employs the previously described attention layer, while the other applies 3D convolutions. The outputs from both branches are concatenated to generate the final video-based diagnosis.

Another approach is introduced in [230], where single-image annotated data is utilized to assist in training attention weights for ultrasound video classification models. The authors use a ResNet-50 [96] as a frame encoder to extract features from video frames, which are subsequently passed through an aggregation attention layer. The same frame encoder is also employed to extract features from a still image dataset, enabling the computation of class-wise feature centers. These centers are integrated into a coherence loss to guide the attention module, ensuring the assignment of reliable attention weights to individual frames.

Transformer-Based Networks

With the remarkable success of Transformer neural networks and self-attention mechanisms, initially introduced for text data by [251], numerous researchers have proposed adaptations to deploy it for video data.

In this context, video frames can be compared to words in a sentence. With effective tokenization of video frames, classical Transformer blocks can be directly applied to video analysis without structural alterations. However, the high dimensionality of video frames necessitates additional preprocessing steps to make this approach computationally feasible, what may limit the applications due to memory limits.

The common approach to address this challenge is to utilize a dedicated CNN to extract features from each frame, which serve as input tokens for the Transformer block [199, 4]. These tokens are augmented with positional embeddings and then passed through a standard Transformer block with self-attention operations. To generate the final diagnostic prediction, either the features corresponding to the [CLS] token (noted as 0 in Figure 2.18) or the aggregated output features are passed through linear layers. Figure 2.18 illustrates this process.

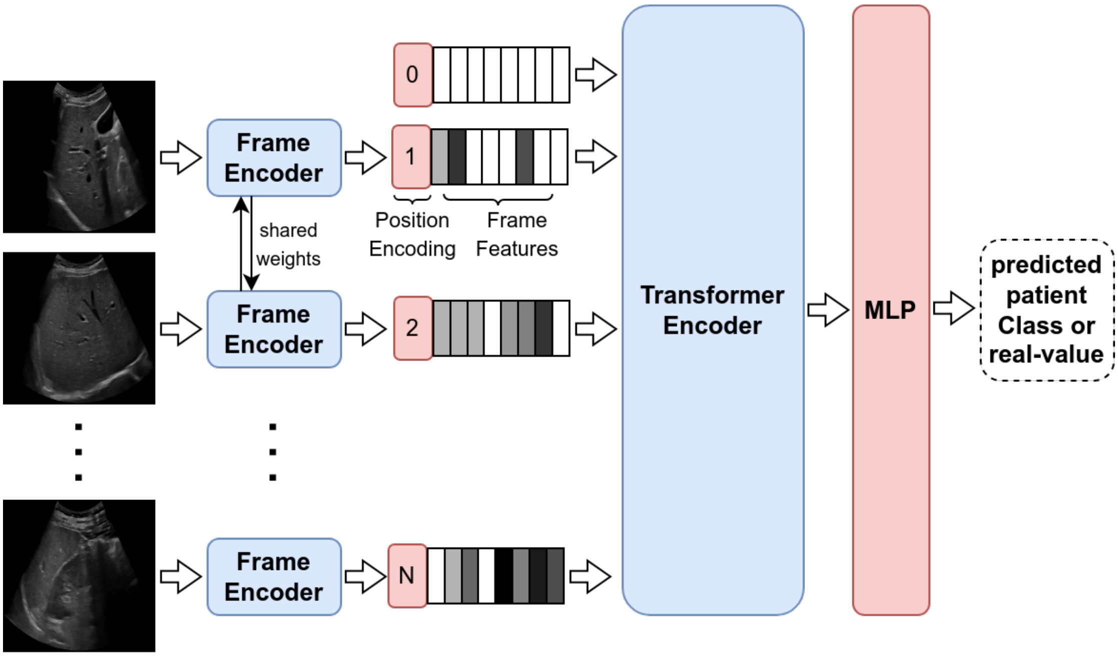


Figure 2.18 – A typical pipeline where CNN-extracted frame features, combined with positional embeddings, are processed by a Transformer block to produce diagnostic predictions.

In the UVT model [199], the authors utilized the encoder portion of a ResNet autoencoder [96] to extract frame-level features, which were then fed into a BERT encoder [61]. This architecture effectively predicted EF values while also identifying keyframes of critical importance. Similarly, [4] employed a ResNet-18 [96] to extract frame features, which were subsequently passed through a custom self-attention block designed to assign higher weights to regions around pre-determined reference frames.

Despite their successes, a notable limitation of these methods is the absence of a self-supervised pre-training step, a technique that has significantly advanced language models by enabling them to adapt to new tasks with minimal labeled data.

Graph Neural Network

Graph Neural Networks (GNNs) are designed to capture temporal dependencies between frames in a video by representing them as a graph. In this structure, frames are modeled as graph nodes, while edges represent the relationships between these frames.

In GNNs, after extracting frame features using the neural network C, each frame I_t in a video V with N frames is represented as $z_t = C(I_t)$. Thus, the entire video V can be described as a set of features $Z = \{z_1, z_2, \cdots, z_N\}$, where $z_i \in \mathbb{R}^d$, and d denotes the dimensionality of the extracted features

GNNs model video analysis as a graph $G = (V, E)$, where each vertex v_i corresponds to a frame I_i of the input video, and each edg $e_{i,j}$ represents the relationship between frame I_i and I_j. The frame embeddings (z_i) are used as node features of $H = \{h_1, h_2, \cdots, h_N\}$, typically represented as $H \in \mathbb{R}^{N \times d}$.

The primary objective of GNNs is to transform the initial node feature representation, $H^0 = \{h_1^0, h_2^0, \cdots, h_N^0\} = Z$ into a more informative one tailored for the specific video analysis task. At each step, the node features $H^t = \{h_1^t, h_2^t, \cdots, h_N^t\}$ are updated to $H^{t+1} = \left\{h_1^{t+1}, h_2^{t+1}, \cdots, h_N^{t+1}\right\}$, as illustrated in Figure 2.19.

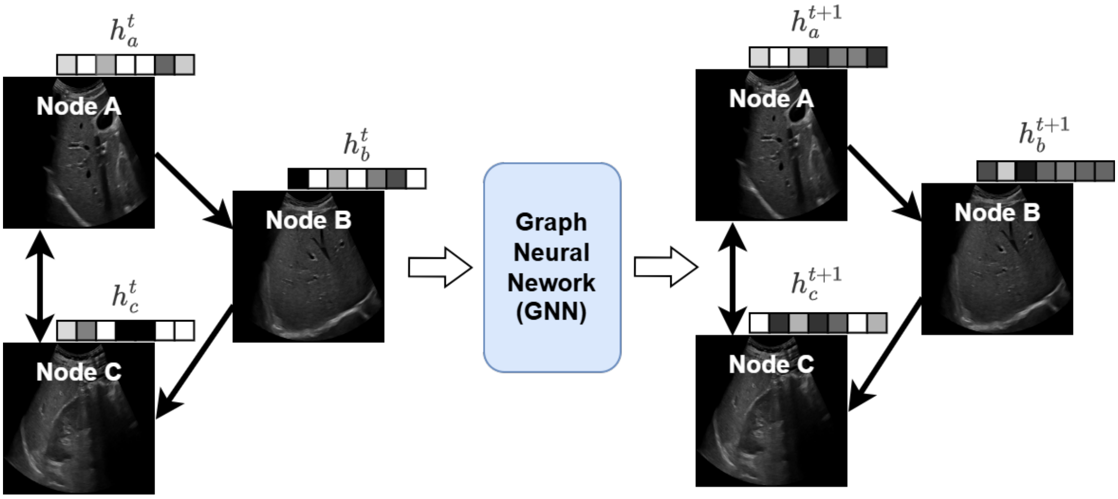


FIGURE 2.19 – An example of a GNN architecture applied to video analysis, showing node feature updates through layers.

While multiple strategies exists to transform H^t in H^{t+1}, one concise approach is described in [252]. This method applies a learnable linear transformation via a weight matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ to each graph node. Then, an attention mechanism $a : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is applied to model relationships between nodes :

$$e_{i,j} = a\left(\mathbf{Wh}_i^t, \mathbf{Wh}_i^t\right) \tag{2.16}$$

This attention mechanism can be implemented as a simple feedforward neural network (as in [252]) or as a dot product. To constrain the graph connections (focusing only on neighboring frames, for example), a masking operation $\mathcal{N}_i$ can be applied.Additionally, normalization is performed using the softmax function :

$$\alpha_{ij} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \tag{2.17}$$

The updated node features are computed as a weighted sum of connected node features in the graph, followed by a non-linear activation function σ :

$$\mathbf{h}_i^{t+1} = \sigma \left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} \mathbf{h}_j^t \right) \tag{2.18}$$

Finally, after processing the graph through l GNN layers, the final node features $h_i^{(l)}$ are used for video assessment. These features can be combined using a linear model W_{out} (a linear neural network, averaging, or attention mechanisms) to perform the desired video diagnosis task.

$$\hat{y} = f\left(W_{\text{out}} H^{(l)}\right) \tag{2.19}$$

A groundbreaking application of GNNs to ultrasound video diagnosis is presented in [173], where the authors estimate Ejection Fraction (EF) from echocardiography videos. The proposed framework, EchoGNN, begins by using a custom 3D CNN Video Encoder to extract features from individual frames. These frame features serve as node embeddings in a complete graph structure, where a neural message-passing GNN [79] updates the node features in a single step. Finally, the updated node features are passed to a Graph Regressor to produce real-valued EF predictions. The detailed architecture is illustrated in Figure 2.20.

Their work achieved comparable performance to other EF estimation models, with the added advantage of explainability through the analysis of graph nodes, which remain unaltered by the 3D Video Encoder.

Other works also utilize GNNs for EF estimation in echocardiography videos [241, 242]. Similar to previous approaches, these methods first extract frame features using backbone networks such as MobileNet v2 [211] and ResNet-based 3D CNNs [92]. However, unlike [173], where graph nodes represent individual frames, these methods model graph nodes as myocardial border points for each frame. This design allows the GNN to focus on optimizing node features for accurate left-ventricle segmentation, facilitating the identification of key frames and improving EF estimation.

In [241], the authors take this a step further by feeding the GNN outputs into a large language model, which generates natural-language explanations for

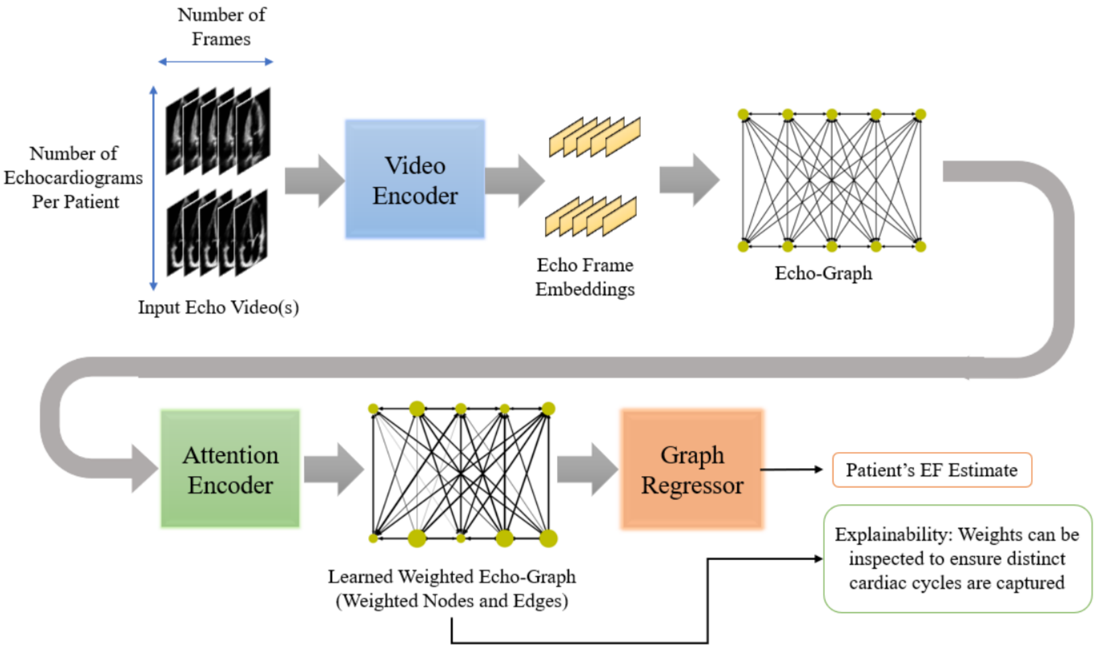

Number of Frames
Number of Echocardiograms Per Patient
Input Echo Video(s)
Video Encoder
Echo Frame Embeddings
Echo-Graph
Attention Encoder
Learned Weighted Echo-Graph (Weighted Nodes and Edges)
Graph Regressor
Patient's EF Estimate
Explainability: Weights can be inspected to ensure distinct cardiac cycles are captured

FIGURE 2.20 – EchoGNN has three main components. (1) Video Encoder : encodes video frames into vector embeddings while preserving the temporal dimension ; (2) Attention Encoder : infers weights over the nodes (video frames) and edges (relationships among frames) of the echo-graph ; (3) Graph Regressor : estimates EF using the inferred weighted graph. Figure and description extracted from [173].

the neural network's EF estimations, enhancing model interpretability and clinical applicability.

2.2.1.2 Two-Stream Networks

Two-stream neural networks, initially introduced in 2015 [223, 277], are a significant approach in video understanding. These architectures are characterized by two separate input branches (streams) : one dedicated to processing spatial information and the other focused on temporal information from video inputs.

In the context of ultrasound video analysis, the spatial branch typically processes b-mode video inputs (sequences of 2D images), extracting spatial feature representations using neural networks such as 3D CNNs, 2D CNNs combined with LSTMs, or attention-based models. Meanwhile, the temporal branch processes optical flow images, using similar network architectures to generate motion feature representations. Finally, the spatial and temporal features are fused or concatenated to estimate the video-level diagnosis.

In [73], the authors adopted this exact approach for classifying echocardiography videos, as depicted in Figure 2.21. To compute robust optical flow maps to use as input for the temporal stream, they utilized FlowNet2 [111], a neural network trained to extract optical flow from videos. Both streams (spatial and temporal) processes features extracted by a custom 2D CNN autoencoder, processed by LSTMs and combined using attention blocks. Finally, the features from the two streams (spatial and temporal) are concatenated to provide the final video-level diagnosis.

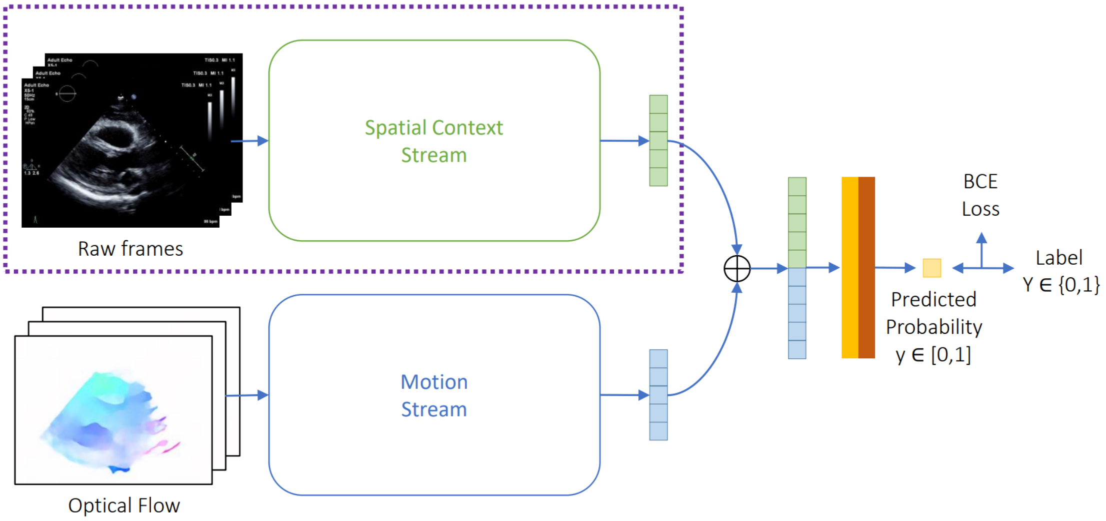


Figure 2.21 – Architecture of the two-stream neural network proposed in [73]. B-mode ultrasound frames and optical flow images are processed in separate spatial and temporal branches to extract features. These features are fused to produce the video-level diagnosis.

Another well-established two-stream neural network architecture is the Slow-Fast model [70]. This model utilizes frame rate sampling to define two streams : a slow stream (low frame rate) for processing spatial information and a fast stream (high frame rate) for processing temporal dynamics. Both streams are built using modified 3D ResNet encoders [71], with lateral connections enabling information sharing between the streams at multiple resolutions. This architecture was applied to ultrasound videos by [151] for classifying lung tumors with data acquired using a radial probe.

Other studies in ultrasound video analysis replace the motion stream with additional ultrasound modalities, similar to the approaches discussed for single-image cases in Section 2.1.1.4. For example, in [88, 145], the authors replaced the temporal stream with a CEUS stream to diagnose breast and renal lesions, respectively. Similarly, [264] employed two-stream neural networks for the automatic diagnosis of prostate cancer, replacing the temporal stream with an SWE video stream.

[54] while not a true two-stream network, the authors integrate it in R2+1D ResNet [247] by introducing a segmentation mask to it, which is used as a teacher

for a student network which receives solely the rgb video as input.

2.2.1.3 3D Neural Networks

Another approach to processing a sequence of ultrasound images is to treat it as 3D volumetric data, concatenating 2D frames to form a 3D volume with dimensions (W, H, C, F), where W and H represent the image width and height, C denotes the number of channels, and F corresponds to the number of frames. These models process spatial-temporal data simultaneously, normally with 3D Convolutions or specialized Transform-based architectures.

The most established architecture of this type consists of networks utilizing 3D convolutions, commonly referred to as 3D CNNs [246, 247]. These networks closely resemble traditional CNNs but employ layers composed of 3D convolutional filters. These filters operate across the W, H, and F dimensions with a sliding window, allowing the network to reason about spatial and temporal information simultaneously. The final fully connected dense layers and loss functions remain the same as those used in traditional 2D CNN-based networks. Figure 2.22 illustrates the core principle of 3D convolutions.

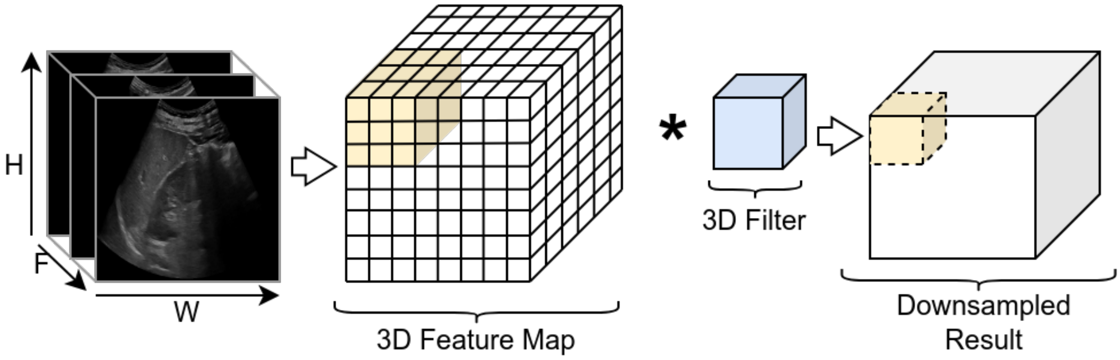


FIGURE 2.22 – Illustration of 3D Convolutions, inspired from [185].

Several authors have proposed CAD systems utilizing volumetric ultrasound data and 3D CNNs to address a wide range of pathologies [297, 292, 101, 152]. For example, in [292], the authors employ a 3D ResNet-50 [247] to detect lesions in breast ultrasound videos, demonstrating the superiority of video-based models compared to single-image approaches. Similarly, in [101], a self-supervised pretraining strategy using the same 3D ResNet-50 architecture [247] is proposed, significantly enhancing diagnostic performance in echocardiography video analysis.

Transformer-based neural networks have also been applied to volumetric data analysis, showcasing their potential in tasks such as echocardiography and thyroid diagnosis [5, 68, 174, 104]. As previously discussed, adapting Transformer models for imaging data primarily involves converting the input into a suitable sequence of tokens compatible with the Transformer architecture. In earlier sections, we highlighted methods that generate independent tokens for each video frame. Alternatively, tokens can be extracted from 3D patches of the input video, enabling the model to capture spatial and temporal relationships from the first layers of the model.

For example, in [5], the authors addressed the echocardiography EF estimation problem using a variant of the ViViT model [11]. Their approach involves extracting input tokens from tubelets, which are non-overlapping spatiotemporal patches spanning consecutive frames of the input video. These patches are then projected into the desired embedding dimension using a linear layer. Positional embeddings are added to the tokens, and the remainder of the Transformer architecture closely follows the ViT architecture [64]. Figure 2.23 illustrate the tubelets/token strategy.

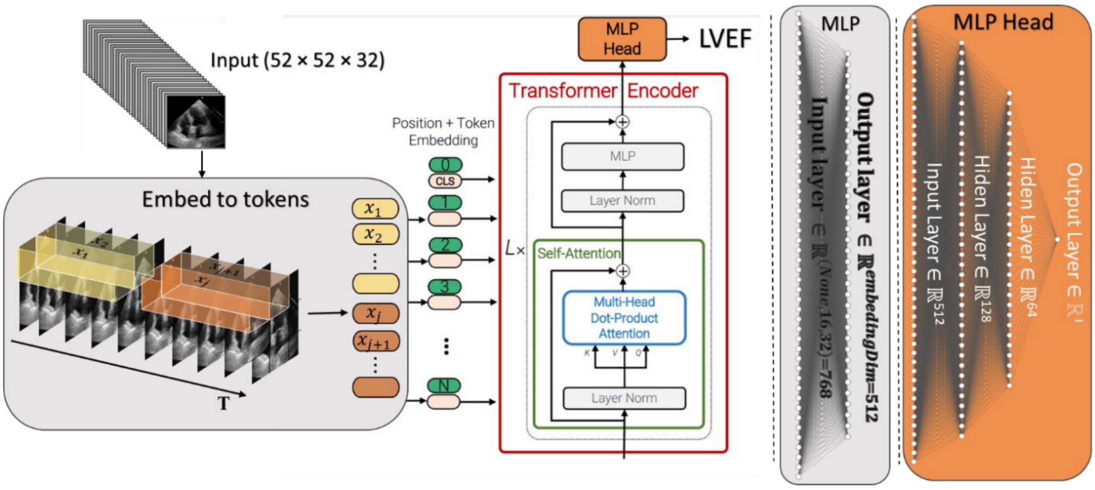


Figure 2.23 – Extraction of spatiotemporal tokens from tubelets in the video model proposed by [5], based on the ViViT architecture.

A similar approach is adopted in [174] for EF estimation, employing a video regression architecture based on the Uniformer [146]. This model combines the capabilities of 3D convolutions for capturing local spatiotemporal context and Transformers for modeling global dependencies. In this architecture, video tubelets are processed using 3D convolutional layers, transforming spatiotemporal patches into feature tokens. The Uniformer then uses these tokens in its hybrid Transformer-CNN architecture to produce the final EF estimation.

Another notable Transformer model for analyzing volumetric data is the Video Swin Transformer [163], which was adapted for EF prediction in ultrasound by [68]. The core principle of this model is that self-attention is computed locally within spatial-temporal window neighboring tokens, which are extracted from 3D patches of the input video using linear layers. To improve efficiency and enable multi-scale representation, a Patch Merging Layer is used after each Transformer block. This layer concatenates and downsamples neighboring tokens, reducing spatial-temporal dimensions. While attention is restricted to tokens within a local window, the Shifted Windows mechanism offsets the windows between layers, allowing the model to capture relationships across new regions. Figure 2.24 illustrates the Patch Merging and Shifted Windows mechanisms.

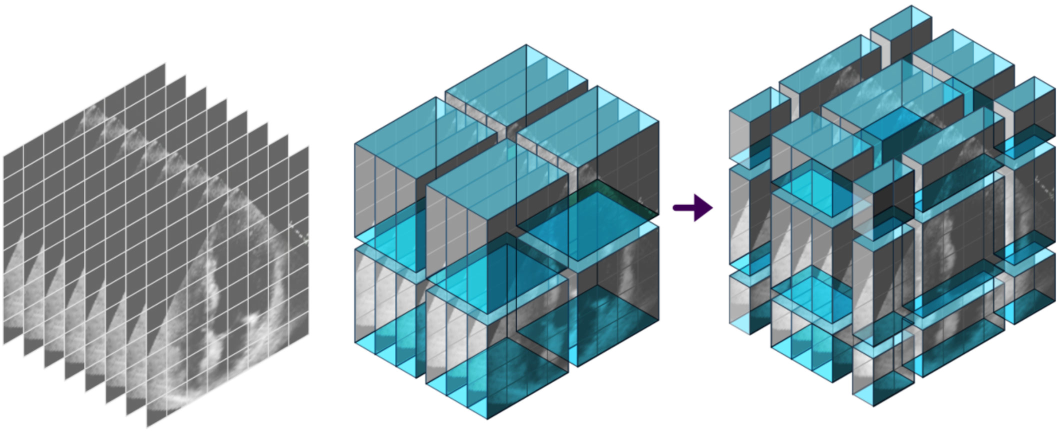


Figure 2.24 – Patch Merging and Shifted Windows mechanisms in the Video Swin Transformer. Patch Merging downsamples spatial-temporal dimensions, while Shifted Windows ensure expanded attention regions across Transformer layers. Figure extracted from [68]

Other approaches have explored combining 3D CNNs and Transformer blocks for the analysis of ultrasound video data. In [104], the authors integrate a 3D ResNet [71] with the Video Swin Transformer [163] to diagnose thyroid nodules. Similarly, [186] incorporates a modified 3D ResNet architecture, replacing certain layers with self-attention mechanisms, to predict fetal birth weight from ultrasound videos.

2.2.2 Guided Video-Based Diagnosis Models

All the end-to-end models discussed in the previous section rely exclusively on video-level labels for training. While these models have demonstrated strong performance as reported in their respective studies, they often don't consider valuable auxiliary information that could guide the learning process more effectively. Integrating additional data (like frame-level annotations, segmentation masks, or clinical data) has the potential to enhance both training efficiency and inference accuracy, leading to more robust and interpretable models.

The guided video-based diagnosis models typically consist of a video classification or regression neural network (as discussed in the previous section) and an auxiliary neural network trained on an external dataset. This auxiliary neural network is used either as a frame-selection tool (as discussed in Section 2.1.2 regarding image selection) or as a source of additional information fed to the video classifier during training, inference, or both.

A good example that incorporates both aspects was introduced by [270], where the authors proposed a guided approach for diagnosing liver lesions in ultrasound. Their framework provide guidance from three DeepLabv3 models [200], which perform segmentation of the scan region, liver, and lesions in still images. Only frames containing detected lesions are forwarded to the next step, while the segmentation masks are used to exclude pixels that do not correspond to the liver parenchyma or lesions.

The filtered images are then processed independently by a Frame Extractor CNN (DenseNet121 [103]), and the resulting features are passed through LSTM blocks to incorporate temporal information. Additionally, the lesion segmentation masks are used to estimate mass sizes, which are used for generating a mass-attention guidance signal. Finally, this attention signal is applied to the outputs of the LSTM layers, giving higher weights to larger lesions, and the combined features are passed through MLP layers to produce the final lesion diagnosis. The second part of the system (without the segmentation models) is illustrated in the Figure 2.25.

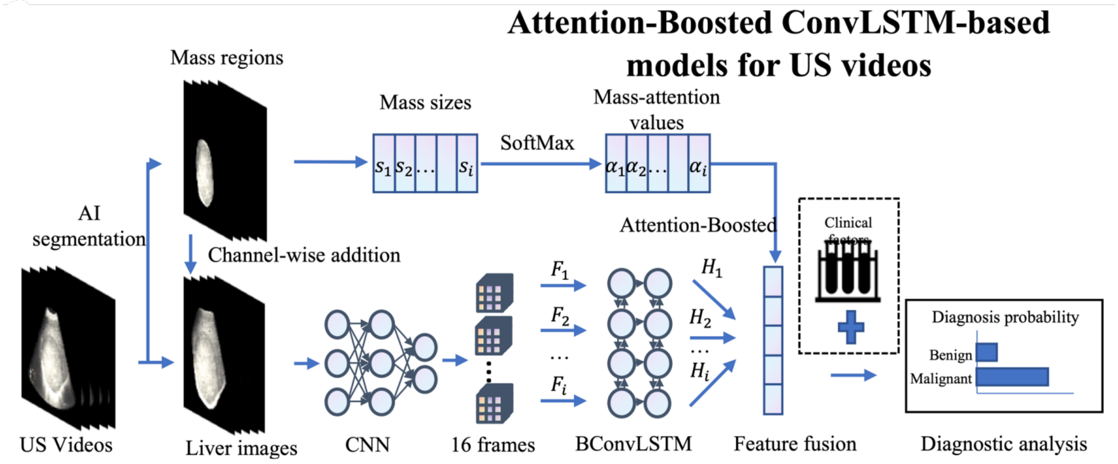


FIGURE 2.25 – Illustration of the second part of the guided liver lesion diagnosis system proposed by [270]. Features extracted by a DenseNet121 from filtered frames are passed through LSTM blocks for temporal modeling. A mass-attention guidance signal, generated using lesion segmentation masks and size estimates, is applied to the LSTM outputs, prioritizing larger lesions. The combined features are processed by MLP layers to produce the final diagnosis. Image extracted from [270].

In [258], the authors perform the diagnosis of breast lesions in ultrasound videos using a lesion detector and a keyframe localization model as guidance. The lesion detector is based on Faster R-CNN [197]. Given a video input, features are extracted from the detected lesion ROIs in individual frames and fed into an LSTM model. This model is connected to fully connected layers to output a single score indicating whether the current frame is a keyframe. Since each input video contains only one keyframe, pseudo-labels are generated for the remaining frames based on their proximity to the annotated keyframe. The entire guidance system is trained end-to-end. Finally, the authors sample a clip around it and perform lesion diagnosis using a modified 3D CNN (C3D [246]).

Other works in this category include [134], where the authors employ breast lesion segmentation maps to support the diagnosis of breast ultrasound lesion characterization in videos. In [133], the authors introduce a method for detecting cardiac abnormalities in fetal ultrasound videos. Their approach is trained exclusively on healthy cases and integrates a YOLO object detector [195] to guide the detection process.

2.3 Research Gaps

Limitations of Human-Annotated Data

The first point to address is the challenge of acquiring high-quality datasets necessary for training and evaluating AI-based CAD systems in ultrasound. Gold-standard diagnostic labeling techniques, such as histopathology from biopsies or MRI, are impractical for large-scale use due to their invasiveness and associated costs. As an alternative, most studies rely on visual annotations provided by radiologists or other healthcare practitioners.

Interpreting ultrasound images visually is inherently challenging, introducing a significant degree of subjectivity in annotation labels. Even among radiologists with similar levels of expertise, their labels can vary substantially. This variability can hinder the development of deep learning models, which depend on consistent and reliable annotated data for training. An open challenge remains in finding effective methods to reduce subjectivity in visual annotations, minimizing the impact of annotator bias.

Challenges in Untrimmed Video Training and Inference

A significant challenge in developing AI-based CAD systems capable to be deployed in untrimmed ultrasound video data. Unlike single-image or trimmed video datasets, untrimmed ultrasound videos contain a high amount of diagnostically irrelevant frames, including frames affected by artifacts, noise, or background content. Existing methods either require strong supervision, such as frame-level annotations or segmentation masks, or rely on external guidance systems to identify diagnostically relevant frames. These approaches are labor-intensive and are difficul to be extended to new pathologies, requiring careful design.

Additionally, current methods face difficulties in balancing computational efficiency with the requirement to process long video sequences while preserving both spatial and temporal resolution. Developing models capable of automatically identifying diagnostically relevant frames with minimal or no explicit guidance, while also enabling real-time inference, remains an open challenge.

Integration in Real-World Applications

To democratize access to ultrasound screening, CAD methods must provide enough automation to be operable with ease by non-experts. These systems should guide operators during the screening process by automatically identifying diagnostically relevant information, thus eliminating the need for specialized expertise. Furthermore, the system must be capable of indicating when a diagnosis is not feasible, prompting the user to capture higher-quality images or seek further assistance.

While various works in the literature address some of these functionalities individually, they often lack integration into a comprehensive system. Developing a unified CAD framework that meets these requirements, while being trainable under medical data constraints and adaptable to a wide range of pathologies or conditions in b-mode ultrasound, remains an open research challenge.

3. CVL+RankNet : A New Approach to Label Images for Computer-Assisted Diagnosis

Chapter summary

A substantial amount of annotated data is essential to develop AI-based Computer-Assisted Diagnosis (CAD) systems, which largely rely on deep learning models. This data is used not only for training but also for testing and performance monitoring as demanded for certification as a medical device. The challenge of obtaining accurate and representative labeled datasets to train such models imposes significant limitations to advancing CAD with abdominal ultrasound. Additionally, access to patient records is often restricted due to stringent data protection laws, such as the General Data Protection Regulation (GDPR) in Europe. Consequently, the labeling task is typically performed by specialized annotators relying on visual inspection.

Under these conditions, annotators often struggle to determine the stage or severity of the pathology by visually inspecting ultrasound images. For instance, a pathology regularly diagnosed visually using US is non-alcoholic fatty liver disease (NAFLD), also called liver steatosis, which is characterized by the accumulation of fat tissue in the liver and can lead to various complications if not addressed. The difficulty of obtaining objective visual annotations that convey not only the presence or absence of the disease but also the severity hinders the development of AI-based solutions to assist in early diagnosis and to track the progression after clinical treatments.

In this chapter, we assess the reliability of visual annotations for labeling liver steatosis cases in abdominal ultrasound images. We measure the precision of visual annotations by comparing them to histopathological examinations, and study the reasons behind the errors in visual annotations. We then propose a new annotation method for CAD, called Comparative Visual Labeling (CVL), based on relative annotations. They significantly improve annotation accuracy and consistency between annotators, while also providing continuous-valued labels that correlate very well with disease progression measured by histopathology.

3.1 Additional background

3.1.1 Our case study application : Liver steatosis detection in B-mode ultrasound :

Liver steatosis is a common condition characterized by fat accumulation within liver cells. It is particularly prevalent, affecting approximately 25% of the global population [10], with a higher prevalence in developing countries [275]. Steatosis is associated with either alcohol consumption or metabolic conditions, known as Non-Alcoholic Fatty Liver Disease (NAFL), and the latter may develop chronic conditions such as type 2 diabetes and metabolic syndrome [16]. If HAFL is not identified and treated early, it can initiate an inflammatory process known as Non-Alcoholic SteatoHepatitis (NASH), and causes liver injury, leading to scaring produced by the response of hepatic stellate cells and then liver fibrosis fibrosis. If this process continues, it can evolve into liver cirrhosis, which severely impairs liver function and may lead to liver cancer like hepatocellular carcinoma (HCC) [38, 202, 136, 39]. Figure 3.1 illustrates this process. As a consequence, early detection of liver steatisis is essential, as well as severity grading for treatment monitoring.

As stated in the background section, B-mode ultrasound (US) screening is the preferred imaging modality to diagnose many abdominal conditions due to its convenience, low cost, and non-invasiveness [29]. Liver steatosis is one of various conditions regularly detected with B-mode abdominal ultrasound.

In general, although b-mode US visual examination for steatosis is convenient and harmless to the patient, it requires a skilled clinician to operate the US probe and produce an accurate diagnosis. As with US-based diagnosis in general, interpreting US images and anatomical/pathological features requires significant expertise, is highly operator-dependent, and necessitates years of training to achieve confidence and accurate diagnosis skills [63].

Another significant challenge in US-based liver steatosis diagnosis is the fact that a clinician’s mental decision boundary when differentiating between healthy and pathological subjects differs among clinicans. In the early stages of hepatic steatosis, distinguishing between healthy and pathological imaging characteristics can be particularly challenging, often relying heavily on the radiologist’s experience, which can vary widely, and there may be up to 20% underestimation of liver steatosis [127]. In the study by [140], an 87% agreement was observed between MRI-PDFF (defined below) and B-mode ultrasound annotations by radiologists, with ultrasound diagnostics showing a notable tendency to underestimate the severity of steatosis.

Due to the above issues, the need for skilled clinicians to perform a US examination has led to a global shortage of access to care, and many people are screened less frequently than necessary or not at all. This lack of early liver steatosis screening leads to fewer patients being diagnosed at the early stages of the disease, where timely intervention could significantly improve outcomes and reduce the costs associated with intensive care.

The above limits have motivated research in the past few years in AI-assisted liver steatosis detection in B-mode ultrasound images presented in section 2.1. An important consideration in all such methods is how to obtain ground-truth labels, required for training and validating CAD systems

3.1.2 Obtaining ground-truth labels for liver steatosis : Approaches and open challenges

Three methods have so far been used to obtain ground-truth labels for liver steatosis CAD : 1) Histopathological labeling, 2) MRI labeling and 3) visual assessment of ultrasound images (referred to as 'Visual labeling'). We briefly discuss each method and the limitations we aim to overcome with comparative visual labeling.

Histopathological labeling. Histopathological examination is the gold standard method used in clinical practice to precisely quantify the severity of liver steatosis. This method involves extracting a liver sample via fine-needle aspiration biopsy, which is then microscopically analyzed in a laboratory [18, 238]. The standard biomarker is the *Percentage of Fatty Hepatocytes* (*PFH*), giving the proportion of liver cells (hepatocytes) that contain excessive fat deposits. Patients with PFH below 5% are classified as healthy (Figures 3.2(a) and 3.2(b)), whereas those with a PFH of 5% or more are classified as pathological (images 3.2(c) to 3.2(h)).

However, histopathological labeling presents various practical and ethical challenges. First, it is an invasive procedure requiring image-guided biopsy, which carries risks of complications and is generally avoided unless necessary. Second, it is costly, requiring medical intervention for sample collection and the expertise of a histopathologist for diagnosis. Thirdly, since a biopsy is typically performed only when there is a suspicion of liver abnormalities, there is an inherent population bias towards pathological patients, which may affect the accuracy and generalizability of models trained with such data [250].

Additionally, there is a risk of label misalignment, where the ultrasound data may be acquired at a different location to the biopsy sample. When the disease is in its early stage and not well diffused through the liver, this can result in systematic labeling errors.

MRI labeling. An alternative method to obtain labels is with Magnetic Resonance Imaging Proton Density Fat Fraction (MRI-PDFF). MRI-PDFF is a quantitative liver fat concentration biomarker obtained during a magnetic resonance (MRI) scan, making it a non-invasive option. It is calculated as the fraction of liver proton density attributable to fat, derived from the ratio of fat to water images acquired during the MRI procedure. Several studies have demonstrated a strong correlation between MRI-PDFF and histopathology. In [82], 635 participants underwent MRI-PDFF and histological examination. The study reported that the sensitivity and specificity of MRI-PDFF for classifying liver steatosis were 0.93 and 0.94 for healthy vs. Grade 1–3, 0.74 and 0.90 for healthy vs. Grade 2–3, and 0.74 and 0.87 for healthy vs. Grade 3. In addition, a meta-analysis of 24 studies with 2,979 NAFL patients, reported in [83], confirms the accuracy of MRI-PDFF. The study found that MRI-PDFF had high diagnostic accuracy with a Hierarchical Summary Receiver Operating Characteristic (HSROC) of 0.97 for detecting steatosis at Grade 1 or higher, 0.91 for Grade 2 or higher, and 0.90 for Grade 3 or higher.

However, using MRI to obtain labels presents two main challenges : the high costs associated with MRI and the need to enroll patients for MRI scans that are not required in their standard clinical pathway. Additionally, similarly to histopathology, MRI-based labeling may also result in label misalignment.

Visual labeling. In this approach, clinicans or trained annotators label the US data by directly inspecting the US data according to clinical guidelines. In the literature, one form of visual labeling has been attempted, where a binary label is assigned to each image according to the visual evidence presented in each image. We refer to this visual labeling method as Single-image labeling (SVL). Table 3.1 shows three cues that have been used for SVL : the echogenicity of the liver parenchyma, the visibility of the liver diaphragm, and the visibility of the hepatic vessels, with the echogenicity of the liver parenchyma being the most informative cue.

In practice, these cues cannot be used to reliably obtain continuous labels to quantify pathology severity, unlike PFH. Nevertheless, there exist some guidelines to use the cues in Table 3.1 to coarsely categorize severity into three grades as follows :

1. **Mild liver steatosis (Grade 1) :** This occurs when the PFH is between **5% and 33%.** In ultrasound images, this is chacterized by a minimal increase in hepatic echogenicity, with normal visualization of the diaphragm and intra-hepatic vessel borders
2. **Moderate liver steatosis (Grade 2) :** This occurs when the PFH is between **34% and 66%.** In ultrasound images, this corresponds to a moderate increase in hepatic echogenicity and slightly impaired visualization of the intra-hepatic vessels and diaphragm
3. **Severe steatosis (Grade 3) :** This occurs when the PFH **above 66%.** In ultrasound images, this is marked by a significant increase in echogenicity and poor or non-visualization of the hepatic vessels and diaphragm.

Given the practical constraints of histopathology and MRI-PDIFF labeling, SVL is an attractive approach for generating large-scale datasets without misalignment issues. However, labeling errors caused by inconsistency in US-based visual assessment of the disease could result in suboptimal CAD models and/or unreliable performance evaluation and monitoring—critical aspects when deploying CAD models as medical devices. These errors and uncertainties may arise from the subtle and difficult-to-distinguish pathological differences in B-mode ultrasound images, making consistent and accurate labeling particularly challenging. The challenges of SVL are illustrated in Figure 3.2, which shows examples from subjects in the Byra Dataset [29] with varying degrees of liver steatosis, from 3% PFH (healthy) to 80% PFH (severe) fat infiltration. The differences between pathological findings in neighboring images, particularly between healthy livers and those in the earlier stages of steatosis, are subtle, and they may be severely obscured by variable image quality.

In summary, visual labeling offers huge practical advantages for acquiring datasets necessary to train and evaluate AI models for liver steatosis CAD. However, an open and important research question is how to reduce the errors and biases associated with SVL ?

3.2 Methodology

3.2.1 Section overview

To address the challenges of SVL, we propose a new visual labeling method for assessing pathologies in medical image data, based on comparative assessment. We refer to this as *Comparative Visual Labeling* (*CVL*). In the case of ultrasound image labeling, we propose to label the images in pairs, where the annotator

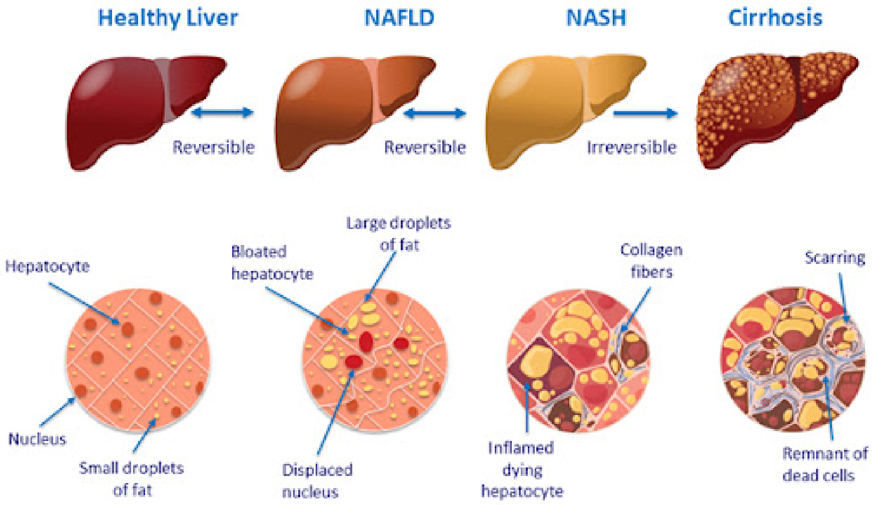
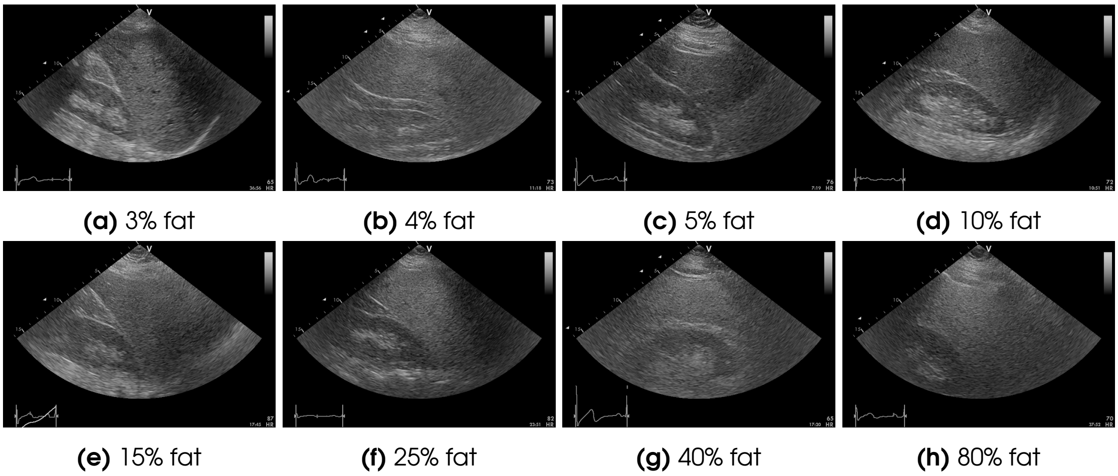


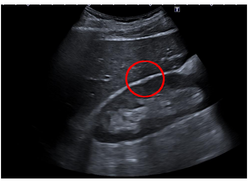
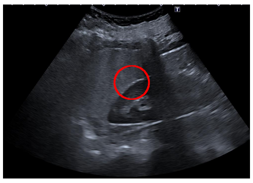
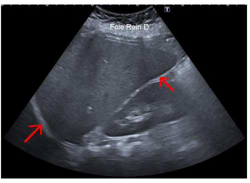
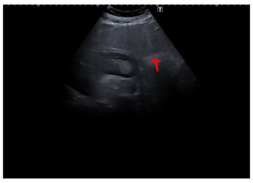
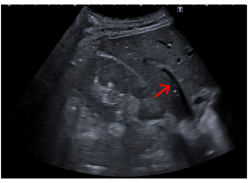
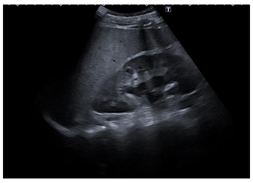
FIGURE 3.1 – Progression of NAFLD from liver steatosis to cirrhosis and potential hepatocellular carcinoma [35].


(a) 3% fat (b) 4% fat (c) 5% fat (d) 10% fat

(e) 15% fat (f) 25% fat (g) 40% fat (h) 80% fat

FIGURE 3.2 – Images extracted from the Byra Dataset [29] showcasing multiple stages of hepatic steatosis. These images demonstrate different levels of fat accumulation in the liver, as verified through histopathological examination of fine-needle biopsy samples. The dataset provides visual evidence of steatosis progression, ranging from early to advanced stages.

Table 3.1 – Guidelines for diagnosing steatosis using b-mode ultrasound images. The primary visual indicator of pathology is increased echogenicity of the liver parenchyma. Images from our property 'Dataset 2', described in detail in Section 3.2.5.2. Diagnosis is typically performed in the liver-kidney view, where the echogenicity of the liver is compared relative to the kidney parenchyma. Figures edited using GIMP's crop and draw tools with GIMP 2.10.28 [240].

Visual Feature	Description	Healthy	Pathological
Liver parenchyma echogenicity	In pathological cases, the echogenicity (brightness) of the liver parenchyma is increased, appearing whiter on ultrasound images. This enhancement is most effectively assessed by comparing it to the echogenicity of the kidney parenchyma.		
Liver diaphragm visibility	In healthy cases, the liver diaphragm is well-defined and appears with a strong white color. However, the presence of fat in the liver can cause the diaphragm to become blurry or even non-visible.		
Liver vessels visibility	The identification of well-defined blood vessels is a sign of a healthy liver. The presence of fat in the liver reduces their visibility, serving as an indicator of steatosis.		

assesses which of the images has a greater indication of the pathology. In principle, CVL can be applied to various pathologies, but here, it demonstrates its value for steatosis CAD. We hypothesize that Comparative Visual Labeling (CVL), which focuses on annotating the relative severity of the disease between image pairs rather than determining absolute disease presence or absence, can provide more objectivity and reduce variability and sensitivity biases in labeling. We also propose a way to easily convert these relative annotations into a single continuous-valued label for each image, we use a Learning To Rank (LTR) approach implemented with RankNet [24].

We propose a method to transform Comparative Visual Labels into continuous-valued per-image pathological scores, using a Learning to Rank approach. In the experimental validation section, we show that the pathological scores correlate extremely well with the gold standard (cell fact percentage assessed with histopathology). The results show, for the first time, that a continuous score can be produced from visual labeling of US data that reflects disease severity. We also show how to utilized CVL to train neural networks for single-image steatosis estimation.

3.2.2 Learning-to-Rank (LTR) and RankNet

Learning-to-Rank (LTR) is a general class of methods to rank data based on comparative relevance measurements. LTR has received much attention in web search engines, where the user inputs a query, and the LTR algorithm ranks and outputs the most relevant web pages. In this context, the goal of an LTR algorithm is to score the web pages according to relevance while ensuring that relevance aligns with the user's goal of obtaining the information they seek on the most relevant pages. LTR is also widely used in e-commerce to rank products according to user behavior. For instance, when a user clicks on a product over others, the system interprets this as a preference, ranking that product higher in future searches. This data helps the LTR algorithm provide personalized results for the user based on their behavior. A similar process occurs on social networks, where LTR ranks content and advertisements, ensuring users see more relevant posts and ads based on their previous interactions. Figure 3.3 illustrate some of these examples.

The objective of an LTR model is as follows. We denote as x_q a query (for example, a keyword search) and a set of input features as $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, where each vector $\mathbf{x}_i$ represents features extracted from the i^{th} input. For example, the features could be the word frequencies extracted from a webpage whose relevance is to be computed. LTR finds a scalar relevance function Φ that computes the relevance of each input with respect to x_q. We denote as s_i the relevance score for the i^{th}, computed using Φ with the following expression :

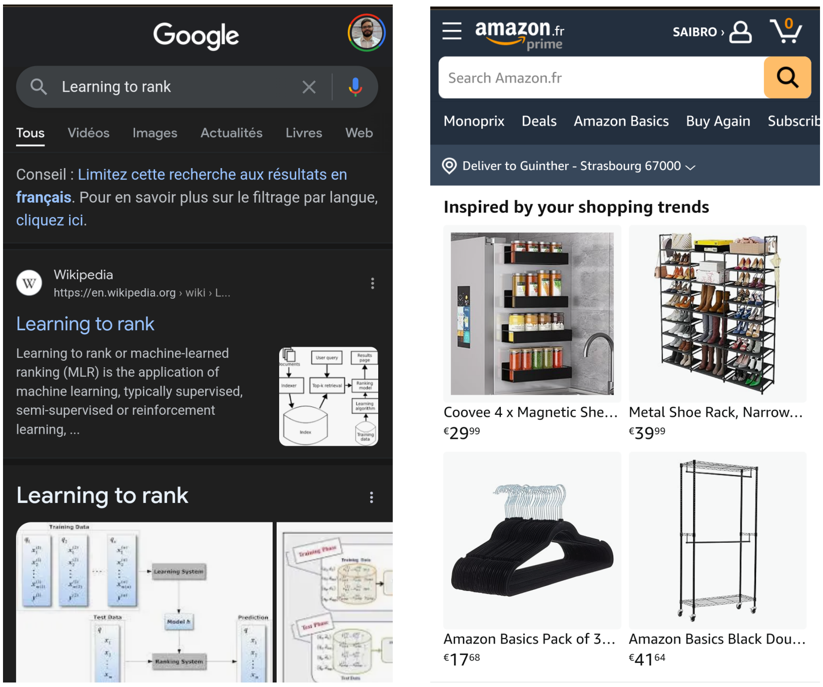
$$s_i = \Phi\left(\mathbf{x}_i; \theta; x_q\right) \tag{3.1}$$

and where θ denotes the learnable parameters of Φ that are optimized during training.

In the LTR literature, learning normally involves fitting θ to labeled training data, by optimizing a loss function of the following form :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}\left(\theta; x_q, X, Y\right) \tag{3.2}$$

where $\mathcal{L}$ denotes a loss function, and Y denotes a set of labels. Unlike standard supervised learning, where labels correspond to the models' expected outputs,



Conseil : Limitez cette recherche aux résultats en français. Pour en savoir plus sur le filtrage par langue, cliquez ici.

Wikipedia
https://en.wikipedia.org › wiki › L...

Learning to rank

Learning to rank or machine-learned ranking (MLR) is the application of machine learning, typically supervised, semi-supervised or reinforcement learning, ...

Learning to rank

Deliver to Gunther - Strasbourg 67000

Inspired by your shopping trends

Coovee 4 x Magnetic She...
€29^{99}

Metal Shoe Rack, Narrow...
€39^{99}

Amazon Basics Pack of 3...
€17^{68}

Amazon Basics Black Dou...
€41^{64}

(a) Google Search **(b)** Amazon Marketplace

FIGURE 3.3 – Example Applications of Learn to Rank (LTR) Algorithms. 3.3(a) a Google web search for the query "Learning to Rank," displaying the most relevant results, and 3.3(b) recommended products on the Amazon Marketplace homepage. Both sites were accessed on September 26, 2024, using a Pixel 7a. The figures were edited using GIMP 2.10.28's crop and draw tools[240].

in LTR, such labels may not always be available. This can result from the inherent difficulty in having users objectively assess the relevance of an input as a scalar function.

As a consequence, several labeling methods have been considered in the LTR literature, grouped into three categories :

1. **Pointwise labels :** The labels are defined for each training example, providing the example's continuous-valued relevance score with respect to x_q.
2. **Pairwise labels :** The labels are defined for pairs of training examples. Each label states which member of the pair has higher relevance with respect to x_q.
3. **Listwise labels :** This extends pairwise labels to multiple training examples. The labels provide the order of relevance of the examples with respect to the query x_q.

The choice of label method determines the type of loss function that can be used. For Pointwise labels, a regression loss is normally used, and commonly implemented with the sum of squared error loss :

$$\mathcal{L}_{SSE}\left(\theta; x_q, X, Y\right) = \sum_{i \in [1,N]} \left(\Phi\left(\mathbf{x}_i; \theta, x_q\right) - y_i\right)^2 \tag{3.3}$$

where $y_i \in Y$ denotes the continuous-valued label for the i^{th} input.

For Pairwise labels, the loss is normally based on variants of the loss proposed in RankNet [24, 26, 25] as follows. We define as $P(i \succ j)$ the probability that the model predicts that training example i should be ranked higher than example j. This is evaluated via Φ using the Sigmoid function σ as follows :

$$P(i \succ j) = \sigma\left(\Phi\left(\mathbf{x}_i; \theta, x_q\right) - \Phi\left(\mathbf{x}_j; \theta, x_q\right)\right) \tag{3.4}$$

We define as y_{ij} the label associated to the pair (i, j) where $y_{ij} = 1$ if the i^{th} input should be ranked higher than j, and $y_{ij} = 0$ otherwise. The pairwise loss evaluates the difference between the predicted probabilities $P(i \succ j)$ and the labels using binary cross entropy as follows :

$$\mathcal{L}_{PW}\left(\theta; x_q, X, Y\right) = -\sum_{(i,j) \in C} y_{ij} \log\left(P(i \succ j)\right) + \left(1 - y_{ij}\right) \log\left(1 - P(i \succ j)\right) \tag{3.5}$$

where $C \subset [1, N] \times [1, N]$ denotes the set of pairs for which a label y_{ij} exists. It is not necessary, or desirable in practice, to obtain labels for all possible pairs. Various methods exist for selecting pairs, including random subset selection, or selection based on minimum spanning trees. We note that, ignoring pais with no difference $y_{ij} = 0$, the input pairs can be reordered *a priorir* so that the pairwise labels are always positive : $y_{ij} = +1$. This reordering simplifies the loss to the following :

$$\mathcal{L}'_{PW}\left(\theta; x_q, X, Y\right) = -\sum_{(i,j) \in C} \log\left(P(i \succ j)\right) \tag{3.6}$$

For Listwise labels, several different loss functions have been proposed, based on either Normalized Discounted Cumulative Gain (NDCG) or Mean Average Precision (MAP). NDCG considers the relevance of training examples, while penalizing incorrect orderings more heavily for items ranked near the top. It also uses a normalized loss to make balance the loss when training with multiple ranked lists or queries. One of the main challenges of NDCG is that, in its original form, it is non-differentiable. This has been overcome using either a differentiable approximation such as SoftRank [239], or using iterative optimization such as LambdaRank [25] or LambdaMART [26].

Additionally, losses based on NDCG tend to favor highly-ranked training examples because, in many applications such as web search, a user is mainly only concerned with the most relevant web pages. This can be a drawback for non-traditional applications like ours, where the most challenging patients to diagnose are those in the early stages, who are not top-ranked with respect the visual extent of the pathology. Additionally, listwise labels may take more time to label compared to pairwise comparisons because of the need to sort items.

3.2.3 Comparative Visual Labeling (CVL)

In CVL, we assign a pairwise label to a pair of medical images, where the label conveys which of the two images shows greater pathology severity. Consequently, these labels express the relative pathology severity of the image pair. We then transform the set of Comparative Visual Labels to absolute, continuous pathology severity scores by training a RankNet. This contrasts SVL, where each image is annotated individually with absolute labels.

The differences between CVL and SVL are illustrated in Figures 3.4 and B.1. We recall that in SVL, binary labels are defined as follows :

$$D_i = \begin{cases} 1, & \text{if } \mathbf{I}_i \text{ appears pathological} \\ 0, & \text{if } \mathbf{I}_j \text{ appears healthy} \end{cases} \tag{3.7}$$

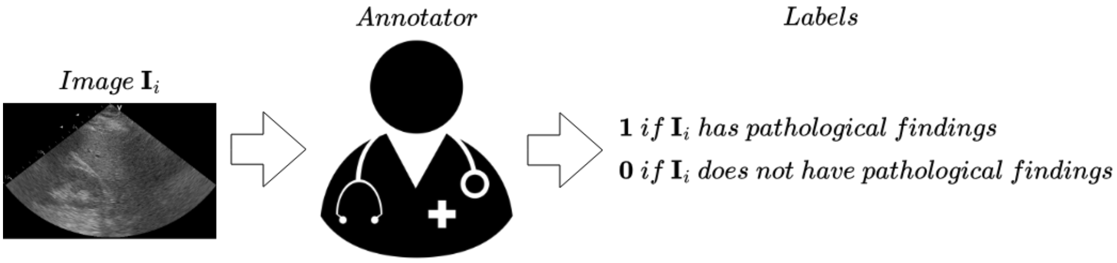

Annotator
Labels
$Image\ \mathbf{I}_i$
$\mathbf{1}\ if\ \mathbf{I}_i\ has\ pathological\ findings$
$\mathbf{0}\ if\ \mathbf{I}_i\ does\ not\ have\ pathological\ findings$

FIGURE 3.4 – Illustration of the Single-Image Visual Labeling (SVL) annotation process. The annotator assigns a binary label $\{0, 1\}$ (healthy versus pathological) to the image based on the visual findings listed in Table 3.1 and their own experience.

In CVL, we define as **S** $= \{I_{i,j} \mid i \in \{1, 2, \ldots, n\}, j \in \{1, 2, \ldots, n\}, i \neq j\}$, a set of image pairs that are to be annotated with CVL. One can view **S** as a graph where each member is a connection between two images, which are the graph's nodes. In practice, **S** does not need to include all possible pairs, which would be prohibitively expensive to annotate. However, **S** must have a single connected component. That is, every image can be linked to every other image, by traversing the edges in the graph. This is important because it is necessary to convert pairwise annotations (which can be seen as attributes on the graph's edges), to pathological score, using a RankNet as described in Section. We discuss two methods to generate **S**, as used in our experiments, in Section 3.2.5.3.

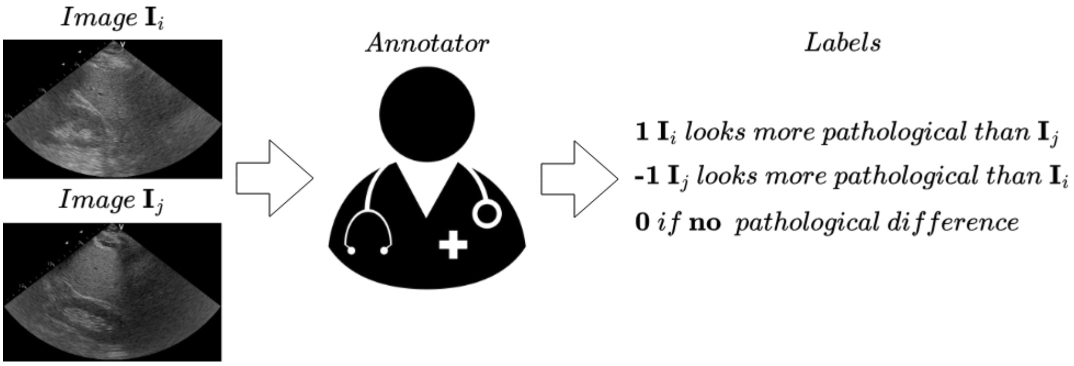

$Image\ \mathbf{I}_i$
$Image\ \mathbf{I}_j$
$Annotator$
$Labels$
$\mathbf{1}\ \mathbf{I}_i\ looks\ more\ pathological\ than\ \mathbf{I}_j$
$\mathbf{-1}\ \mathbf{I}_j\ looks\ more\ pathological\ than\ \mathbf{I}_i$
$\mathbf{0}\ if\ \mathbf{no}\ pathological\ difference$

FIGURE 3.5 – Illustration of the Comparative Visual Labeling (CVL) annotation process. The annotator assigns a binary label the annotator assigns labels belonging to $D_{i,j} \in \{1, -1, 0^+, 0^-\}$ to the image pair $I_{i,j}$ based on which image is perceived as having a higher degree of pathology severity. In our case study of liver steatosis, this assessment is determined by comparing the evidence for the visual indicators given in Table 3.1.

Each image pair $\{I_{i,j}\} \in \mathbf{S}$, has an associated Comparative Visual Label that we define as follows :

$$D_{i,j} = \begin{cases} 1, & \text{if } \mathbf{I}_i \text{ appears more pathological than } \mathbf{I}_j \\ -1, & \text{if } \mathbf{I}_j \text{ appears more pathological than } \mathbf{I}_i \\ 0^-, & \text{if both } \mathbf{I}_i \text{ and } \mathbf{I}_j \text{ are healthy} \\ 0^+, & \text{if both } \mathbf{I}_i \text{ and } \mathbf{I}_j \text{ are pathological and indistinguishable} \end{cases} \tag{3.8}$$

We used the label 0^+ to avoid situations where an annotator would be forced to distinguish disease severity when, in practice, they cannot due to insufficient visual evidence. Table 3.2 shows some examples of Comparative Visual Labels.

3.2.4 Transforming Comparative Visual Labels to continuous pathology severity scores with RankNet

We now present our approach to transform a set of Comparative Visual Labels into per-image, continuous pathology scores, by adapting and training an LTR model implemented with RankNet. We refer to this approach as CVL+RankNet, and it is motivated by two important uses. Firstly, the continuous pathology scores are effectively continuous labels that, as shown in the experimental results section for liver steatosis, correlate well with PFH values in its early stage (which is the most clinically relevant range for screening with CAD). Secondly, the scores can be

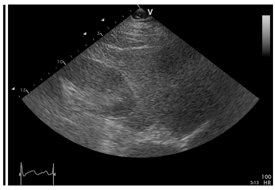
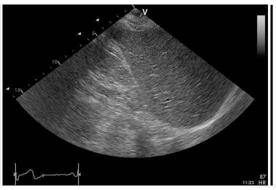
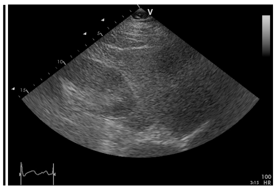
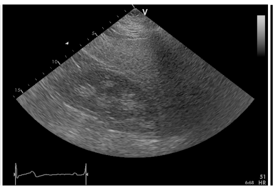
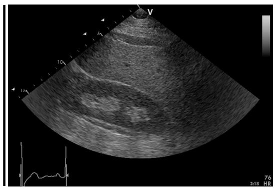
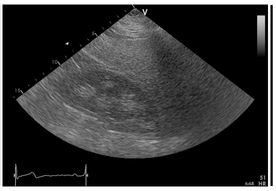
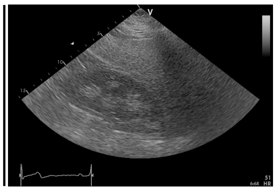
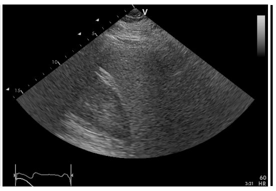
Image 1	Image 2	Label (y_{12})
		0^-
		-1
		0^+
		1

TABLE 3.2 – Examples of Comparative Visual Labels for 'Dataset 1'. The labels are defined in Equation 3.8.

binarized (dichotomized) at any desired specificity/sensitivity threshold, to meet according to the specific application. Unlike SVL, such binarization does not need to be set in stone at the time of annotation, and an adjustment of the threshold does not require any re-annotation by humans.

Term	Semantics : LTR for search engines	Semantics : LTR for Image-based CAD
Datatype	Webpage	Medical image
Relevance score	Webpage relevance	Perceived severity of pathology
$P(i \succ j)$	Probability that webpage i is more relevant than i	Probability that image i shows greater pathology severity than i
Query	User search input	Type of pathology

TABLE 3.3 – Mapping of key terms from Learning-to-Rank (LTR) literature in web search to their application in computer-aided diagnosis (CAD).

The LTR problem. We define our LTR problem as follows : Given a dataset of medical images $\mathbf{I} = \{I_i \mid i \in \{1, 2, \ldots, N\}\}$, where $\mathbf{x}_i$ denotes the features associated to the i^{th} image, our objective is to estimate a relevance score $s_i \in \mathbb{R}^+$ associated with each image, for a query pathology x_q.

In our context, the relevance score corresponds to the severity of the pathology, and it is determined by a function Φ as introduced in Equation 3.1. In Table 3.3, we map the key terminology used in the prior LTR literature focused on web search to the new application of CAD.

We propose to model Φ using a small neural network based on RankNet [135, 24], and trained from Comparative Visual Labels, as depicted in Figure 3.6.

Although other alternative LTR models to RankNet exist, such as LambdaRank [25] and LambdaMART [26], RankNet was chosen for its satisfaction in other ranking tasks and its open-source implementation. RankNet has also shown relatively stable performance without requiring significant hyper-parameter turning, such as the number of hidden layers and other architectural choices.

It is important to emphasize that this approach is not intended to train a model that produces pathology severity scores for unseen images (not contained in the training dataset). Rather, its purpose is only to transform Comparative Visual labels into pathology scores for each image in the training dataset via Φ.

RankNet training methodology. We illustrate the RankNet in Figure 3.6. During training, each training image I_i is assigned to a unique binary string $\mathbf{x}_i \in \{0, 1\}^N$ using one-hot encoding, where N is the total number of images. Consequently, $\mathbf{x}_i = [0, \ldots, 1, \ldots, 0]$ where 1 appears in the i^{th} position in $\mathbf{x}_i$. The string $\mathbf{x}_i$ is passed into a RankNet with N neurons in its input layer (connected to each bit $\mathbf{x}_i$). The RankNet has one output neuron, producing the pathology score s_i. The training forward pass operates as follows. For each image pair (i, j) with a CVL label, the binary strings $(\mathbf{x}_i, \mathbf{x}_j)$ are passed through the RankNet, which outputs their respective scores s_i and s_j. The pair's rank probability $P(i \succ j)$ is then evaluated from s_i and s_j, using Equation 3.4. The rank probabilities of all pairs are then compared against the ground-truth rank probabilities (determined from the pairs' Comparative Visual Labels), using a loss function, detailed in the following section. Once the RankNet is trained, we obtain the real-valued pathological score of each training image, referred to as its **CVL+RankNet** score, as $s_i = \Phi\left(\mathbf{x}_i; \hat{\theta}, x_q\right)$, where $\hat{\theta}$ denotes the RankNet's trained parameters.

The RankNet can be trained by minimizing the loss function with backpropagation and a suitable optimizer such as Adam [131]. We provide our implementation details used for the experiments of this chapter, including training parameter settings, and the hidden layer architecute in 3.4, which were fixed for all experiments.

Loss function. Recall that we propose using 4 comparative visual labels : Two of the labels ($+1$ and -1) indicate a perceivable difference in pathology severity. For these labels, we apply the pairwise Binary Cross Entropy loss as defined in Equation 3.5. In contrast, the other label values (0^- and 0^+) correspond to no perceivable pathology difference. We have considered different options to handle these labels. The first option was to add a loss that encourages the relevance score of each image in the pair to be the same. A naive way to approach this is to add an equality loss for the 0^- and 0^+ labels, such as a sum of squared differences :

$$\mathcal{L}^0_{PW}\left(\theta; x_q, X, Y\right) = \sum_{(i,j)\in C} \mathbb{1}(y_{ij} = 0)\left(\Phi\left(\mathbf{x}_i; \theta, x_q\right) - \Phi\left(\mathbf{x}_j; \theta, x_q\right)\right)^2 \tag{3.9}$$

where $\mathbb{1}$ denotes the indicator function. The losses may be combined using a weighting term $\lambda \mathbb{R}^+$ $\mathcal{L} = \mathcal{L}_{PW} + \lambda \mathcal{L}^0_{PW}$ where λ balances the influence of pairs with no perceivable pathology difference. During training, each training image I_i is assigned to a unique binary string $\mathbf{x}_i \in \{0, 1\}^N$ using one-hot encoding, where N is the total number of images. Consequently, $\mathbf{x}_i = [0, \ldots, 1, \ldots, 0]$ where 1 appears in the i^{th} position in $\mathbf{x}_i$.

However, in our experiments, we noticed that using a positive λ generally resulted in worse performance than $\lambda = 0$. This can be explained by the fact that zero labels are used when an annotator cannot confidently perceive, among two images, a difference in the pathology severity. However, this does not necessarily imply that there is no pathology difference. As such, we view the assignment of the label 0 as a case of missing data, and in our experimental results, we use only $\mathcal{L}_{PW}$ as the loss function.

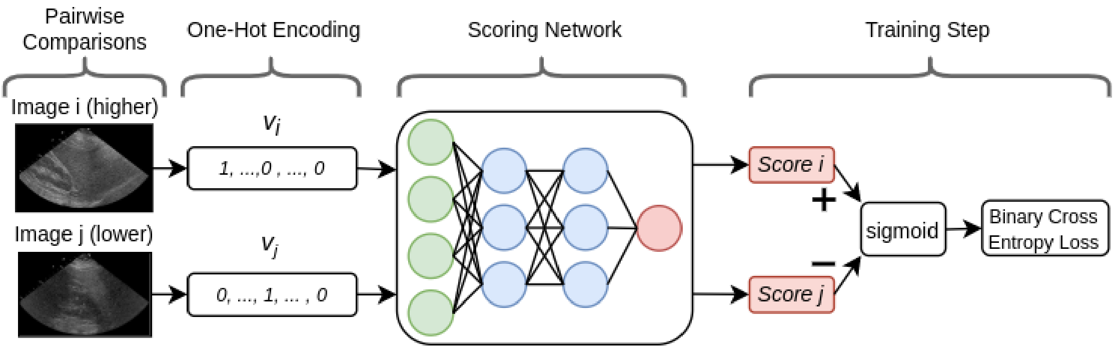


FIGURE 3.6 – Scheme of our implementation of RankNet [24], a LTR neural network trained with pairwise comparison data.

Dichotomizing pathology severity scores. Per-image binary classification labels (healthy vs. pathological) can be produced from the CVL+RankNet scores by applying a threshold τ, as described in Equatioin (3.10). An annotator can decide this threshold as follows : Firstly, the images are presented to the annotator in order of CVL+RankNet score. The annotator selects the first I_a and last I_b image between which they believe the boundary exists between healthy and pathological cases. We then compute τ as $\tau = \frac{1}{2}(\Phi\left(\mathbf{x}_a; \hat{\theta}, x_q\right) + \Phi\left(\mathbf{x}_b; \hat{\theta}, x_q\right))$. In our results section, discuss this method to select τ in greater detail.

$$y_i = \begin{cases} 1, & \text{if } s_i \geq \tau \\ 0, & \text{if } s_i < \tau \end{cases} \tag{3.10}$$

3.2.5 Implementation of CVL+RankNet for labeling and classifying liver steatosis in ultrasound images

3.2.5.1 Section overview

We now describe our implementation of CVL+RankNet for labeling liver steatosis in ultrasound image sets.This section is organized as follows. First, we describe the datasets used for training and validation. Next, we describe the cohort of annotators used to label the data with CVL and SVL. Next, we describe the RankNet implementation and training hyper-parameters (which have been fixed for all experiments). Finally, we describe the AI models trained using to automatically detect liver steatosis from US images, using CVL+RankNet scores as labels.

3.2.5.2 Datasets

Two datasets were used. The first dataset was the Byra Dataset [29] [1], referred herin as Dataset 1. This was the first, and to this date, the only, publicly available dataset consisting of US images paired with ground-truth PFH values from histopathology. The dataset was collected anonymously with written informed consent at the Department of Internal Medicine, Hypertension, and Vascular Diseases at the Medical University of Warsaw in Poland [123, 29], and comprises 55 severely obese patients (mean age 40.1 ± 9.1 years, mean BMI 45.9 ± 5.6, 20% male) collected within two days before bariatric surgery. The B-mode ultrasound data was obtained using the GE Vivid E9 Ultrasound System (GE Healthcare INC, Horten, Norway), equipped with a convex abdominal 2.5 MHz probe with harmonic imaging. Ten B-mode US images of the liver and kidney were captured for each patient in the

[1] The dataset is available for download at : https://zenodo.org/records/1009146.

liver-kidney sagittal plane, at a resolution of 434x636 pixels (550 images in total). The images were very similar, and captured with the probe in the same position. Consequently, intra-patient image variability was small, with only noticeable changes in speckle patterns and mild organ movement due to respiration. Figures 3.8(a) and 3.8(b) showcase examples of healthy and pathological cases, respectively. A liver biopsy was performed on each patient using the subcapsular part of the left liver lobe for histopathological examination. The distribution of liver steatosis in Dataset 1 is provided in Figure 3.7.

The second dataset, referred to as Dataset 2, was collected anonymously with written informed consent at the MIM clinic in Strasbourg, France, and comprises 54 patients collected during routine abdominal examinations. The B-mode ultrasound data was obtained using a Canon Aplio a450 system (Canon Medical Systems, Ōtawara, Tochigi, Japan), equipped with a convex abdominal probe. One US image was collected for each patient, showing the liver and kidney in the liver-kidney sagittal plane, at a resolution of 434x636 pixels. Figures 3.8(c) and 3.8(d) showcase examples of healthy and pathological cases, respectively.

Datasets 1 and 2 differed in three important ways. Firstly, Dataset 1 had ground truth from histopathology, whereas Dataset 2, which represented images from routine ultrasound abdominal procedures, did not. Secondly, the images came from different devices, and thirdly, Dataset 1 consisted of obese patients admitted for bariatric surgery using a prospective data collection protocol. In contrast, Dataset 2 consisted of patients from routine abdominal examinations (not specifically for the purposes of bariatric surgery). As a result, there was strong class distribution difference between datasets, where Dataset 1 had a much higher proportion of pathological cases (69%) compared to Dataset 2 (33% - estimated using SVL as described in the following section).

3.2.5.3 Image pair selection and annotation

For a datset of N images, the number of possible image pairs is $N(N-1)\,/\,2$ (quadratic in N). Consequently, it was important to apply a pair selection strategy to reduce the amount of annotation effort. As discussed in Section 3.2.3, the image pair set can be viewed as a graph whose nodes are images and edges are labels. This graph must have one connected component; otherwise, it is impossible to transform the labels to continuous pathology severity scores with any LTR method, including RankNet, where RankNet training would otherwise be an ill-posed problem. For Dataset 1, where images from the same patient were extremely similar, only the first image of each patient was used for annotation ($N = 55$ images). For Dataset 2, all images were used for annotation ($N = 54$

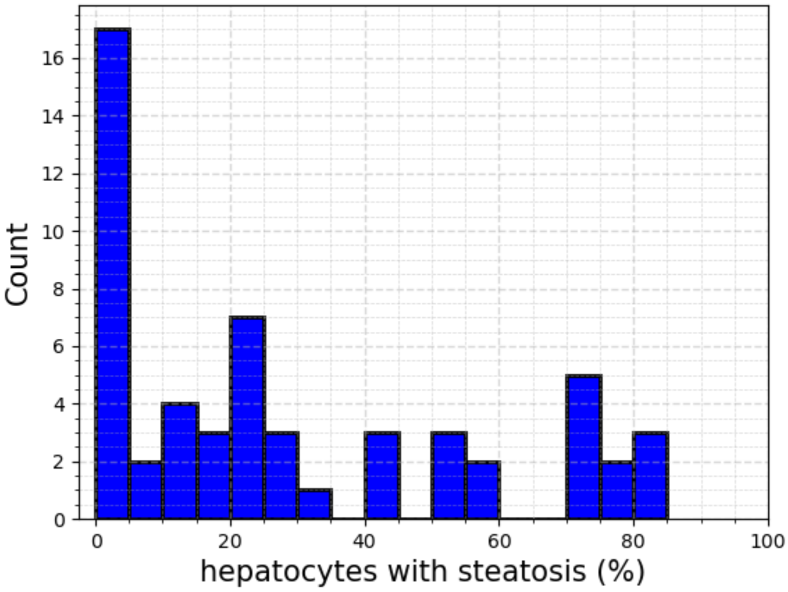


FIGURE 3.7 – Distribution of Percentage of Fatty Hepatocytes (PFH) values for patients in Dataset 1, determined from histopathological examination of liver biopsy samples [29]. 69% of patients had steatosis (PFH > 5%), and there was a broad range of steatosis severity, ranging from mild to severe.

images).

For both datasets, we used a simple technique of pair selection, which produced a single connected component and graphs with redundancy (i.e. edges could be removed while maintaining a single connected component). We exploited this redundancy to assess the impact of reducing the number of image pair annotations using edge pruning, as described later in Section 3.3.3. Pairs were established by taking each image and pairing it with $0 < p \leq N$ other images in the dataset using random selection. In our experiments, we used $p = 18$, representing a large amount of redundancy and generating graphs with one connected component.

Each image pair was then annotated independently by three experienced radiographers from IRCAD's annotation team. The annotator profiles were as follows :

- **Annotator A :** Team leader radiographer with 11 years of experience in ultrasound image analysis.
- **Annotator B :** Radiographer with 7 years of experience in ultrasound analysis.
- **Annotator C :** Radiographer with 6 years of experience in ultrasound analysis.

To perform the annotation tasks, the image pairs were horizontally concatenated and presented to each annotator in random order using the CVAT annotation tool [216]. The concatenated images were annotated with the CVL labels as defined in Equation (3.8) using CVAT.

To measure the potential benefit of label fusion (i.e. combining labels from different annotators), we created additional labels ('Fused Labels') using the majority vote from the annotators. Because we used 4 label classes as defined in Equation 3.8, if each annotator labeled a pair differently, there would be no majority. However, this did not occur. The images of each dataset were also annotated independently by each annotator with SVL. This was performed to analyze label quality of both techniques. Like CVL, label fusion was also performed, creating a fourth set of labels per dataset, using majority voting.

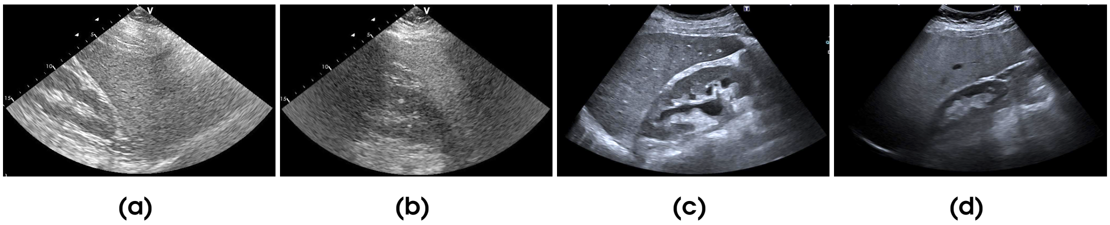


Figure 3.8 – Sample images from Datasets 1 and 2 used to train and evaluate CVL+RankNet. (a) and (b) show healthy and pathological cases from Dataset 1, respectively. (c) and (d) show health and pathological cases from Dataset 2, respectively. To improve visibility, the brightness of the images has been increased by 150% using GIMP's exposure filter (GIMP 2.10.28) [240].

3.2.5.4 RankNet training

We trained eight RankNets, with four RankNets per dataset. The first three RankNets were trained using the Comparative Visual Labels from Annotators A, B and C. The fourth RankNet was trained using Fused Comparative Visual Labels. For each RankNet, the same architecture and training hyper-parameters were used. The architecture involves a single hidden layer, presented in Table 3.4. In the results section, we show that performance was relatively insensitive to three key architecture hyper-parameters.

The RankNets were implemented in Python 3.8 using Keras 2.6, and trained with Adam optimization [131]. This took approximately one minute per RankNet using a standard workstation PC with a consumer-grade GPU (NVidia RTX 3090). Once training was finished, the associated CVL+RankNet scores were computed for each RankNet, by forward passing the training images through the RankNets.

Model : "RankNet_CVL"

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 55)	0	
input_2 (InputLayer)	(None, 55)	0	
Sequential (Sequential)	(None, 1)	6,617	Input_1[0][0] Input_2[0][0]
Subtract (Subtract)	(None, 1)	0	Sequential[0][0] Sequential[1][0]
activation (Activation)	(None, 1)	0	subtract[0][0]

Total params : 6,617
Trainable params : 6,617
Non-trainable params : 0

TABLE 3.4 – The RankNet architecture used to transform Comparative Visual Labels to continuous pathology severity scores.

3.2.5.5 Using CVL+RankNet scores for automatic liver steatosis detection

By pairing the CVL+RankNet scores with their associated images, we could train two kinds of liver steatosis detection models, as shown in Figure 3.9. They were as follows :

- **Regression models** : These models were trained to predict pathology severity from an ultrasound image, using CVL+RankNet scores as training labels.
- **Binary classification models** : These models were trained to classify ultrasound images into two classes : healthy or pathological. The training labels were established by thresholding the CVL+RankNet scores as described in Section 3.2.4.

These models could be implemented and trained using any state-of-the-art approach. We chose the Inception-ResNet-v2 model [233] for its strong performance in image classification, though other competitive network architectures are also suitable.

We configured Inception-ResNet-v2 for binary classification by replacing its final layer with two neurons (one for each class) with Softmax activation. It was trained using the Binary Cross Entropy Loss and class weighting to handle class imbalance. We configured Inception-ResNet-v2 for regression by removing the classification head and replacing it with a regression head. This was implemented using a max pooling layer, then a flattening layer, followed by three fully connected layers with 512, 16, and 1 neurons, respectively. The regression model was trained with the Sooth L1 Loss [80] with $\beta = 1$.

Both models were pre-trained on ImageNet and trained using a standard workstation equipped with a GeForce RTX 3090 GPU and PyTorch 1.7.1. To improve generalization, data augmentation was applied as a pre-processing step, involving random geometric and photometric perturbations of the training images. The same augmentation parameters were used for both networks, and the complete list of image pre-processing operations is provided in Table 3.5. The training hyper-parameters of both models are listed in Table 3.6. Training typically took approximately 20 minutes for each model.

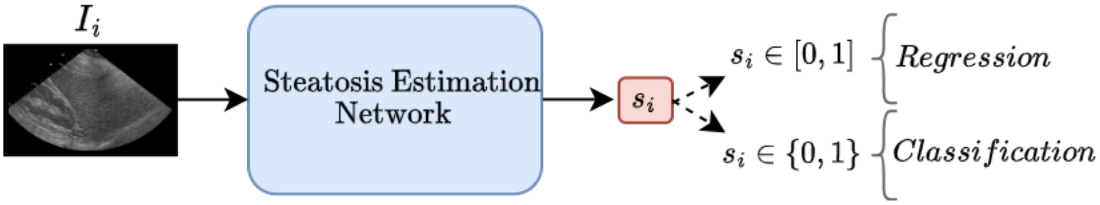


FIGURE 3.9 – Illustration of two neural network models trained to detect liver steatosis from a B-mode ultrasound image using CVL+RankNet scores. Top-right : binary classification (predicting whether an ultrasound image shows the presence or absence of steatosis (pathological vs. healthy). Bottom-right : regression (predicting a continuous-valued pathology severity score from an ultrasound image).

TABLE 3.5 – List of image pre-processing operations with value ranges (where applicable) and application probabilities. All operations have were implemented using Albumentations 1.1.0 [27].

Operation	Value Range	Probability
Crop Ultrasound ROI	—	100%
Rand. horizontal Flip	—	50%
Rand. rotation	[-15º, +15º]	100%
Rand. horizontal Translation	[-10%, +10%]	100%
Rand. vertical Translation	[-10%, +10%]	100%
Rand. scaling	[95%, 105%]	100%
Rand. Gaussian Blur	[0, 2.0]	50%
Rand. Erasing	[2%, 33%]	50%
Resize	[299, 299]	100%

TABLE 3.6 – Training hyperparameters for our three proposed architectures : classification and regression.

Parameter	Value (classification)	Value (regression)
Learning rate	auto lr find	auto lr find
Gradient accumulation	true	true
Class balance	inverse frequency	none
Batch size	64	64
Min. epochs	50	200
Max. epochs	400	400
Early stop min. delta	0.00001	0.00001
Early stop count	7	7
Backbone	Inception-ResNet-v2	Inception-ResNet-v2

3.3 Results

3.3.1 Section overview

We now present our experimental validation of the methodology described in the previous section. This section is organized as follows. First, we analyzing the quality of Single-image Visual Labels and Comparative Visual Labels, against ground truth from histopathology. For both labeling methods, we quantify each annotator's error rates and the Fused Label error rates (established by majority voting). Next, we analyze the quality of CVL+RankNet scores compared to ground truth and measure statistical agreement (correlation) in different severity bands. Next, we present an experiment to study the influence of the number of pairwise comparisons on the quality of CVL+RankNet scores. Finally, we analyze the performance of classification and regression models trained using CVL+RankNet scores as labels.

3.3.2 SVL and CVL error analysis

This analysis involved only Dataset 1 for which ground truth labels were available. The distribution of label errors using SVL is illustrated in Figure 3.10, showing the percentage of mislabeled images, grouped into four categories (severity bands) : healthy ($\leq 5\%$ fatty hepatocytes), mild (grade 1, $5-33\%$), moderate (grade 2, $33-66\%$) and severe (grade 3, $> 66\%$) liver steatosis. For all annotators, most errors occured in the mild band, indicating the difficulty in labeling mild cases, which, however, are the most clinically relevant cases for early disease detection. The results showed a tendency of annotators B and C to label mild liver steatosis as healthy using SVL, reflecting previous findings showing liver steatosis can be underestimated by as much as 20% [127, 142].

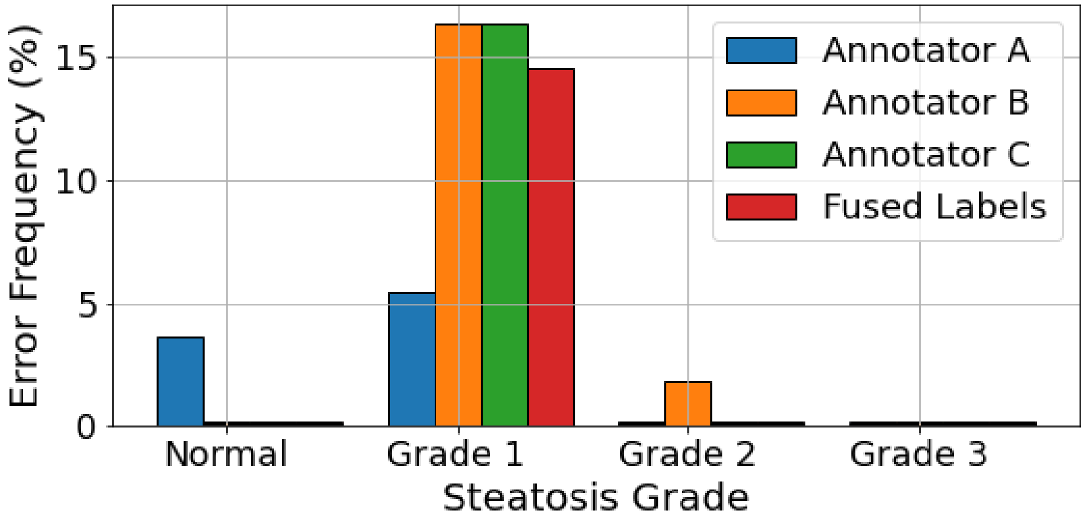


FIGURE 3.10 – SVL error rates of each annotator and Fused Labels. Error rates are grouped into four categories, corresponding to error rates for healthy images and error rates for pathological images in three severity bands (grades) as defined in Section 3.1.2 - Histopathology labeling. Each bar group has 4 bars, representing the error rates of the annotators and also that of Fused Labels (computed using the annotators' majority vote).

Figure 3.11 shows the CVL error rates when annotators used labels $+1$ and -1 (i.e., when they perceived a difference in pathology severity when, according to ground truth, there was no difference). We grouped the errors into five categories (shown by the five groups of bar plots), to highlight that CVL errors depend on the actual difference in pathology severity. To this end, we used five categories : 1) image pairs with PFH differences 1) below 15%, 2) between 15-30%, 3) between 30-50%, 4) between 45-60%, and 5) above 60%. For each category, three error bars are shown, giving the error frequency of each annotator.

The highest error rates occurred in category 1, when the difference in liver steatosis severity was smallest. This was expected, given that this interval contained the most similar pathological grading, which can be hard to distinguish. Nonetheless, the error rates were nevertheless relatively low in category 1, and below $<$ 3.5% for all annotators. We observed a general trend to lower error rates from categories 1 to 5. There were no errors in category 5.

Figure 3.12 shows the distribution of CVL errors when annotators used labels 0^- and 0^+ (the equality labels). For these labels, an error occurred when they perceived no pathology difference, when, according to ground truth, there was a difference. Concretely, the figure shows the probability of label errors on the y-axis,

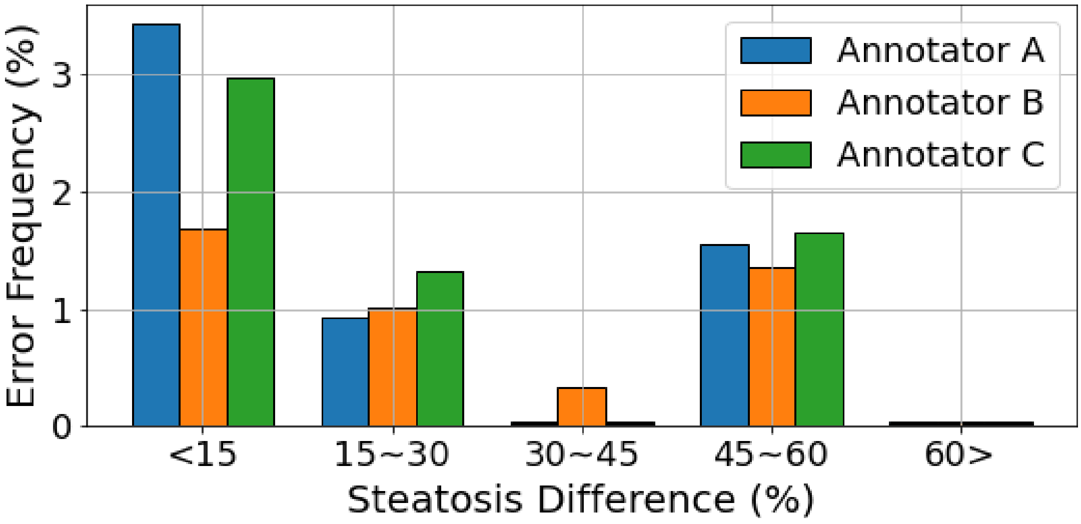


FIGURE 3.11 – Distribution of CVL errors using Dataset 1 for labels $+1$ and -1. The error rates of each annotator are shown, as well as the error rates of Fused Labels using majority voting. To reflect the fact that the error rate depends on the actual difference in pathology severity (measured by the absolute difference in PFH values), we grouped the misclassified image pairs into 5 categories (shown by the five groups of barplots), ranging from a small difference (category 1 : $< 15\%$) to a very large difference (category 5 : $> 60\%$). Solid black horizontal bars represent an error rate of 0.

where a label was considered an error if the difference in PFH values exceeded a threshold τ_{GT}. We ranged τ_{GT} from 1 to 100% (representing the maximum possible difference in PFH values.). For example, with $\tau_{GT} = 20\%$ (which, in practice, is a substantial difference in liver steatosis severity), the label error rates were approximately 0.22 for all annotators. In contrast, with $\tau_{GT} = 10\%$, the label error rates increased to approximately 0.56 for all annotators. For $\tau_{GT} > 73\%$, there were no labeling errors for any annotator. The figure also shows the probability of label errors was remarkably similar between annotators.

The results indicate that when the annotators used equality labels, they were often notable differences in PFH values. Nevertheless, we recall that the annotators used the equality labels when they could not perceive a difference in pathology severity. Therefore, these errors may be explained due to limited visual evidence. This finding also justifies our decision to exclude equality labels in the RankNet loss function (Section 3.2.4 - Loss function).

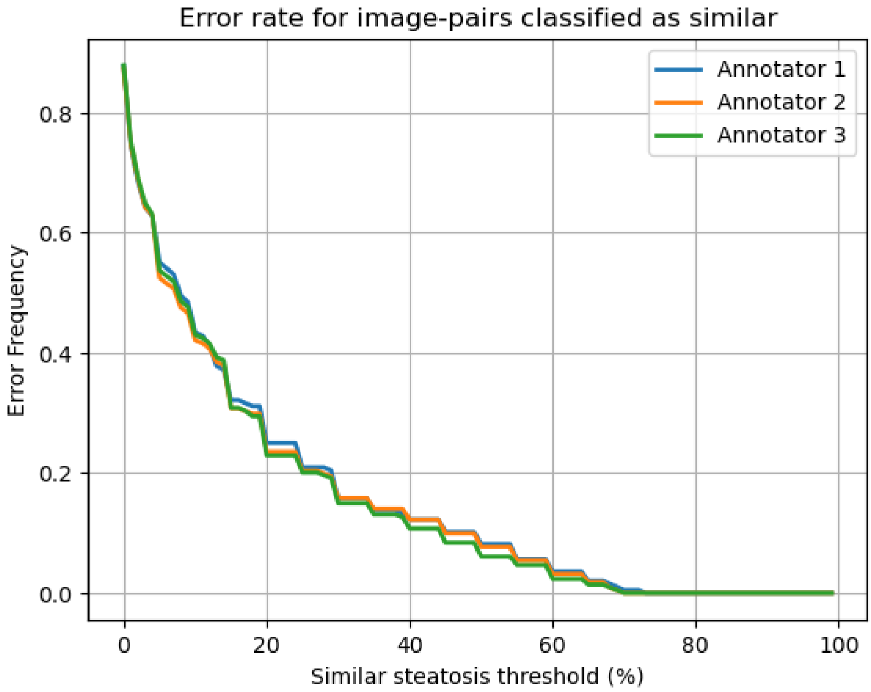


FIGURE 3.12 – Probability of equality label errors for each annotator.An error occurred when they perceived no pathology difference, when, according to ground truth, there was a difference. Concretely, the figure shows the probability of label errors on the y-axis, where a label was considered an error if the difference in PFH values exceeded a threshold τ_{GT}. We ranged τ_{GT} from 1 to 100% (representing the maximum possible difference in PFH values).

3.3.3 CVL+RankNet performance for binary labeling

F1 and ROC-AUC performance metrics. We compared the per-image labels produced by CVL+RankNet against ground truth labels using F1 (the harmonic mean of sensitivity and specificity), and ROC-AUC (area under the receiving operating curve) metrics. Ground truth labels were established by thresholding the image's PFH value at 5% according to clinical guidelines [29]. The CVL+RankNet scores were thresholded by the annotators using the approach described in Section 3.2.4 - Dichotomizing pathology severity scores. We generated ROC curves (and consequently ROC-AUC) for CVL+RankNet, by interpreting the CVL+RankNet scores as detection confidence values. As such, ROC curves could be produced by varying the decision threshold τ (discussed in Section 3.2.4) from 0 (corresponding to 100% sensitivity where all images are classified as pathological) to the maximal value of 100 (corresponding to 100% specificity where all images are classified as healthy).

We used SVL was used as the baseline labeling method. We note that in contrast to CVL+RankNet, it was impossible to assess ROC-AUC with SVL because it produced only a set of binary labels (from which it was impossible to generate a ROC curve). This also highlights a key virtue of CVL+RankNet compared to SVL : the threshold may be adjusted to achieve a desired label sensitivity/specificity without requiring the annotators to relabel images.

Table 3.7 – Label quality metrics (F1 and ROC-AUC) evaluated on Dataset 1 with SVL and CVL+RankNet. The lower 2.5% and upper 97.5% confidence intervals (CIs) are shown in brackets.

Annotator Method	Annotator A	Annotator B	Annotator C	Fused labels
SVL (F1)	0.92 $[0.85, 0.98]$	0.83 $[0.72, 0.92]$	0.85 $[0.75, 0.93]$	0.87 $[0.77, 0.94]$
CVL+RankNet (F1)	0.99 $[0.96, 1.00]$	0.93 $[0.86, 0.98]$	0.93 $[0.86, 0.98]$	0.97 $[0.93, 1.00]$
CVL+RankNet (AUC)	0.99 $[0.90, 1.00]$	0.97 $[0.88, 0.99]$	0.95 $[0.88, 0.99]$	0.99 $[0.89, 1.00]$

The results are shown in Table 3.7 where we observe the following. The F1 performance of SVL varied considerably between annotator A and the other two annotators, and SVL with Fused Labels had a lower F1 performance than Annotator A. This result could be attributed to the lower performance of Annotators B and C, which negatively influenced the fused labels. Compared to SVL, F1 performance of CVL+RankNet was substantially higher for all annotators and Fused Labels.

We assessed statistical significance using the 95% CI of paired differences between SVL and CVL+RankNet labels (implemented with bootstrap resampling with 5000 samples). Significance was found for Annotators A, B and Fused Labels. McNemar's test at ($\alpha = 0.05$) was also performed to assess, for each annotator, whether the differences between their SVL and CVL+RankNet labels were significant. The p-values were $p = 0.059$ (significant), $p = 0.034$ (significant), $p = 0.096$ (not significant), and $p = 0.020$ (significant) for Annotators A, B, C, and Fused Labels, respectively. Table 3.7 shows very strong ROC-AUC performance for all annotators and Fused Labels, with the lowest performance from Annotator C (0.95). However, the ROC-AUC differences were not statistically significant.

Fleiss' Kappa performance metric. We also assessed label quality in terms of inter-rater agreement. This was measured using the Fleiss' Kappa method [74], which determined whether the observed agreement among individual raters was statistically significant or merely due to chance. Fleiss' Kappa was calculated as :

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3.11}$$

where P_o represented the observed label agreement and P_e, represented the label agreement that could be expected by chance. The Fleiss' Kappa value was

0.75 for SVL (considered `substantial agreement') and 0.84 with CVL+RankNet (considered `almost perfect agreement'). Consequently, not only did the binary label quality improve using CVL+RankNet, but so did the agreement between annotators. Both of these are important virtues of CVL+RankNet.

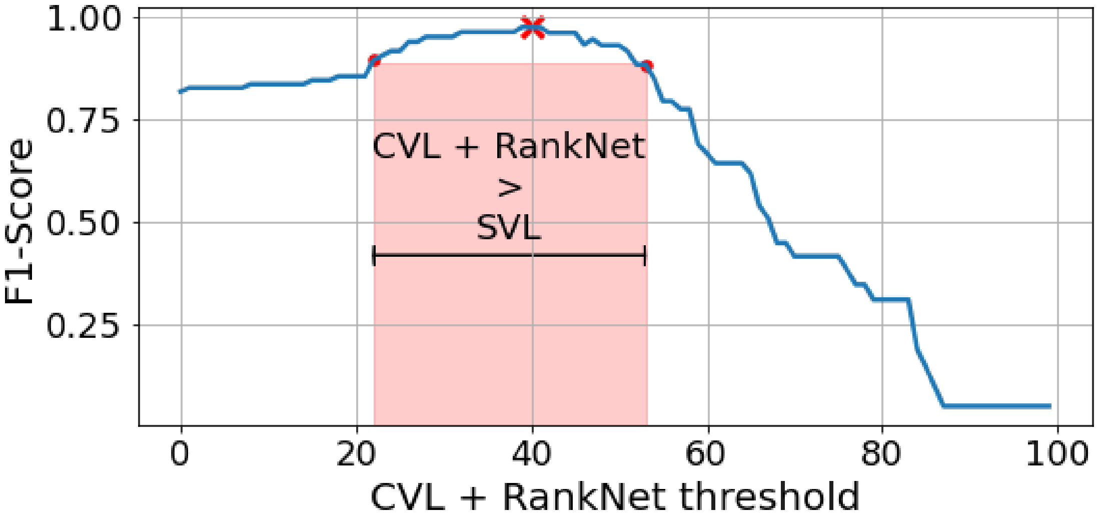


FIGURE 3.13 – The range of CVL+RankNet thresholds yielding a better F1-score compared to SVL. The pink rectangle shows the range, and the red cross shows the CVL+RankNet threshold selected by the annotators).

Decision threshold sensitivity. We also studied how CVL+RankNet performance was affected by the choice of the decision threshold τ. For brevity, we studied this using only Fused Labels, and the results are shown in Figure 3.13. The range of τ yielding a higher F1 score compared to SVL (also using Fused Labels) is shown in the pink rectangle. Considering that the CVL+RankNet scores ranged from 0 (healthy) to 100 (maximum pathology severity), one can see a relatively wide range of thresholds where CVL outperformed SVL. The red cross shows the threshold selected by the annotators, which was very close to the optimum threshold.

RankNet architecture sensitivity. CVL+RankNet performance is affected by choices in the RenkNet architecture, and consequently a performance sensitivity analysis was performed for 3 key hyper-parameters (the number of hidden layers H, the number of neurons per hidden layer K, and the dropout rate d. We evaluated performance using Fused Labels by conducting a hyper-parameter grid sampling, ranging H from $1 \rightarrow 5$, K from $32 \rightarrow 512$, and d from $0 \rightarrow 0.9$). We found that dropout had the strongest influence, but when it was kept below 0.7, the F1-score were relatively stable across all configurations (mean F1 of 0.976 and standard deviation of 0.0089).

Sensitivity to the number of pairwise comparisons A drawback of CVL+RankNet compared to SVL is that it requires more labels (one per image pair, versus one per image). Therefore, we studied the relationship between label quality (measured with the F1 metric) as a function of the average number M of pairwise comparisons per image. We started with the original set of image pairs, and we then pruned image pairs using the method described in Section 3.3.3 until $M = 6$. We then trained the RankNet using the labels of the pruned image pair set and measured its F1 performance. We continued the process of pruning, RankNet training, and performance measurement until $M = 1$, which represented an extremely sparse set of image pairs. To account for the fact that the label pruning process was stochastic due to randomized image pair selection, we repeated the process 10 times, allowing us to measure the distribution of F1 scores for each M. We recall that to measure F1 performance, a decision threshold τ was required at each repetition. It was infeasible to have the annotators manually perform this, so we set the threshold automatically as the one with the maximal F1 score among all possible thresholds via an exhaustive search.

The results are shown in Figure 3.15, which shows four graphs. Each graph shows the F1 distribution as a function of M, for each annotator (graphs (a)-(c)) and with Fused Labels (graph (d)). The horizontal lines in each graph show the F1 performance of the corresponding annotator using SVL. The blue line shows the mean F1 score of CVL+RankNet, and the light blue zone shows one standard deviation from the mean. We notice the following. Firstly, with only 2 pairwise comparisons per image, the average CVL+RankNet F1 scores exceeded the CVL F1 scores for all annotators. Secondly, CVL+RankNet performance appeared to level off at approximately $M = 5.5$. Thirdly, beyond $M = 5.5$, CVL+RankNet performance tended to a similar value in all graphs (F1 score of approximately 0.975). This indicated that CVL+RankNet resulted in highly consistent binary labels, of superior quality to SVL, with only 5.5 comparisons per image.

3.3.4 CVL+RankNet performance for continuous pathology severity assessment

We also investigated the correlation between the raw (non-binarized) CVL+RankNet scores, and PFH values. The results are shown in Figure 3.14 as a scatter plot. Each point represents an image from Dataset 1, with its PFH value on the x-axis and its CVL+RankNet score (computed from Fused Labels) on the y-axis.

The figure shows 4 horizontal bands representing normal and three liver stea-

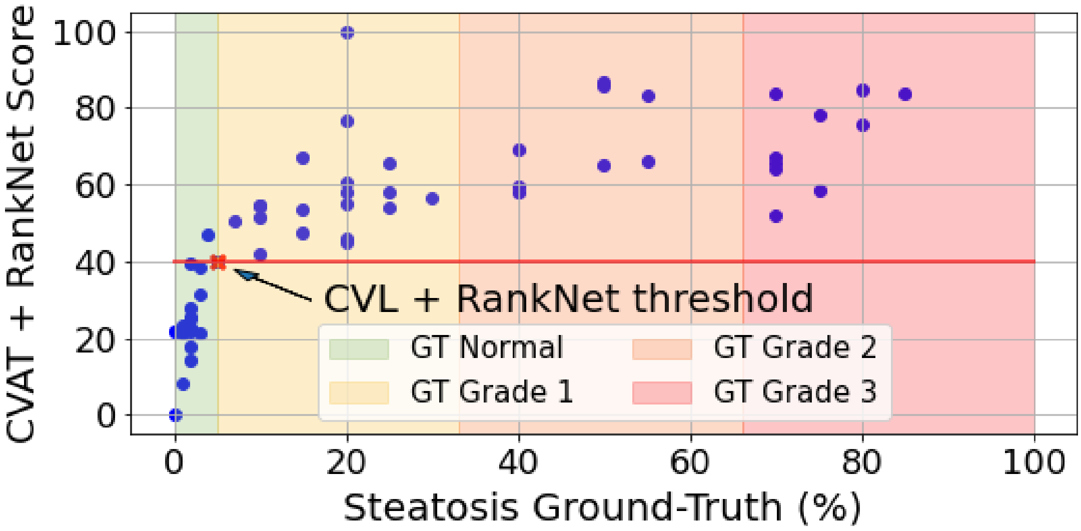


FIGURE 3.14 – Relationship between CVL+RankNet scores computed using fused CVL labels (y-axis), and ground truth PFH values (x-axis). The colored regions indicate the steatosis severity bands : Healthy (green), Grade 1 (yellow), Grade 2 (orange), and Grade 3 (light pink). The red line represents the CVL+RankNet score threshold, selected by the annotators, to separate healthy from pathological images.

tosis severity grades. The horizontal red line shows the CVL+RankNet threshold selected by the annotators (39.78). We observed the following from the graph : 1) CVL+RankNet scores differentiated healthy cases from pathological (Grade 1 or above) with almost perfect separation (1 false positive, 1 false negative, indicated by the red cross. 2) CVL+RankNet scores could not differentiate pathology grades. Nevertheless, in clinical practice, the ability to differentiate healthy patients from those with mild steatosis (grade 1) is substantially more important for disease screening and early detection with ultrasound, compared to differentiating severity grade with ultrasound.

The graph also showed a non-linear relationship between CVL+RankNet scores and PFH values. This was expected because the CVL+RankNet scores were not specifically calibrated [249] against PFH values. Non-linear correlation was assessed with Spearman's rank correlation coefficient ρ [189], measuring the strength and direction of a monotonic relationship between two variables, as follows :

$$\rho = \frac{\sum_i (s_i - \bar{s}) \left(h_i - \bar{h}\right)}{\sqrt{\sum_i (s_i - \bar{s})^2 \sum_i \left(h_i - \bar{h}\right)^2}} \tag{3.12}$$

where s_i represents the CVL+RankNet score for image i, and h_i represents its PFH value. Strong correlation was observed with $\rho = 0.87$ and correlation was highly statistically significant ($p = 7.6e^{-18}$).

3.3.5 Using CVL+RankNet for training liver steatosis detection models

Using Dataset 1, we trained and validated regression and classification models as described in Section 3.2.5.5. Due to the relatively small number of patients in Dataset 1 (55), leave-one-out cross-validation (LOOCV) was used, and performance was evaluated using ROC-AUC.

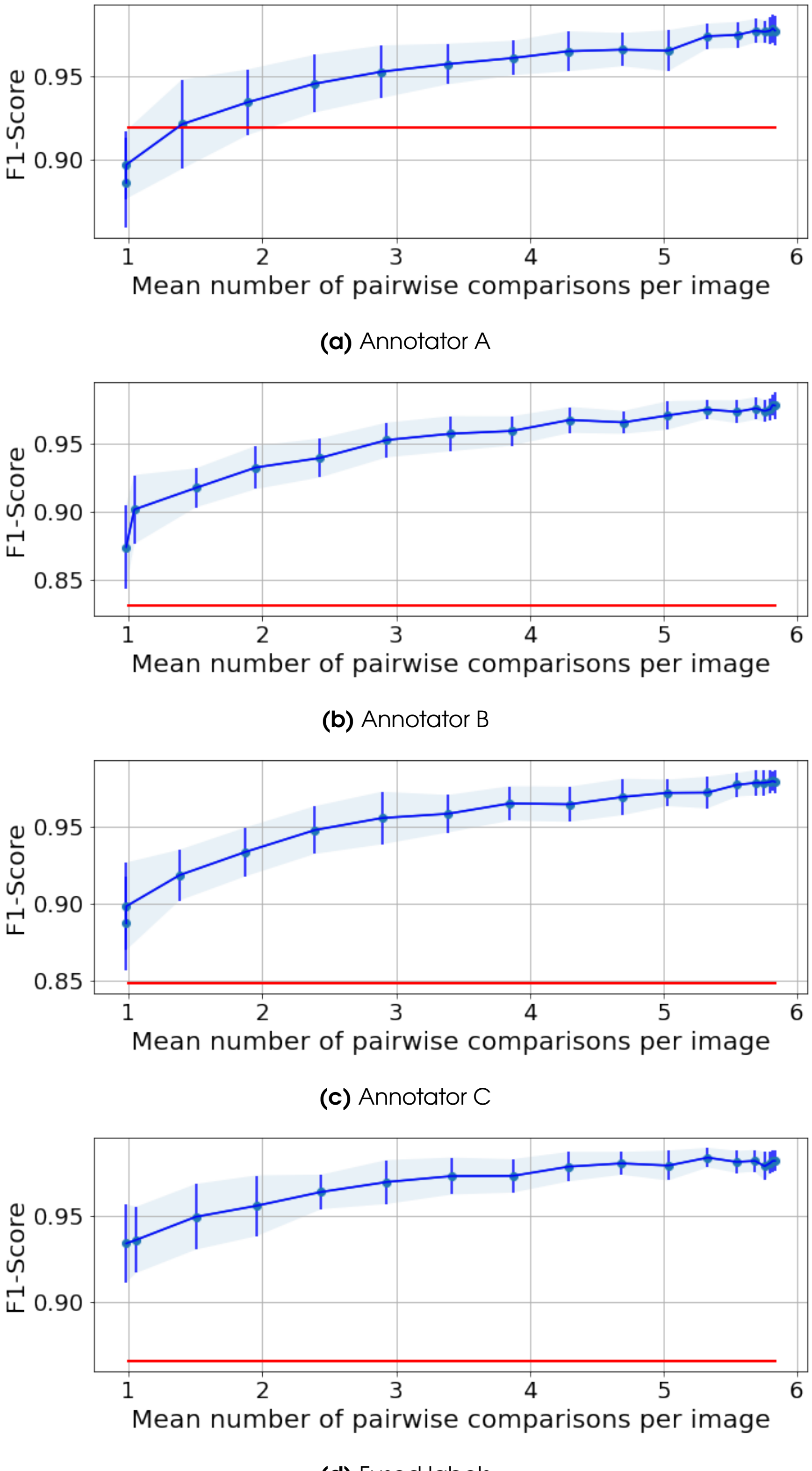
The results are presented in Table 3.8, where each row represents a model configuration and each column represents the source of training labels. There were four configurations : **Classification : SVL** (classification trained with SVL labels), **Classification : CVL+RankNet** (classification trained with CVL+RankNet labels), **Regression : GT** (regression trained with PFH values as labels) and **Regression : CVL+RankNet** (regression trained with CVL+RankNet scores). Consequently, there were 5 sources of labels : (labels from each annotator (CVL and SVL), Fused Labels (CVL and SVL), and PFH values (ground-truth)). For a fair comparison, the same architecture (Inception-ResNet-v2) was used for all configurations, the training parameters were the same as described in Section 3.2.5.5.

From Table 3.8, one can see that all configurations achieved similar ROC-AUC scores. We recall that the CVL+RankNet labels were more accurate than SVL labels as shown previously in Table 3.7. However, that did not translate to significantly better model model performance in terms of ROC-AUC. This indicated that the models had some inherent robustness to training label errors in this task. However, the limited size of Dataset 1 made it difficult to draw firm conclusions about the impact of label errors on model performance, requiring further research. The regression and classification models also performed similarly.

TABLE 3.8 – ROC-AUC scores of classification and regression model configurations. GT represents the histopathological results.

Training labels / Config.	Annot. A	Annot. B	Annot. C	Fused labels.	GT
Classification : SVL	0.93	0.93	0.95	0.93	0.93
Classification : CVL+RankNet	0.92	0.95	0.92	0.91	0.93
Regression : GT	-	-	-	-	0.92
Regression : CVL+RankNet	0.92	0.94	0.89	0.91	0.92

We also performed a cross-dataset performance comparison, using histogram equalization as an image pre-processing step to reduce the domain gap [200].



(a) Annotator A

(b) Annotator B

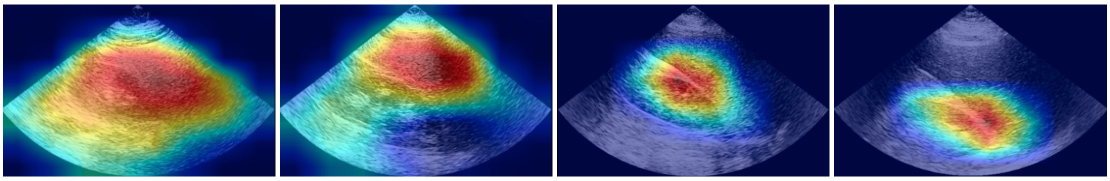
(c) Annotator C

(d) Fused labels

FIGURE 3.15 – F1 Performance of CVL+RankNet (blue curve) as a function of the number of pairwise comparisons. The red line is the reference value for single-image annotations and the blue zone represents 1 standard deviation.

Using Dataset 2, two classifier models were trained using SVL and CVL+RankNet labels. Inference was then performed on Dataset 1, and ROC-AUC was measured using its ground-truth labels. ROC-AUCs were 0.89 (CVL+RankNet) and 0.86 (SVL), and the difference was not statistically significant ($p = 0.34$). This result agreed with the findings described earlier using Dataset 1 and LOOCV.

To verify the models learned appropriate task features, visual explanation maps generated by Grad-CAM [217] were produced for **Classification CVL+RankNet** (Figure 3.16). Pixels with higher influence on the classification decision are illustrated in red (positive influence) and blue (negative influence). The model generally assigned greater influence to the liver parenchyma and the liver/kidney interface, agreeing with the regions that human experts focus on when diagnosing liver steatosis. This suggests that the model effectively learned the task and identified visual features were consistent with experts.



(a) #27 (TP) **(b)** #55 (TP) **(c)** #5 (TN) **(d)** #8 (TN)

FIGURE 3.16 – Representative explanation maps using Grad-CAM [217]. Two true positive (TPs) images are shown, as well as two true negatives (TNs). Pixels with higher influence on the model's output are illustrated in red (positive influence) and blue (negative influence).

3.4 Conclusion and Future Work

In this chapter, we explored the challenges of acquiring high-quality visually-labeled data for training, testing, and monitoring deep learning models in medical image-based diagnosis, which remains a major barrier to the clinical adoption of AI-based CAD systems in ultrasound. We presented a novel and simple labeling technique for diagnostic image labeling using Comparative Visual Labeling (CVL) and RankNet (CVL+RankNet). This method demonstrated a significant enhancement in label quality, particularly for the early detection of steatosis, a critical global health issue, in ultrasound images, compared to Single-image Visual Labeling (CVL). Various benefits of CVL+RankNet have been revealed in our experimental validation, that included very strong ROC-AUC performance and strong correlation with PFH values in healthy and mild steatosis cases, making it not only a method for model training but also, a practical and scalable method to obtain ground-truth labels for model validation, without the need for biopsy sampling and histopathology. We showed that regression and classification models could be trained using these labels, and their performance was very similar to training using PFH values as labels. Additionally, agreement between different annotators was found to be very high using CVL+RankNet compared to SVL using the Fleiss' Kappa metric.

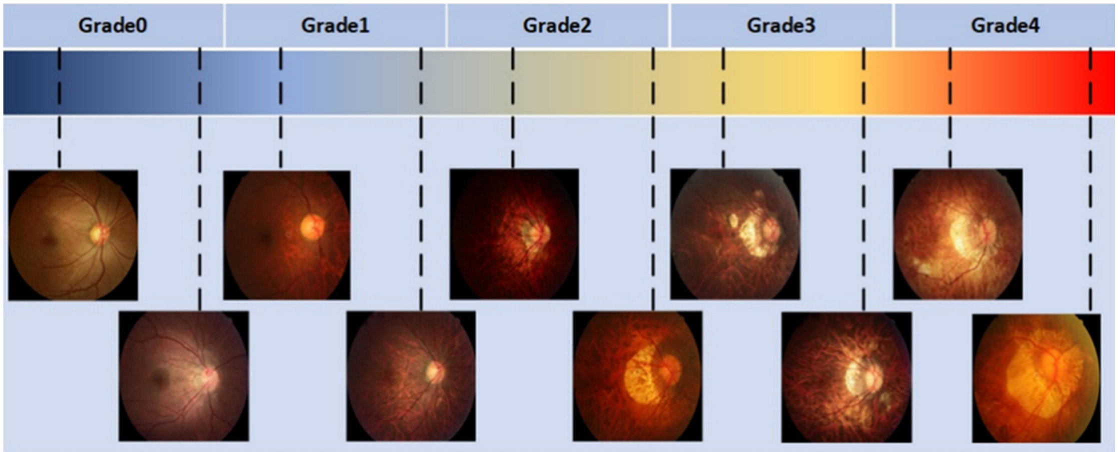


FIGURE 3.17 – Ranking problem in myopic maculopathy diagnosis using retinal fundus images [231]. Figure reproduced from their publication.

This research was published and presented at the 2022 Medical Image Computing and Computer-Assisted Intervention (MICCAI) international conference in Singapore. While we have presented an innovative solution to the challenge of visual labeling for medical image diagnosis, there are several limitations and avenues for future research. The main ones are as follows :

- Validation was limited to a single disease : We evaluated the method only for liver steatosis, but further validation is necessary for other imaging diagnosis tasks and other modalities, including video.
- Increased annotation effort compared to SVL : Despite edge pruning, CVL required more annotations compared to SVL (one per image pair versus one per image).
- CVL+RankNet score calibration : An important future step is to calibrate the scores, especially in the healthy/mild range, with PFH values.
- Liver steatosis detection models performance : Their performance was not shown to be superior when trained using CVL+RankNet scores compared to SVL, despite CVL+RankNet scores having fewer labeling errors.

Considering the first item, some recent articles have built on our research and extended the use of CVL for other pathologies. Specifically, for diagnosing myopic maculopathy in retinal fundus images [178], illustrated in Figure 3.17, and to estimate the severity of ulcerative colitis in endoscopic images [122]. These extensions indeed show great promise for applying the method to other diseases.

[178] also proposed some novel methodology extensions that included leveraging the rank relationship between reference images and new query images through a network with self-attention blocks. Additionally, they implement a loss function in the latent space to align the latent representations with the severity scores of the pathology.

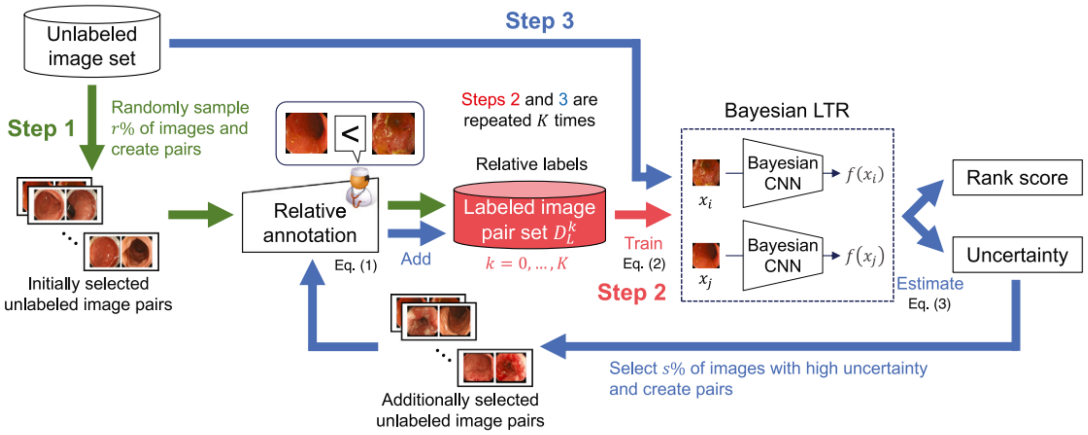

Unlabeled image set
Step 3
Step 1
Randomly sample r% of images and create pairs
<
Steps 2 and 3 are repeated K times
Bayesian LTR
Relative labels
Relative annotation
Labeled image pair set D_L^k
$k = 0, \ldots, K$
Eq. (1)
Add
Train
Eq. (2)
Step 2
Bayesian CNN
x_i
$f(x_i)$
Bayesian CNN
x_j
$f(x_j)$
Rank score
Uncertainty
Estimate
Eq. (3)
Initially selected unlabeled image pairs
Additionally selected unlabeled image pairs
Select s% of images with high uncertainty and create pairs

Figure 3.18 – Deep Bayesian active learning-to-rank for relative severity estimation ; step 1 (green arrows) : generating a small number of pairs using randomly selected images from an unlabeled image set and annotating these pairs for the initial training ; step 2 (red arrow) : training the Bayesian CNN using the labeled image pair set ; step 3 (blue arrows) : selecting high-uncertainty images from the unlabeled image set to create pairs and attaching relative labels to the pairs. Figure and description extracted from [122].

Considering the second item, [122] employed an active learning framework with Monte Carlo Dropout (MCD) to select image pairs for annotation. A Bayesian CNN was trained using CVL with a loss function similar to the one described in Equation 3.5. Once trained, the network was applied to individual images to estimate uncertainty, which was then used to generate new image pairs for annotation. The process iterated to progressively refine the model's performance. Figure 3.18 illustrates the authors' scheme for these steps from [122].

Considering the third item, model calibration would be an essential next step in using CVL+RankNet as an alternative method to quantify fat cell percentage compared to biopsy and histopathology. While this has huge potential as a clinical tool in its own right, accurate calibration must be performed, which may potentially be achieved using a simple regression approach.

In conclusion, by introducing CVL, and a way to convert its labels to continuous pathology severity scores with RankNet, we have shown a promising approach for enhancing label quality in early disease detection. This was demonstrated in this chapter using liver steatosis, and extended to other diseases in follow-up works.

We believe the advancements discussed here will contribute meaningfully to the ongoing development and evaluation of CAD models using visual labeling.

4. DR-Clips : A novel frame-guidance approach for computer-assisted diagnosis with untrimmed ultrasound videos

4.1 Chapter summary

In the previous chapter, we addressed a key challenge in ultrasound CAD : enhancing the quality and objectivity of visual diagnostic labels. In this chapter, we focus on another critical challenge : training video classification models using untrimmed videos paired with video-level diagnostic labels. Unlike image-level labels, which correspond to information in a specific image or video frame, video-level labels describe the overall content of a video and are not aligned with individual frames. Unlike trimmed videos, which are manually edited to isolate specific segments displaying pathology, untrimmed videos retain their full duration, including segments with limited or no diagnostic relevance.

Video classifiers designed for untrimmed videos offer several notable advantages. These include reducing annotation effort, utilizing more diagnostically relevant information from training and testing videos, enabling cross-modal training labels, and supporting real-time inference without requiring clinicians to manually select relevant segments. While state-of-the-art video classifiers are theoretically capable of processing untrimmed ultrasound videos with video-level labels, they are highly susceptible to overfitting and face significant challenges in generalizing effectively, particularly when applied to the small datasets commonly encountered in CAD research.

To address these limitations, this chapter introduces DR-Clips, a novel methodology that enhances the performance and generalization of ultrasound video classifiers trained on untrimmed videos with video-level labels. At the core of DR-Clips is a neural network called the Frame Relevance Assessor (FRA), which automatically identifies a set of diagnostically relevant frames and sorts them in relevance order, forming what we term a *diagnostically relevant clip* (DR-Clip). Importantly, the FRA does not need to be perfect, and to achieve robustness and generalization, we train the video classifier using randomized DR-Clip selection. This approach significantly improves performance, demonstrating the effectiveness of DR-Clips in overcoming the challenges associated with untrimmed videos and small datasets in CAD research.

We apply DR-Clips to liver and kidney pathology classification in abdominal ultrasound. However, a major challenge in this area is the lack of publicly available video datasets. To address this, we describe a new dataset that extends Dataset 2 from the previous chapter to include video data and video-level labels. The methods and results presented in this chapter using this dataset have been submitted to the International Journal of Computer-Assisted Radiology and are currently under major revision.

4.2 Additional background in pathology classification with abdominal ultrasound

4.2.1 Benefits of video classification

In computer-assisted diagnosis (CAD) with abdominal ultrasound data, three main categories of pathology classification models have been proposed :

1. *Supervised image classification* : Models are trained to perform pathology classification from single images, using training datasets comprising images and image-level labels. Labels relate to the visual content of each image.
2. *Supervised trimmed video classification* : Models are trained to perform pathology classification from trimmed videos (sections of ultrasound videos that have been manually cropped to moments that confirm the presence or absence of a pathology). They are trained on datasets comprising trimmed videos and video-clip-level labels. Labels relate to the visual content of each trimmed video.
3. *Supervised untrimmed video classification* : Models are trained to perform pathology classification from untrimmed videos (ultrasound videos that have not been manually cropped to diagnostically relevant segments). They are trained on datasets comprising untrimmed videos and video-level labels. Labels relate to the visual content of an entire video.

Most previous approaches in abdominal ultrasound CAD used category 1) (image classification models trained with image-level labels) [29, 237, 213, 53, 267, 50, 269]. These models were predominantly designed for liver pathology classification, especially for tasks such as liver steatosis detection [29, 237, 213] and focal liver lesion analysis [53, 267]. Models have also been developed for kidney pathology classification, including detecting chronic kidney disease and other renal pathologies such as cystic or obstructive nephropathy [229, 138, 294, 221, 177]. Other abdominal organs such as the gallbladder and spleen have been explored to differentiate of pathological conditions like cholecystitis for the gallbladder and splenic abnormalities [116, 117, 245, 157].

The above works depended on manually curated training and test datasets of high-quality, diagnostically relevant images selected by experts. For example, in liver steatosis classification [29], the datasets were composed of images captured in standardized (canonical) views, ensuring clear visualization of both the liver and kidney parenchyma. However, this dependency has four key limitations concerning (*i*) annotation effort, (*ii*) domain gap, (*iii*) limited exploitation of video data, and (*iv*) label alignment. We discuss each limitation below and how video classifiers trained on video-level labels can significantly help. Figure 4.3 illustrates video classification at a high level.

Annotation effort. Creating ultrasound datasets can be accomplished either prospectively or retrospectively. In prospective data collection, operators acquire data specifically for annotation and model development purposes. This approach typically yields higher-quality data that may be easier to annotate ; however, it requires adherence to a specific acquisition protocol, which can be challenging to implement in clinical settings. In contrast, retrospective data collection uses data recorded during routine clinical practice, which is later analyzed and annotated. The primary advantage of retrospective collection is its scalability. However, it often results in large volumes of frames with limited or no diagnostic value, as illustrated in Figure 4.2. Maneuvering the proof to obtain clear imaging windows to either confirm the presence or absence of pathology is challenging due to the high variance in patient anatomy and varying degrees of operator competence. This leads to many frames with limited or no diagnostic information. Consequently, only a few frames may be diagnostically relevant in retrospectively collected videos, leading to a needle-in-a-haystack problem. To generate image-level labels from retrospective data, it is necessary to manually sift through the data to identify diagnostically relevant frames—a time-intensive process.

In contrast, video-level labels significantly reduce this annotation burden by eliminating the need for manual frame selection. Annotators no longer need to evaluate individual frames to confirm diagnostic relevance or ensure that selected frames adequately capture intra-patient variability (e.g., probe positioning, breathing states, or applied pressure). By streamlining such decisions, which are often subjective and nuanced, video-level labeling provides a more efficient and practical alternative.

Domain gap. When building single-image datasets from prospectively or retrospectively collected data, a significant domain gap often exists between the curated image distribution and those captured during live ultrasound procedures. For instance, a single-image steatosis detector trained on Dataset 1 [29] might fail when applied to images lacking clear kidney visibility, as it was exclusively trained on liver-kidney plane images. This domain gap arises because curated datasets, typically composed of diagnostic-quality images, fail to capture the variability and challenges encountered in real clinical practice. As a result, this mismatch can lead to out-of-domain errors, where a classifier makes incorrect predictions with high confidence on unseen or mismatched data [272].

In contrast, video datasets naturally capture a broader variability that better aligns with live procedure videos, potentially mitigating out-of-domain errors. Video data enables models to learn from diverse contexts and conditions, improving their robustness and real-world applicability.

Label alignment. Image-level labels are inherently tied to the specific diagnostic indications present in individual frames. In contrast, diagnostic labels derived from other modalities (e.g., histopathology, laboratory results, or electronic health records) typically pertain to a patient or case as a whole, rather than to specific frames in an ultrasound video. Training single-image classifiers with these inter-modal labels requires labor-intensive manual selection of frames to align the labels accurately. By treating inter-modal labels as video-level labels, models can be trained directly on entire ultrasound videos, bypassing the need for manual frame selection and label alignment, while preserving the diagnostic information embedded across the video.

Exploitation of video Data. Image classifiers are inherently limited to the information present in the individual frames used for training, often missing the broader diagnostic context. In contrast, video classifiers can exploit the full temporal and spatial data present in videos. By aggregating information across frames, video-level models can account for temporal dynamics and contextual details, which may improve diagnostic performance by incorporating insights that single-image models are unable to capture.

4.2.2 Benefits of untrimmed video classification

Trimmed videos involve selecting diagnostically relevant time windows, which offers improvements over single-frame selection. For instance, it reduces the annotation burden by removing the need to select individual frames precisely and allowing for more video data and temporal context to be exploited. However, it introduces new challenges and does not fully address the above limitations.

Trimming videos still requires manual effort and subjective judgment to identify relevant time windows. Additionally, focusing exclusively on diagnostically relevant segments creates a domain gap, as models trained on trimmed videos may struggle to reason about non-relevant segments or contextual information present in the full video.

In contrast, leveraging untrimmed video data bypasses the need for manual clip selection. Instead, the model is trained to automatically identify and prioritize diagnostically relevant regions within the entire video. This approach captures the natural variability of ultrasound data and eliminates the reliance on manual video trimming, potentially providing a more robust, scalable solution for real-world applications by enabling the model to learn from both relevant and non-relevant video segments.

4.2.3 Relevant prior art in trimmed and untrimmed ultrasound video classification

Various ultrasound video classification models have been investigated for numerous clinical applications [68, 199, 173, 172, 21, 77, 17, 186, 258, 230, 270].

The prior work on ultrasound video classification can be divided along three axes : 1) clinical tasks, 2) model architecture, and 3) the method used (if any) to handle untrimmed videos.

Clinical tasks. The majority of works have focused on cardiac and lung diseases, which were fundamentally stimulated by two recent public datasets : the EchoNet-Dynamic Dataset [182] for cardiac diagnosis, and lung datasets [21, 207] for COVID diagnosis. A range of papers apply and adapt existing video classifiers to ultrasound video classification using these datasets [68, 199, 173, 95, 118, 102, 172, 21, 77, 17, 151, 67]. Ultrasound video classification has also been explored for fetal biometry [188, 186, 187, 193], thyroid nodule analysis [258], breast lesions detection [230, 108] and liver lesion detection [270].

Model architectures. Various model architectures have been explored in the above papers, either as direct applications from the general computer vision literature or as adaptations tailored to specific use cases. These include Graph Neural Networks [173], 3D Convolutional Neural Networks (CNNs) [77], 2D CNNs combined with Long Short-Term Memory (LSTM) to capture temporal context [17], and CNNs-based feature extraction combined with transformers to also capture temporal context [199, 186]. Vision transformers have also been considered, especially Video Swin Transformers [68]. Currently, there is no consensus on the best architecture for ultrasound video classification.

Method to handle untrimmed videos. Only three of the above methods have been demonstrated to work on untrimmed videos [270, 258, 77]. They all use a mechanism to filter out non-diagnostically relevant frames called a `guidance algorithm.' The guidance algorithms were different and trained with varying degrees of human supervision. We go into further details of these methods and their limitations, as they represent the closest research to our proposed methodology.

Xu et al. [270] implemented a frame-guidance algorithm for focal liver lesion detection and malignancy/benign classification. A three-step cascaded pipeline was proposed, where each step filtered out non-relevant ultrasound images from the previous step. In step 1, each image was automatically segmented into three regions : ultrasound liver pixels, non-ultrasound pixels, and `other'. Existing segmentation models were evaluated for this task, including FCN, U-Net, and DeepLabV3, and trained using ground-truth segmentation masks from radiologists. In step 2, the segmented images containing segmented liver pixels were passed to a 2D image classifier (DenseNet121), which was trained to distinguish between liver images containing focal liver masses and those without masses. In step 3, images with detected masses were passed to another segmentation network that segmented the mass. Finally, images with segmented masses from step 3 were passed to a video classification network (LMC-Net) to distinguish malignant from benign masses. The LMC-Net was fed the images, the segmentation masks, and (optionally) other clinical variables from the patient's electronic health record. The results in Xu et al. [270] demonstrated significant improvements in focal liver lesion detection and malignancy classification compared to prior methods, making a compelling case for frame-guided inference. However, the pipeline was complex, requiring multiple levels, and it was trained using pixel-level (segmentation) and image-level labels. Furthermore, a limitation of a cascaded pipeline like this is its vulnerability to errors at each step. The weakest link in the sequence constrains the overall performance. For example, if liver segmentation failed in step 1 due to low ultrasound image contrast, subsequent steps could not compensate for this failure.

Wang et al. [258] implemented frame guidance in a two-step approach for thyroid nodule detection and malignancy/benign classification. In the first step, a 2D object detector (Faster-RCNN) was trained to detect thyroid nodules in 2D ultrasound images using bounding-box labels. This was trained on images containing thyroid nodules, selected by radiologists. In the second step, images with detected nodules were inputted into an LSTM-based model and trained to classify those that were highly diagnostically relevant (`keyframe images'). Finally, a temporal window of 32 frames was extracted about each detected keyframe, and passed to a 32-frame video classification model (a simplified C3D CNN [246]), trained to classify malignant versus benign nodules. The labels for the C3D CNN came from the patient's clinical diagnosis in their electronic health record. The results of [258] were encouraging, however, similarly to [270], frame-guidance was implemented with multiple stages and trained with multiple levels of supervision : bounding-box level labels for mass detection, and image-level labels for keyframe selection.

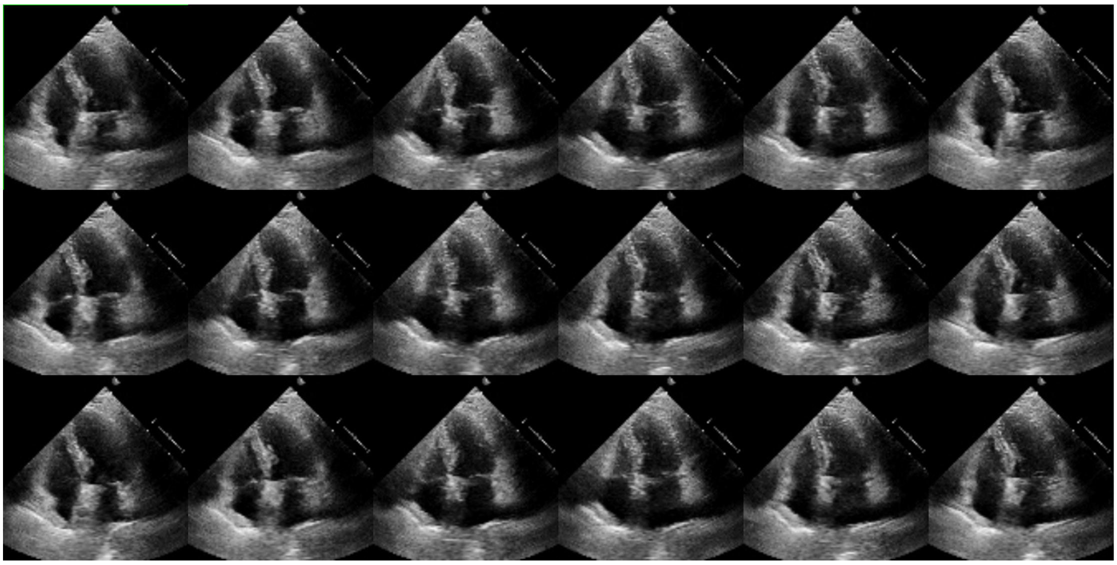


FIGURE 4.1 – Example of a trimmed ultrasound video from the EchoNet-Dynamic Dataset [182], illustrating a sequence where most frames are relevant for the diagnostic task. Frames are sampled at a 6-frame interval from video '0X1A0A263B22CCD966.avi' and resized to 256x256 using OpenCV [22].

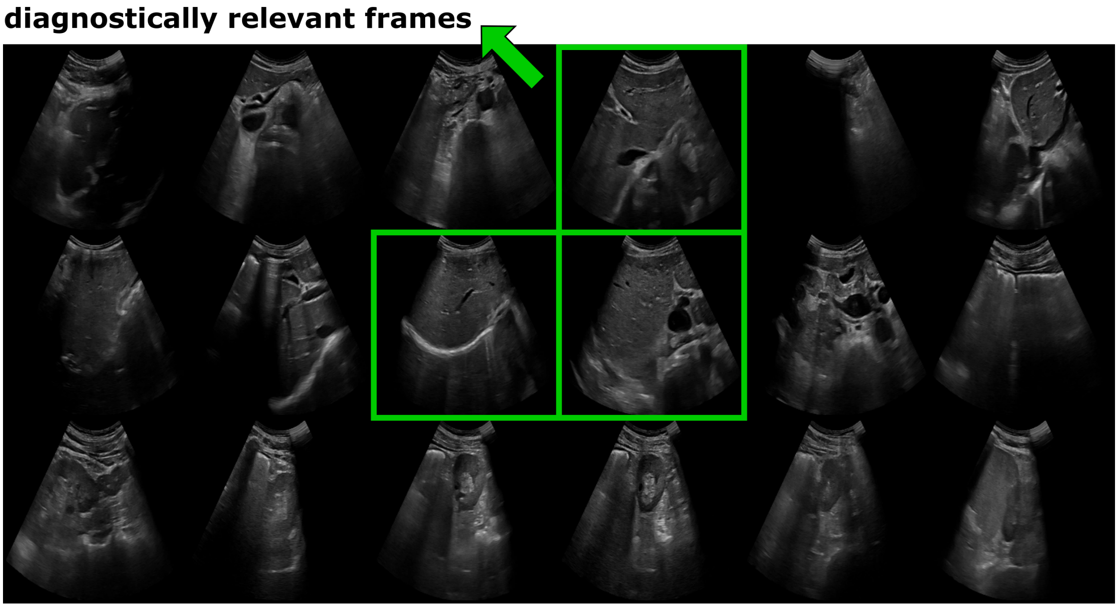


FIGURE 4.2 – Example of an untrimmed ultrasound video from an abdominal screening from our 'MIM-US-107 Video Dataset' (resized to 256x256 using OpenCV [22]), highlighting the challenge of non-relevant frames. In this sequence, only a few frames are pertinent to diagnosing liver pathologies, while the majority are irrelevant.

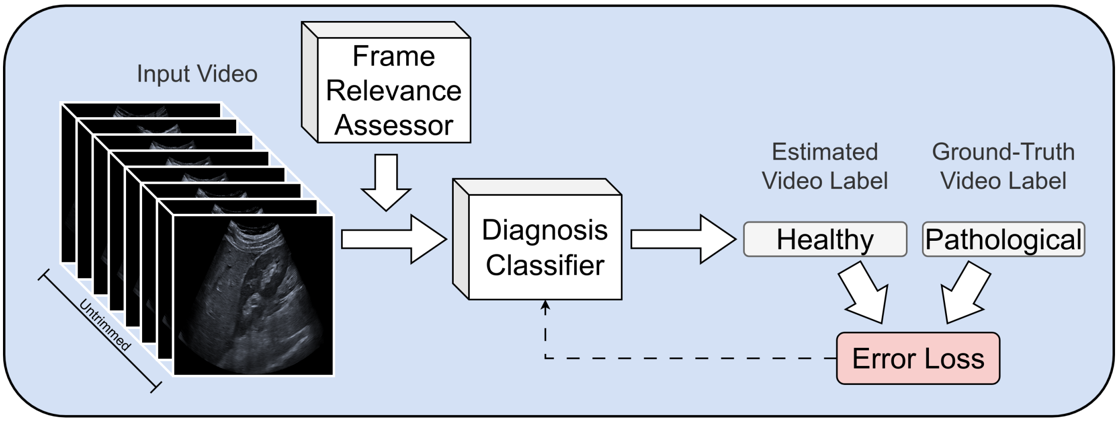


FIGURE 4.3 – Diagram illustrating the integration of a Frame Relevance Assessor (FRA) into video classification for ultrasound (US) CAD. The FRA evaluates the diagnostic relevance of each frame in untrimmed video data, guiding the video classification model during training and inference.

4.3 Methodology

4.3.1 Overview of DR-Clips

The key challenge addressed by the proposed methodology is adapting state-of-the-art video classifiers to handle untrimmed ultrasound training videos, supporting inference on untrimmed videos, and reducing the supervision required for frame guidance.

Figure 4.3 provides a high-level overview of our method, which integrates two deep neural networks : the Frame Relevance Assessor (FRA) and a Diagnostic Classifier. The FRA is an image regressor that predicts the diagnostic relevance of individual ultrasound frames, assigning scores on a scale from 0 (not relevant) to 1 (highly relevant).

The Diagnostic Classifier functions as a video classifier, processing a set of frames identified as diagnostically relevant by the FRA. These frames are provided to the Diagnostic Classifier as a DR-Clip, where the frames are in ordered by relevance, enabling the Diagnostic Classifier to focus on diagnosis rather than simultaneously handling diagnosis and frame relevance evaluation. The Diagnostic Classifier is trained using video-level labels, which can be obtained through visual annotation or cross-modal sources. In this chapter, we implement DR-Clips for binary diagnostic tasks (distinguishing between healthy and pathological cases). However, the method is flexible and can be extended to more complex scenarios, such as employing a multi-class Diagnostic Classifier to grade disease severity or a multi-label Diagnostic Classifier to detect multiple pathologies within a single model. Our approach employs a single frame-guidance step via the FRA, resulting in a streamlined pipeline that can be implemented using most state-of-the-art video classifier models for the Diagnostic Classifier and state-of-the-art image regression models for the FRA.

The concept of relevance in ultrasound CAD is complex and multifaceted. It encompasses factors such as image quality, probe position, presence of artifacts, organ visibility, and pathology indicators. Pathology indicators can be direct (e.g., the visibility of a solid mass) or indirect (e.g., dilation of the intrahepatic portal vein, indicative of cirrhosis). However, relevance is not limited to positive diagnoses ; it also includes factors that support a negative (healthy) diagnosis. Importantly, these may not simply be the absence of pathology indicators. For example, the absence of ultrasound signal attenuation in the liver does not necessarily indicate a healthy liver. Poor visibility due to factors like visceral fat can obscure the organ and confound diagnosis, even in the absence of pathology.

Designing the Frame Relevance Assessor (FRA) presents a critical trade-off. On one hand, a high-precision FRA aims to select only the most diagnostically relevant frames, reducing the burden on the downstream video classifier (which we recall, processes only the frames selected by the FRA). On the other hand, a lower-precision FRA requires less supervised training but relies on a video classifier robust enough to handle DR-Clips that include irrelevant frames. We hypothesize that this trade-off can be mitigated by pairing a lower-precision FRA with a robust video classifier trained to tolerate irrelevant frames within DR-Clips, thereby achieving strong overall diagnostic performance.

To achieve this, we design the FRA to evaluate general-purpose features such as image quality, probe position, presence of artifacts, and organ visibility. These attributes are common across many abdominal pathologies, allowing the FRA to generalize effectively. This approach minimizes the need for extensive pathology-specific annotations and may avoid retraining the FRA when the system is expanded to include new pathologies. The details of how FRA training data was acquired and labeled are provided later in Section 4.3.4.3.

4.3.2 Model training

The FRA is implemented as a standard image regression deep neural network, trained using supervised learning on a dataset comprising ultrasound images and associated relevance labels. In contrast, the Diagnostic Classifier uses a novel training methodology described in this section and summarised in Figure 4.4(a).

Training the Diagnostic Classifier starts with a trained FRA and a dataset of V untrimmed training videos with associated video-level diagnostic labels. There are two main steps : 1) DR-Clip generation, which extracts DR-Clips from the training videos, and 2) Model training, where the Diagnostic Classifier is trained using DR-Clips and the video labels associated with each DR-Clip. We now present these steps using italic fonts method hyper-parameters, including their default values fixed in all experiments.

The Diagnostic Classifier can be implemented with a state-of-the-art video classification deep neural network, such as Video Swin Transformer [163], without modification to its architecture.

4.3.2.1 DR-Clip generation

First, the frames in each training video are passed through the FRA, generating per-frame diagnostic relevance scores $\hat{r}_k^v \in [0, 1]$ where v indexes over training videos and k indexes over the video's frames. A set of *L=500* DR-Clips are then randomly sampled from each training video, denoted as the set $\mathbf{C}^v = \{C_1^v, C_2^v, \dots C_L^v\}$. Each member of $\mathbf{C}^v$ is a DR-Clip, comprising a set of frames drawn from the v^{th} video, sorted in descending frame relevance score. Each DR-Clips C_k^v has a random length N_k^v drawn uniformly from the range of $N_{min} = 1$ to $N_{max} = 32$. This variability makes the network robust to clips of different lengths at inference time. The frames in each DR-Clip are randomly selected with uniform probability and without replacement. We do not filter out low-relevance frames as a pre-processing step before assembling DR-Clips. This is to ensure the video classifier is trained on DR-Clips with variations in frame relevance to tolerate imperfect relevance predictions from the FRA. We define as y_v the diagnostic label of the v^{th} video. Each DR-Clip $C_{k\in[1,2,\dots L]}^v$ shares the same diagnostic label of the video from which it is generated.

We then assign a weight $w_k^v \in \mathbb{R}^+$ to each DR-Clip, used to prune the inital pool of DR-Clips before training the Diagnostic Classifier. It is also used in the Diagnostic Classifier's loss function. The weight is computed in two steps; First, the relevance scores of the frames within the DR-Clip are normalized using a Shifted Sigmoid function $\sigma(\cdot, a, b)$ with parameters *a* $= 10$ and *b* $= 0.5$ as

$$\sigma(x, a, b) = \frac{1}{1 + e^{-a(x-b)}} \tag{4.1}$$

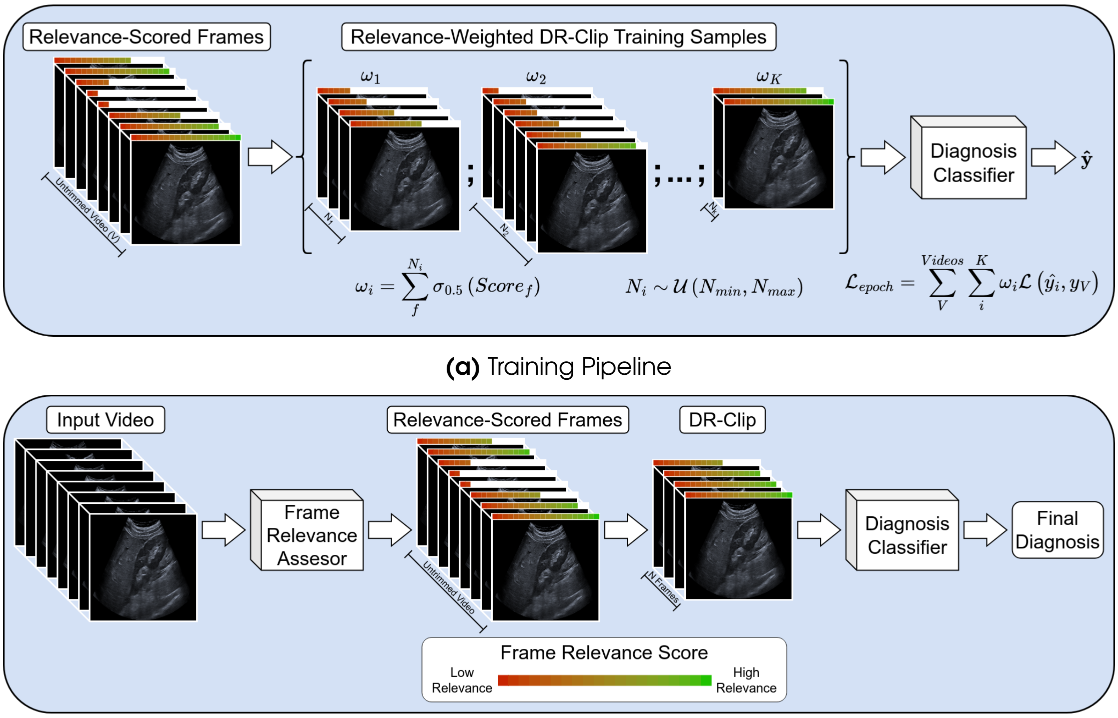
The weight of the DR-Clip is then calculated as the average of its normalized relevance scores :

$$w_k^v = \frac{1}{N_j^v} \sum_{i=1}^{N_j^v} \sigma(\hat{r}_{i,j}^v, a, b) \tag{4.2}$$

where $\hat{r}_{i,j}^v$ denotes the estimated relevance of the i^{th} frame in the k^{th} DR-Clip generated from the v^{th} training video. The purpose of the normalization step is to attribute higher weights to DR-Clips that contain highly relevant frames, even if there are only a few such frames in the DR-Clip.

The final step in DR-Clip generation is clip pruning, which removes DR-Clips with low weights. This step is important to address the sparsity of diagnostically relevant frames, a key challenge in untrimmed training video data. Without pruning, the training dataset may become cluttered with an abundance of irrelevant DR-Clips. This can undermine the training process for the Diagnostic Classifier in two significant ways. First, DR-Clips that contain no relevant frames are inherently unclassifiable, providing no meaningful signal for the classifier to learn. Second, forcing the model to process such clips may divert its focus from learning effectively from meaningful DR-Clips. Clip pruning mitigates this by ensuring that training is concentrated on DR-Clips with relevance.

DR-Clip pruning can be implemented in various ways. In our experiments, we adopt *top-k pruning*, where, for each video, only the DR-Clips with the top K



(a) Training Pipeline

(b) Inference Pipeline

FIGURE 4.4 – Proposed pipeline : Video classifiers for US CAD using Diagnostically-Relevant Clips (DR-Clips).

weights are retained for training. The choice of K involves a critical trade-off. If K is set too low, intra-video variability may be reduced, which could lead to model overfitting and increase reliance on the FRA's recall. Conversely, if K is set too high, the dataset may include too many irrelevant DR-Clips, diluting the quality of the training data. To balance these considerations, we use $K = 10$ as the default in our experiments, resulting in KV DR-Clips available for training.

4.3.2.2 Diagnostic Classifier training

The Diagnostic Classifier is implemented as a DR-Clip video classifier. For binary classification tasks, the binary cross-entropy loss can be used. Drawing inspiration from Curriculum Learning [257], we propose to adapt it by weighting the loss contribution from each training DR-Clip by its average relevance, w_k^v. Curriculum Learning involves training a model on examples in a structured order, usually starting with simpler cases and gradually progressing to more complex ones. This approach has been shown to enhance learning efficiency, improve generalization, and accelerate convergence [257].

In our context, the weight w_k^v, derived from FRA scores, reflects the ease of classifying a DR-Clip, with higher weights assigned to simpler DR-Clips. By weighting the loss of a DR-Clip with w_k^v, we bias the learning process toward focusing on easier DR-Clips during the early stages of training. The resulting loss function called the *DR-Clip loss* is written as :

$$\mathcal{L}_{\text{DR-Clip}}(\Theta) = \sum_{v=1}^{V}\sum_{k=1}^{K} w_k^v \mathcal{L}\left(\Theta(C_v^j), y_v\right) \tag{4.3}$$

where $\Theta(C_v^j)$ denotes the models classification prediction for the DR-Clip C_v^j.

Most video classifier models, such as a Video Swin Transformer [163], require input videos to have a fixed number of frames. To this end, as a pre-processing step, zero padding is applied to the end of any DR-Clip with fewer than N_{max} frames. Zero padding is applied at the end since, by definition, they are the least relevant frames in the DR-Clip (containing no visual information). We go into specific configurations and training details of the Video Swin Transformer used in our experiments in Section 4.3.4.4.

4.3.3 Model inference

Inference consists of two steps, as shown in Figure 4.4(b). First, each frame in the test video is processed by the FRA, giving a set of frames and their associated relevance scores : $\mathcal{I} = \{(I_1, \hat{r}_1), (I_2, \hat{r}_2), \ldots, (I_M, \hat{r}_M)\}$ where M is the number of frames in the test video. Then the top-N most relevant frames are selected to form an *inference DR-Clip* $\mathcal{C}_{infer}$, as :

$$\mathcal{C}_{infer} = \{I_k \mid k \in \operatorname{argsort}_{i=1}^{M}(r_i)[1:N]\}, \tag{4.4}$$

where $\operatorname{argsort}_{i=1}^{M}(r_i)[1:N]$ represents the indices of the N highest relevance scores sorted in descending order. Next, we pass $\mathcal{C}_{infer}$ through the Diagnostic Classifier to generate the diagnostic prediction in a single forward pass.

$$\hat{y_v} = \Theta\left(C_v^{inf}\right) \tag{4.5}$$

The method can be applied to two kinds of test videos. The first kind (batch) is when the video has a fixed duration. The second kind (online) is for real-time inference with live video streamed from an ultrasound device, where M increases over time. In this scenario, $\mathcal{C}_{infer}$ functions as a buffer that stores the most relevant frames encountered up to the current moment. Since $\mathcal{C}_{infer}$ has a fixed size, the inference time of the video classifier remains constant regardless of the stream's duration

4.3.3.1 Combining the FRA with an image classifier

We also present a variant of the approach using an image classifier as the Diagnostic Classifier, such as Inception-ResNet-v2 [233]. In this setup, the Diagnostic Classifier performs classification on individual images, rather than on DR-Clips. We propose this variant for two main reasons. Firstly, to investigate whether the FRA can be used to train an image-level diagnostic classifier on untrimmed training videos, using frames automatically extracted by the FRA. Secondly, to compare its performance against a video classifier using DR-Clips, as described above.

In the training step, the image classifier is trained on all images whose relevance, as predicted by the FRA, exceeds a threshold $\tau_r = 0.6$.

During inference, each frame in a test video is first processed through the FRA. Frames with a relevance score exceeding the threshold τ_r are passed to the classifier. The resulting frame-level predictions are then aggregated using one of two fusion methods :

- *Max Fusion* : The final prediction is the classification label with the highest confidence score among all selected frames.
- *Mean Fusion* : For each label, the mean confidence score across all selected frames is calculated, and the label with the highest mean confidence is chosen as the final prediction.

Any image classification neural network architecture can be used, and we present our choice and implementation details used in our experiments in Section 4.3.4.4.

4.3.4 Abdominal ultrasound datasets and model training details

4.3.4.1 Datasets

This study used anonymized US B-mode abdominal data obtained retrospectively from the Saint-Anne MIM Clinic (Strasbourg, France) during routine US abdominal examinations from January 2022 to January 2023. All patients who underwent abdominal ultrasound examinations as part of their routine care were included. The acquisitions were acquired using a Canon Aplio a450 device and collected with written informed patient consent. The data comprised two categories - video recordings and still images. We constructed two datasets for each category :

The MIM-US-107 Video Dataset

Source data. This dataset contained 107 abdominal B-mode ultrasound videos, with one video per patient. Videos that included Contrast-Enhanced Ultrasound (CEUS) or Doppler imaging where excluded. The clinician typically recorded more than one video throughout their procedure (an average of 6.6 videos per patient, standard deviation : 3.5). We concatenated all patient videos to form a single video file per patient. The average number of video frames per patient was 339.8 (standard deviation : 182.2). The videos were annotated as described in

Annotation. The videos were annotated by the same team of three radiographers (Annotators A, B, and C) from the previous chapter, each with 10-17 years of experience in ultrasound image analysis. Annotations were performed using the CVAT annotation platform [216]. An independent radiographer reviewed the annotations, and any disagreements were discussed to reach an annotation consensus.

The radiographers carefully inspected each video and assigned video-level labels according to established liver, kidney, or biliary system pathologies that are detectable in b-mode ultrasound. In total fourteen labels were used in this dataset :

- Six labels were used for positive liver pathology findings : Liver steatosis, Liver solid masses, Liver cystic masses, Liver metastases, Liver fibrosis, and Hepatomegaly.
- Four labels were used for positive kidney pathology findings : Kidney cystic mass, Hydronephrosis, Chronic kidney disease, and Nephrolithiasis.
- Two labels were used for positive biliary system pathology findings : Gallstones (cholelithiasis) and Bile duct dilation.
- Two labels (Healthy liver and, Healthy kidney) —were used when the liver or kidney appeared healthy in the video, respectively.

The positive pathology labels were not mutually exclusive (since a patient could have multiple pathologies).

Whenever one of the above labels was assigned, the radiographer was also instructed to select up to five keyframes per video, corresponding to images they considered highly relevant in their decision to assign a label. Radiographers were asked to supply diverse keyframes where possible. We used these keyframes to compare baseline methods requiring image-level supervision. They were not used to train our proposed method.

Task definitions. Figure 4.5 provides the frequency of each label, and Figure 4.6 shows representative keyframes of each label. Due to the limited dataset size, as shown in Figure 4.5, most pathologies appeared in only a few videos. As such, training models to detect each pathology as a separate class was not feasible. Instead, we combined labels to generate two diagnostic tasks :

- The **Liver Task**, which involved video classification with two labels : healthy liver versus liver damage (steatosis or fibrosis), which is a key indicator of underlying conditions, especially nonalcoholic fatty liver disease (NAFLD) or viral hepatitis.
- The **Kidney Task**, which involved video classification with two labels : healthy kidney versus renal structural abnormality (kidney cystic masses or hydronephrosis).

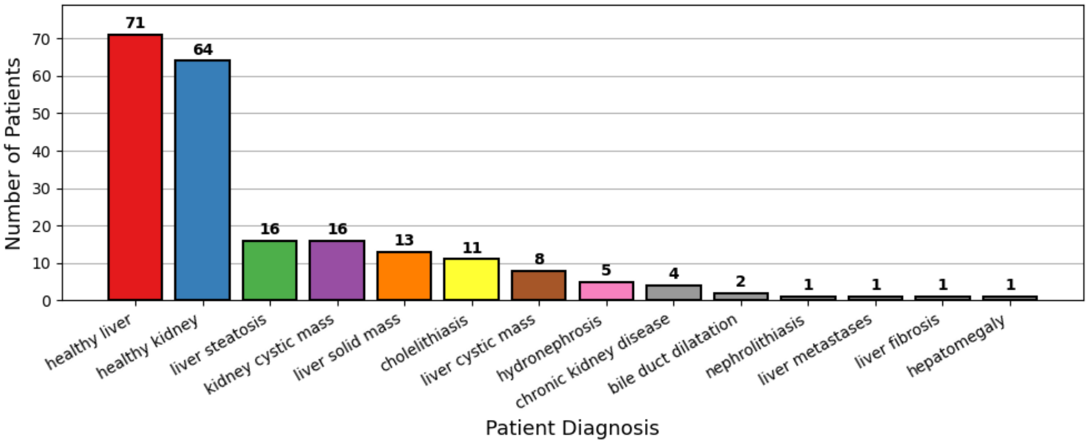
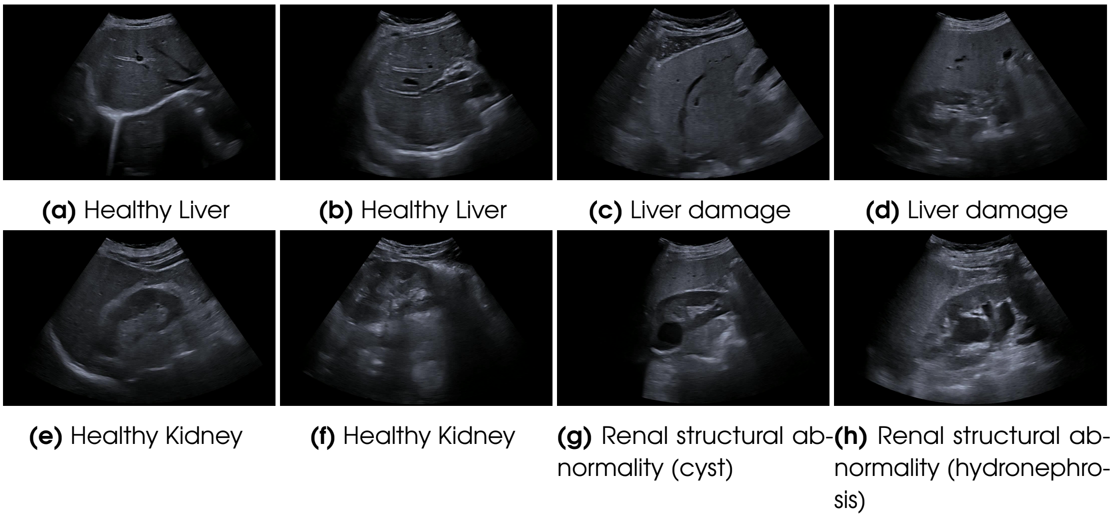


FIGURE 4.5 – Class distribution MIM-US-107 Video Dataset. Each bar represents a pathology, showing the number of videos containing each pathology.

The MIM-US-473 Still Image Dataset

Source data. This retrospective dataset comprised anonymized still-image snapshots routinely captured during standard clinical practice as part of patients' electronic health records. These images, a regular component of medical documentation, were used by clinicians to support their findings and document examinations. A total of 7,924 B-mode images were from 473 patients. No patient featured in both the still image dataset and the video dataset. This dataset was used exclusively to train the FRA.

Annotation. The radiographers evaluated each image by assigning an ordinal relevance score on a scale from 0 to 3, where 0 being irrelevant and 3 being highly



(a) Healthy Liver **(b)** Healthy Liver **(c)** Liver damage **(d)** Liver damage

(e) Healthy Kidney **(f)** Healthy Kidney **(g)** Renal structural abnormality (cyst) **(h)** Renal structural abnormality (hydronephrosis)

Figure 4.6 – Labeled videos from the MIM-US-107 Video Dataset, showing the keyframes selected by a radiographer to support their diagnosis.

relevant. As explained in Section 4.3.1, these relevance scores were based on general diagnostic criteria, combining two aspects :

- Organ Visibility : How much of the liver or kidney is visualized clearly in the image.
- Presence of Artifacts : Distortions or obstructions that could interfere with interpretation.

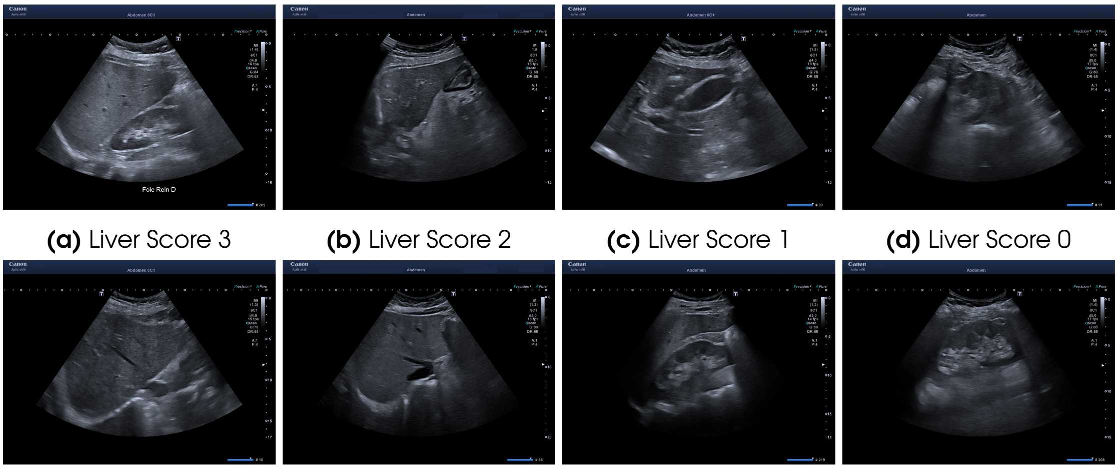
The specific relevance criteria for the Liver and Kidney Tasks are outlined in Table 4.1. Figures 4.7 and 4.8 show representative examples of annotations corresponding to each relevance score defined in Table 4.1. These examples illustrate significant variability in organ visibility due to the viewpoint, shape, and size of the respective organs. Additionally, it is important to note that an image categorized as 'Score 0' for one organ might be annotated as 'Score 3' for the other organ. As a consequence, the relevance criteria are task-specific.

4.3.4.2 Dataset labeling

In this section, we describe the datasets used for exploring and evaluating video classification models on untrimmed abdominal B-mode ultrasound data. Two distinct datasets, both anonymized and sourced from our partner hospital, are employed. The first dataset, referred to as the **'MIM-US-107 Video Dataset'**, consists of untrimmed ultrasound screening videos from 107 patients. It serves as the foundation for training and evaluating video-based diagnostic models. The second dataset, known as the **'MIM-US-473 Still Image Dataset'**, includes ultrasound snapshot images from 473 patients, captured in an unstructured manner during

Table 4.1 – Image relevance definitions and criteria for The Liver and Kidney Tasks (Section 4.3.4.1). The scores are ordinal and divided into four categories based on standard practice in US examination : 0 (not relevant) - indicates insufficient visual information, 1 - (mildly relevant) indicates some relevant visual information but likely insufficient for a healthy or pathological diagnosis, 2 (relevant) - indicates sufficient information, and 3 (highly relevant) indicates near-optimal information. NL and NK provide the number of times the score was assigned for the Liver and Kidney tasks.

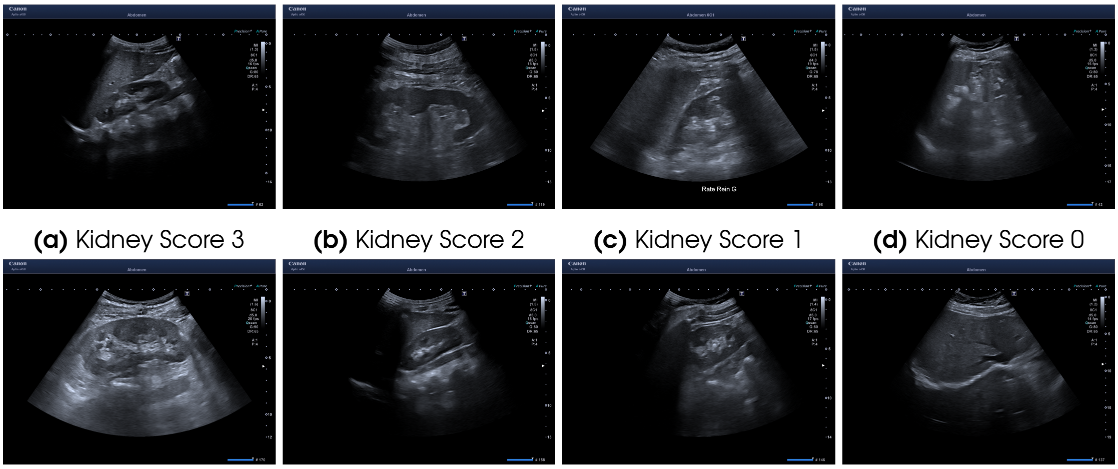
Score	Definition for Liver Task	NL	Definition for Kidney Task	NK
Score 3	*Artefacts* : non-existing or minimal in liver parenchyma *Organ Visibility* : 1. Liver-Kidney plane (liver at least 30 % image), or 2. Visible liver (at least 50% of the US), visible portal vein	151	*Artefacts* : non-existing or minimal in kidney parenchyma 1. Kidney centered screen, 2. Long and short axis visible, 3. Clear visualization of the renal pelvis and renal calyces, 4. Well-defined contours of the dilated collecting system, 5. Contrast between the fluid-filled structures and normal kidney tissue	204
Score 2	*Artefacts* : small amount in liver parenchyma *Organ Visibility* : 1. Liver-Kidney plane (liver bigger than kidney) , or 2. Visible liver (at least 50% of the US)	580	*Artefacts* : small amount in kidney parenchyma 1. Good visualization of the kidney (at least 50% of the US window) 2. Long and short axis seen, and 3. Partial visualization of the renal pelvis and renal calyces	428
Score 1	*Artefacts* : can be significant *Organ Visibility* : 1. Partial part of the liver being identifiable	458	*Artefacts* : can be significant *Organ Visibility* : 1. Partial part of the kidney being identifiable	286
Score 0	*Artefacts* : can be significant, limiting diagnosis *Organ Visibility* : 1. No or poor visualization of liver parenchyma	2200	*Artefacts* : can be significant, limiting diagnosis *Organ Visibility* : 1. No or poor visualization of kidney parenchyma	2473



(a) Liver Score 3 **(b)** Liver Score 2 **(c)** Liver Score 1 **(d)** Liver Score 0

(e) Liver Score 3 **(f)** Liver Score 2 **(g)** Liver Score 1 **(h)** Liver Score 0

FIGURE 4.7 – Examples of relevance score annotations for the Liver Task. The images illustrate relevance scores ranging from 'Score 0' (not relevant) to "(highly relevant).



(a) Kidney Score 3 **(b)** Kidney Score 2 **(c)** Kidney Score 1 **(d)** Kidney Score 0

(e) Kidney Score 3 **(f)** Kidney Score 2 **(g)** Kidney Score 1 **(h)** Kidney Score 0

FIGURE 4.8 – Examples of relevance score annotations for the kidney. The images illustrate diagnostic relevance scores ranging from 'Score 0' (not relevant) to "(highly relevant).

routine abdominal examinations. This dataset is used to train the Relevant Frame Assessor (FRA), enabling the automatic generation of diagnostic relevance scores.

4.3.4.3 FRA Implementation and training

In our experiments, we modeled the FRA with a fine-tuned Inception-ResNet-v2 backbone [233]. This was mainly motivated based on its strong performance in the previous chapter. It was adapted by replacing the classifier head with a regression head. Two FRAs were trained : one for each task, using all images in the **MIM-US-473 Still Image Dataset** as training data. The ordinal labels of each image were linearly rescaled to regression targets $\{r_1, r_2, \ldots, r_N\}$ in the range 0.0 to 1.0, where N denotes the number of training images. The models were trained with the Mean Smooth L1 loss function :

$$L_{FRA} = \frac{1}{N} \sum_{i}^{N} \mathit{SmoothL1}\left(\Phi_{RFA}\left(I_i\right), r_i\right) \tag{4.6}$$

where I_i denotes the i^{th} training image, $\Phi_{RFA}\left(I_i\right)$ denotes the model's predicted relvance score, and *SmoothL1* is the Smooth L1 function :

$$\mathit{SmoothL1}\left(x, y\right) = \begin{cases} 0.5(x-y)^2/\beta, & \text{if } |x-y| < \beta \\ |x-y| - 0.5 * \beta, & \text{otherwise} \end{cases} \tag{4.7}$$

The term β is a smoothing parameter set to the standard value of $\beta = 1$. Following best practice, geometric and photometric data augmentation were applied to increase generalization. However, we limited it to mild augmentation because strong changes, such as pronounced image cropping, may significantly affect the image's relevance. The training hyperparameters, including augmentation parameters, are presented in Table 4.2.

TABLE 4.2 – Frame Relevance Assessor (FRA) training hyperparameters

Network Parameter	**Value**
Pre-trained	imagenet
Backbone	inceptionresnetv2
Dense Layers	[512, 16, 1]
Learning Rate	10^{-5}
Training Epochs	116
Training Loss	Smooth L1 Loss
Early Stop Patient Epochs	10
Min. Epochs	100
Max. Epochs	1000

4.3.4.4 Diagnostic Classifier implementation and training

Video classifiers. In our experiments, two Video Swin Transformer models were trained (small variant, pre-trained on ImageNet 22k [59]), referred to as **Video Swin + DR-Clips**. One model was trained per task, using the Adam optimizer and the

DR-Clip loss as defined in Equation 4.3. Class imbalance reweighting was applied to handle the higher proportion of healthy versus pathological videos, as illustrated in Figure 4.6. Geometric and photometric data augmentation techniques were applied to improve model generalization, and, like the FRE, it was limited to mild augmentation. The augmentation parameters were randomized at the DR-Clip level (all frames in a DR-Clip received the same augmentation parameters). Full details on the training hyperparameters used in our experiments, including data augmentation parameters, optimization schedule, and early stopping criteria, are provided in Table A.1 in the appendix.

Single-image classifiers. The single-image versions of the Diagnositic classifer, discussed in Section 4.3.3.1), were implemented as follows. For each of the two tasks, two models were trained : an Inception-ResNet-v2 image classifier [233], referred to as **Resnet (Auto)**, and a Swin Transformer using a video size of 1 frame, referred to as **Swin (Auto)**.

4.4 Results

In this section, we present our experimental results. We compare our method against several competitive baselines, including both single-image and video-based models. Additionally, we provide a dedicated performance analysis of the FRA. All models were trained on a task-by-task basis (Liver Task and Kidney Task) using 10-fold cross-validation. Fold splits were generated randomly and on a per-patient basis (so no patient appeared in training and test splits simultaneously). The main performance metric was ROC-AUC (Receiver Operating Characteristic - Area Under the Curve).

4.4.1 Baseline methods

We compared our method against two categories of baselines :

- *Unguided Video Classifiers* : These were existing video classifiers that were trained and evaluated directly on the untrimmed videos from the MIM-US-107 Video Dataset. Similar to our proposed methodology, these classifiers did not have access to annotated video keyframes in the dataset. They were used as baselines to evaluate the impact of incorporating automatic video frame guidance via our FRA on video classification performance.
- *Manually-Guided Single-Image Classifiers* : These models were trained and evaluated using manually selected keyframes from the videos. The keyframe

selection process followed the methodology described in Section 4.3.4.1 - The MIM-US-107 Video Dataset : Annotation. These baselines were used to assess the performance gap between classifiers that relied on automatic keyframe selection (via the FRA) and those that used manually selected keyframes.

Unguided Video Classifiers. Seven state-of-the-art video classifiers were evaluated using the authors' publicly available code. They have have been presented in more detail in the Background Chapter. Four of the methods - Video Swin [68], VideoMAE V2 [256], MViT V2 [148], and X3D [69] — were selected from the general computer vision literature. These models have demonstrated strong performance in video classification across various domains. The remaining three methods — BabyNet [186], EchoGNN [173], and UVT [199] — were specifically designed for ultrasound video classification.

Because some of the methods were not tested in the binary classification setting, some minor adaptations were required. In **BabyNet** [186], only one view is considered. In **UVT** the output is adapted for binary classification.

All the above methods perform inference on untrimmed videos by dividing the videos into fixed-sized clips. The clip predictions are then combined with a fusion method. For a fair comparison, clip sizes of 32 frames were used for all models, which is the most common size used in the above methods. We compared two fusion methods : **max fusion**, where final video classification is computed from the clip with the highest classification confidence, and **mean fusion**, where final video classification is computed from the mean classification confidence of all video clips. The models were trained using the authors' code with their default training hyperparameter, reproduced in Table A.1 of the appendices.

Manually-guided single-image classifiers. These classifiers were trained on keyframes from the MIM-US-107 Video Dataset, which were manually selected by radiographers as described in Section 4.3.4. Inference was performed on untrimmed test videos, with final predictions aggregated, where we compared max fusion and mean fusion. To ensure a fair comparison with our approach detailed in Section 4.3.4.4, the same image classifier model was used (ResNet) and we refer to this baseline as **ResNet (Manual)**. It was trained in the same conditions and training hyperparameters as **ResNet (Auto)**.

4.4.1.1 Main results

The results are summarized in Table 4.3, which shows each method's ROC-AUC score, averaged over each fold, alongside the ROC-AUC standard deviation across folds (bracketed). Considering the Single-image models, we observe the following :

1. *Strong performance of **Resnet (Manual)*** : We recall this is a single-image approach trained on manually selected healthy/pathological keyframe images. It performed relatively in both the Liver and Kidney Tasks, consistently achieving AUC scores above 0.91 with either fusion strategy. This finding indicates that when a single-image classifier is trained on carefully selected keyframe images, it can perform well on the tasks, which implies there is limited benefit in temporal video features (which are not exploited by a single-image approach).
2. *Performance differences of Max and Min Fusion with **Resnet (Manual)*** : The fusion strategies for the Liver Task yielded very similar performance (mean ROC-AUCs of 0.96 and 0.95 for mean and max fusion respectively). In contrast, Max Fusion yielded a higher mean ROC-AUC for the Kidney Task than Mean Fusion (0.98 versus 0.91). The performance difference can be attributed to the nature of the pathologies. In the Liver Task, the pathology is generally diffuse (liver damage due to steatosis or fibrosis typically affects a large proportion of the liver), so most frames showing the liver consistently represent the same class. This uniformity allows both Mean Fusion and Max Fusion to perform well, as Mean Fusion benefits from averaging consistent predictions across frames, while Max Fusion reliably identifies the correct class from the most confident frame. In the Kidney Task, kidney cysts are spatially localized and appear only in specific frames. The lower performance of Mean Fusion can be explained by the fact that it averages classifications from frames without the pathology, diluting the confidence of the correct class. Max Fusion, however, focuses on the frame with the highest pathology confidence, which is more likely to capture the localized nature of the pathology.
3. *Performance difference of **ResNet (Auto)** and **Swin (Auto)** across tasks, due to noisy image-level labels* : These models achieved similar performance on the Liver Task using Mean Fusion, with ROC-AUCs of 0.94. However, on the Kidney Task, although Swin (Auto) outperformed ResNet (Auto), both models underperformed compared to ResNet (Manual). This disparity is primarily due to the issue of noisy labels. In the Kidney Task, ResNet (Auto) is trained with many noisy labels because the FRA selects frames that include the kidney, regardless of whether it appears healthy or pathological. Since the pathological label is applied to all selected frames during training, frames

depicting healthy regions of the kidney are mislabeled, introducing noise and degrading the model's inference performance. In contrast, this issue is substantially less significant in the Liver Task. The pathology in the Liver Task is diffuse, meaning that the FRA predominantly selects frames that consistently represent the pathological label. As a result, the training labels for the Liver Task are less noisy, enabling better model performance.

Considering the video models, including our proposed approach (DR-Clips), we observe the following :

4. *Strong performance of some unguided video classifier baselines in the Liver Task* : The unguided video classifiers **MViTv2**, **Video Swin**, and **X3D**, using Max Fusion, achieved mean ROC-AUCs of 0.90, 0.91, and 0.94, respectively. The best results are comparable to those of **ResNet (Auto)** and **ResNet (Manual)**, which achieved ROC-AUCs of 0.94 and 0.95, respectively.
5. *Weak performance of unguided video classifier baselines in the Kidney Task* : In the Kidney Task, none of the methods produced strong results comparable to **Resnet (Manual)**. The three strongest methods (**Video Swin**, **MViTv2** and **X3D**) produced mean ROC-AUCs between 0.72-0.74.
6. *Potential reasons for the performance gap* : The stronger performance in the Liver Task suggests that untrimmed video classification is inherently easier for this task compared to the Kidney Task. This can be attributed to the diffuse nature of liver pathology, which is detectable across many frames within a video. In this case, the abundance of frames with consistent features reduces the risk of learning irrelevant patterns and enables classifiers to generalize effectively, even without precise frame selection. In the Kidney Task, however, the pathology is localized and may only appear in a small subset of frames. As described above, the majority of frames may lack pathological features or even show healthy tissue, leading to a higher proportion of irrelevant data. This may increase the risk of overfitting, where classifiers learn spurious or irrelevant patterns from the abundance of unrelated frames. Additionally, the smaller size of the kidney as an organ could exacerbate this problem by further reducing the visibility and frequency of pathological features, making it harder for unguided classifiers to focus on diagnostically relevant information. Ultimately, the diffuse pathology in the Liver Task makes the problem more forgiving, while the Kidney Task requires better focus and better handling of irrelevant data.
7. *Strong performance of Video Swin + DR-Clips in both tasks* : The combination of DR-Clips and Video Swin achieved mean ROC-AUCs of 0.97 for the Liver Task and 0.92 for the Kidney Task, demonstrating substantial improvement over Video Swin, particularly for the Kidney Task. The model achieved results very

close to **ResNet (Manual)**, highlighting the effectiveness of frame guidance provided by DR-Clips.

Video Swin + DR-Clips outperformed **Swin (Auto)** in the Kidney Task, with mean ROC-AUCs of 0.92 compared to 0.81. This highlights the limitations of single-image classifiers trained with automatic relevance filtering. Both **Video Swin + DR-Clips** and **Swin (Auto)** use the same classification model and rely on the same relevance filtering model (the FRA) to identify diagnostically important frames. However, **Video Swin + DR-Clips** may outperform **Swin (Auto)** for two plausible reasons :

Firstly, **Video Swin + DR-Clips** processes batches of frames, potentially enabling the model to leverage relationships among frames and draw on collaborative information. This allows the model to capture patterns and context shared across multiple frames, which may improve its ability to distinguish pathological features. In contrast, **Swin (Auto)** treats each frame independently, which prevents it from utilizing inter-frame consistency or contextual information.

Secondly, as discussed in Item 3, while relevance filtering reduces irrelevant data, treating each frame independently as fully representative of the target label (as in **Swin (Auto)** increases the likelihood of label noise. This noise can arise when selected frames from pathological cases depict healthy tissue or ambiguous features, leading to incorrect labeling during training. In contrast, **Video Swin + DR-Clips** processes batches of frames selected by DR-Clips. For a DR-Clip to be mislabeled in a pathological case, all frames within the batch would need to depict healthy tissue, which is statistically less likely. This batch-based approach likely reduces the overall probability of label noise, enabling the model to learn the underlying task more effectively. The video capabilities of DR-Clips, combined with our proposed training strategy, effectively overcome the state-of-the-art limitations for training and inference with untrimmed abdominal ultrasound videos.

4.4.2 Secondary Analysis

In our secondary analysis, we conduct a feature correlation analysis to evaluate the effectiveness of the FRA, and we we examine the effect of DR-Clips length on performance.

4.4.2.1 How effective is the Frame Relevance Assessor ?

Quantifying the performance of the FRA poses a challenge, as it is trained to generate continuous relevance scores. Since we do not have continuous valued

TABLE 4.3 – Comparison of the proposed method (DR-Clips) against state-of-the-art single-image and video classification models. Fold-averaged values are presented with standard deviations in brackets.

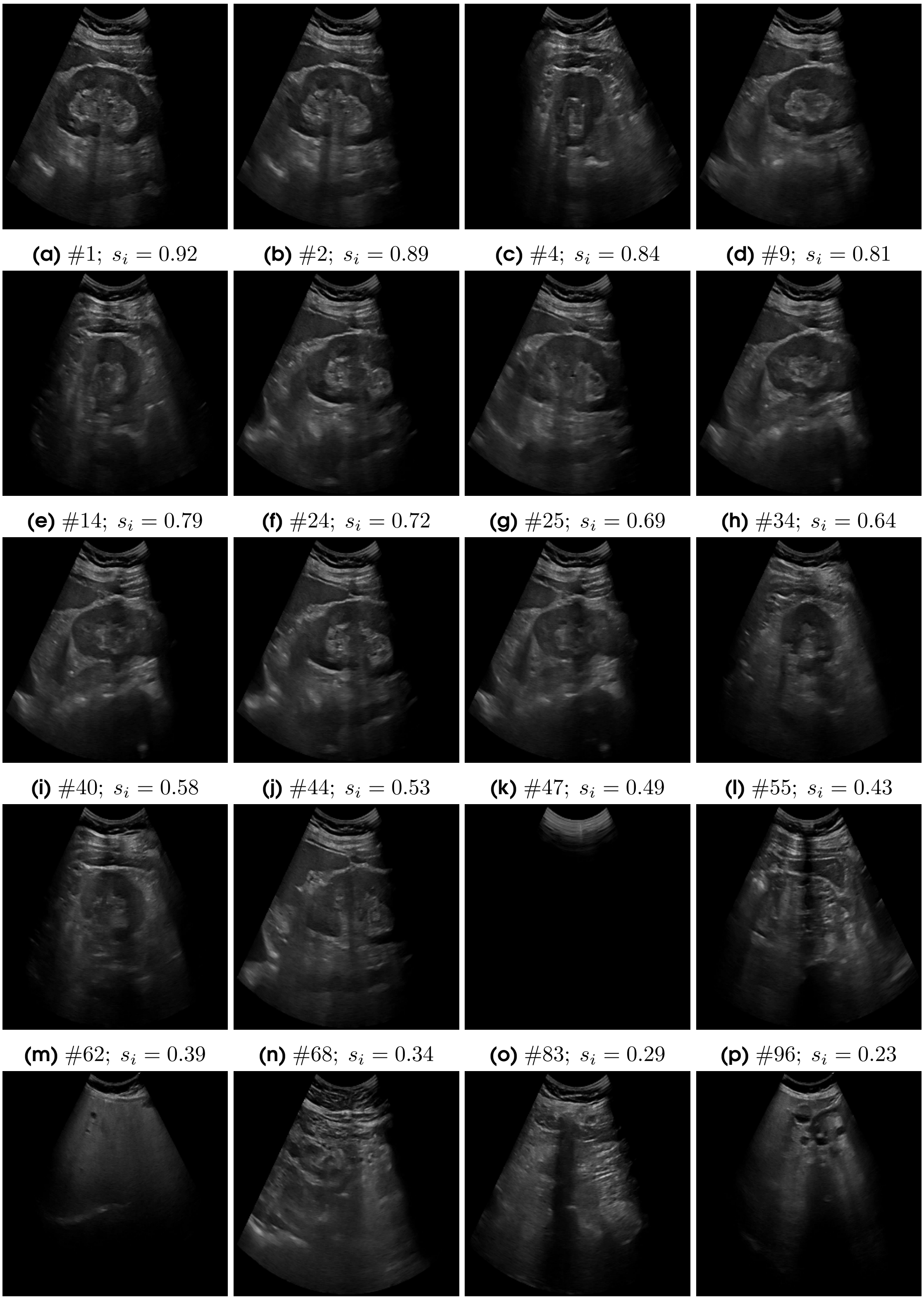
		Liver ROC-AUC (std)		Kidney ROC-AUC (std)	
Model Category	Model name	mean fusion	max fusion	mean fusion	max fusion
Single-image	Resnet (Manual) [233]	0.96 (0.11)	0.95 (0.15)	0.91 (0.25)	0.98 (0.05)
	Resnet (Auto) [233]	0.94 (0.17)	0.94 (0.13)	0.71 (0.26)	0.53 (0.24)
	Swin (Auto) [162]	0.94 (0.17)	0.89 (0.16)	0.81(0.21)	0.69 (0.22)
Video	BabyNet [186]	0.89 (0.25)	0.87 (0.30)	0.61 (0.21)	0.59 (0.31)
	EchoGNN [173]	0.82 (0.29)	0.77 (0.30)	0.55 (0.19)	0.60 (0.24)
	UVT [199]	0.80 (0.29)	0.74 (0.24)	0.68 (0.17)	0.67 (0.16)
	Video Swin [68]	0.89 (0.30)	0.91 (0.17)	0.70 (0.29)	0.72 (0.22)
	VideoMAE V2 [256]	0.80 (0.37)	0.87 (0.26)	0.56 (0.25)	0.60 (0.20)
	MViTv2 [148]	0.88 (0.30)	0.90 (0.26)	0.74 (0.19)	0.64 (0.26)
	X3D [69]	0.85 (0.24)	0.94 (0.13)	0.73 (0.23)	0.69 (0.26)
DR-Clips	Video Swin + DR-Clips	0.97 (0.09)[1]		0.92 (0.13)[1]	

test labels, we cannot used standard regression metrics such as Mean Absolute Error. Instead, we present results qualitatively, as well as quantitatively with feature correlation analysis.

Qualitative results are presented in Figures 4.9 and 4.10. These figures were generated using the predicted relevance scores of the first video in the MIM-US-107 Video Dataset, having a diagnosis of liver steatosis. Its frames were binned 5 categories in the range of 0.0 to 1.0 in intervals of 0.1. Then 5 random frames of each bin were selected and shown in relevance descending order in the figures. One can see that higher-relevance images consistently displayed better organ visibility, with minimal shadows and artifacts, making them more suitable for diagnosis.

Figure 4.11 shows the distribution of normalized relevance scores predicted by the FRA for the Liver and Kidney tasks, for every frame in the MIM-US-107 Video Dataset. The distributions were stratified into healthy and pathological patients. One can see there were a substantial amount of low relevance frames for the Kidney task (a normalized relevance score of 0.3 corresponds to Label 1 (low relevance) according to the annotation criteria in Table 4.1. This likely contributed to the difficulty of the kidney Task, compared to the Liver Task, as reflected in our experiments. The proportion of irrelevant healthy kidney images was greater compared the proportion of irrelevant pathological kidneys. This may reflect the fact that in pathological cases, more video time was spent inspecting the kidney, compared to healthy case.

In our feature correlation analysis, we evaluated whether frames with visual



(a) $\#1;\ s_i = 0.92$ **(b)** $\#2;\ s_i = 0.89$ **(c)** $\#4;\ s_i = 0.84$ **(d)** $\#9;\ s_i = 0.81$

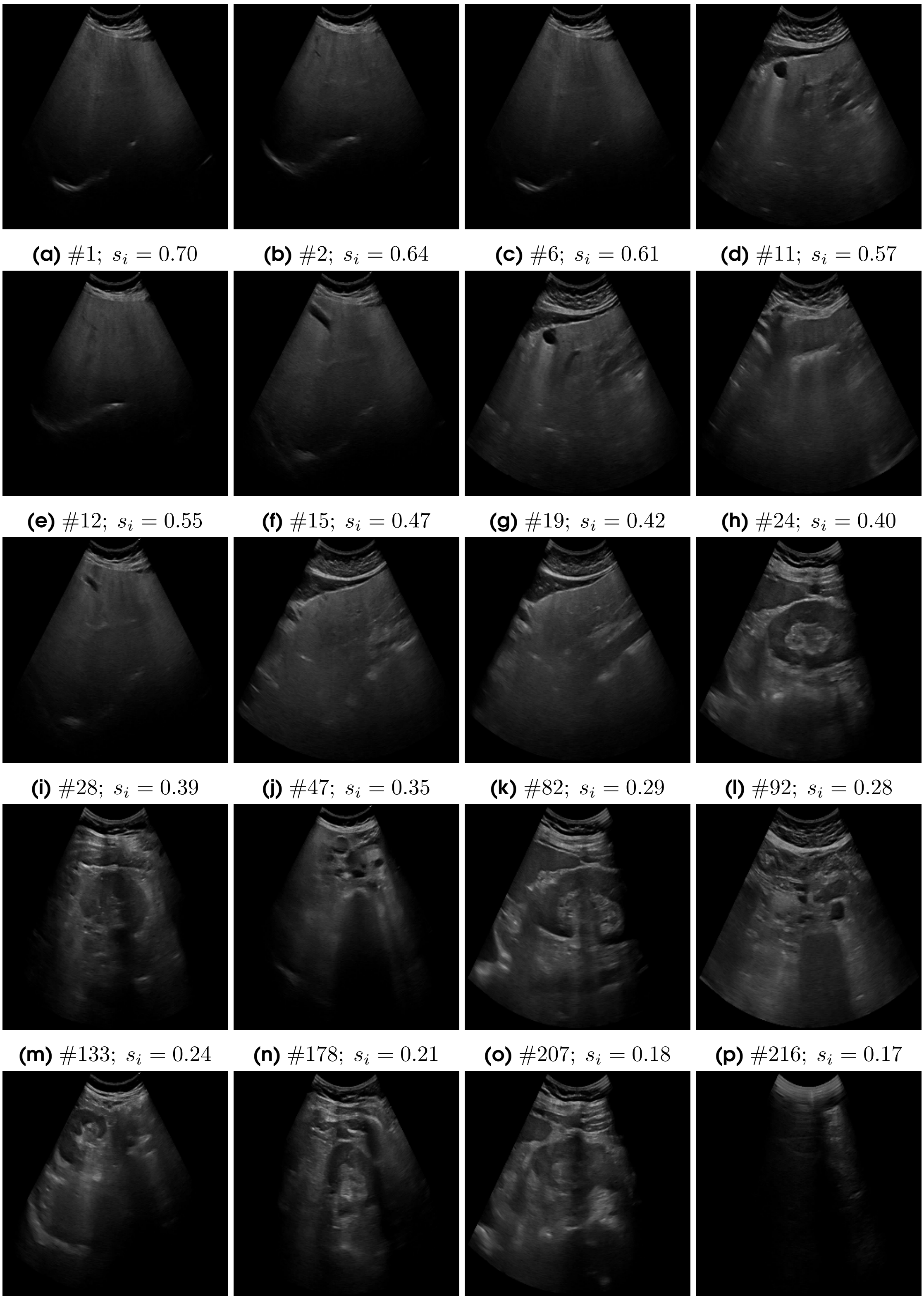
(e) $\#14;\ s_i = 0.79$ **(f)** $\#24;\ s_i = 0.72$ **(g)** $\#25;\ s_i = 0.69$ **(h)** $\#34;\ s_i = 0.64$

(i) $\#40;\ s_i = 0.58$ **(j)** $\#44;\ s_i = 0.53$ **(k)** $\#47;\ s_i = 0.49$ **(l)** $\#55;\ s_i = 0.43$

(m) $\#62;\ s_i = 0.39$ **(n)** $\#68;\ s_i = 0.34$ **(o)** $\#83;\ s_i = 0.29$ **(p)** $\#96;\ s_i = 0.23$

(q) $\#112;\ s_i = 0.19$ **(r)** $\#157;\ s_i = 0.15$ **(s)** $\#244;\ s_i = 0.08$ **(t)** $\#302;\ s_i = 0.04$

FIGURE 4.9 – Frames from patient PID1 in our video dataset, displaying normalized relevance scores (s_i) estimated by our FRA on the Kidney Task. These scores reflect the algorithm's evaluation of the clinical relevance of each frame for diagnosing healthy and pathological kidneys.



(a) $\#1;\ s_i = 0.70$ **(b)** $\#2;\ s_i = 0.64$ **(c)** $\#6;\ s_i = 0.61$ **(d)** $\#11;\ s_i = 0.57$

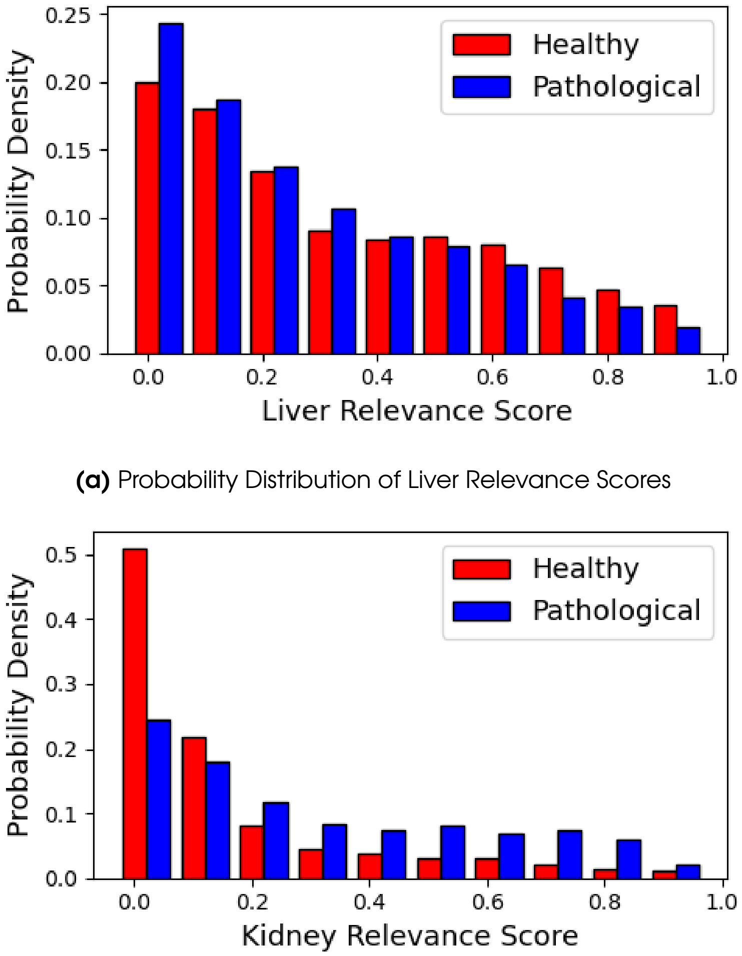
(e) $\#12;\ s_i = 0.55$ **(f)** $\#15;\ s_i = 0.47$ **(g)** $\#19;\ s_i = 0.42$ **(h)** $\#24;\ s_i = 0.40$

(i) $\#28;\ s_i = 0.39$ **(j)** $\#47;\ s_i = 0.35$ **(k)** $\#82;\ s_i = 0.29$ **(l)** $\#92;\ s_i = 0.28$

(m) $\#133;\ s_i = 0.24$ **(n)** $\#178;\ s_i = 0.21$ **(o)** $\#207;\ s_i = 0.18$ **(p)** $\#216;\ s_i = 0.17$

(q) $\#243;\ s_i = 0.15$ **(r)** $\#301;\ s_i = 0.11$ **(s)** $\#344;\ s_i = 0.09$ **(t)** $\#395;\ s_i = 0.06$

FIGURE 4.10 – Frames from patient PID1 in our video dataset, displaying the normalized relevance scores (s_i) predicted by our FRA on the Liver Task. These scores reflect the algorithm's evaluation of the clinical relevance of each frame for diagnosing healthy and pathological livers.



(a) Probability Distribution of Liver Relevance Scores

(b) Probability Distribution of Kidney Relevance Scores

FIGURE 4.11 – Distribution of normalized relevance scores estimated by the FRA for the Liver and Kidney tasks.

features similar to the manually selected keyframes in the video dataset were assigned higher relevance scores by the FRA. Image similarity was measured based on distances in a reduced visual feature space, as outlined below in three steps.

1. *Feature Extraction* : Each frame in the video dataset was processed through a pre-trained Inception-ResNet-v2 model [233], which was trained on ImageNet-22k [59]. We extracted the outputs of the first dense layer as high-dimensional visual features.
2. *Dimensionality Reduction* : Principal Component Analysis (PCA) was applied to reduce the dimensionality of these features, resulting in a compact visual feature representation for each frame.
3. *Distance Calculation* : In the reduced feature space, we calculated the nearest neighbor distances between all video frames and the manually selected keyframes, representing how visually similar each frame is to the

keyframe set.

Figure 4.12 presents a scatter plot, where each point corresponds to a frame in the video dataset. The x-axis shows the FRA-predicted relevance score for the frame, and the y-axis shows its distance to the keyframe set in the feature space. A strong inverse correlation would indicate that frames deemed highly relevant by the FRA are visually similar to those selected by human experts.

The Pearson correlation coefficients for this relationship were -0.84 for the Liver Task and -0.70 for the Kidney Task, demonstrating a strong negative correlation. This suggests that the FRA effectively identifies frames with features closely resembling those of the manually selected keyframes, with particularly high agreement for the Liver Task.

4.4.2.2 Effect of DR-Clip length at inference

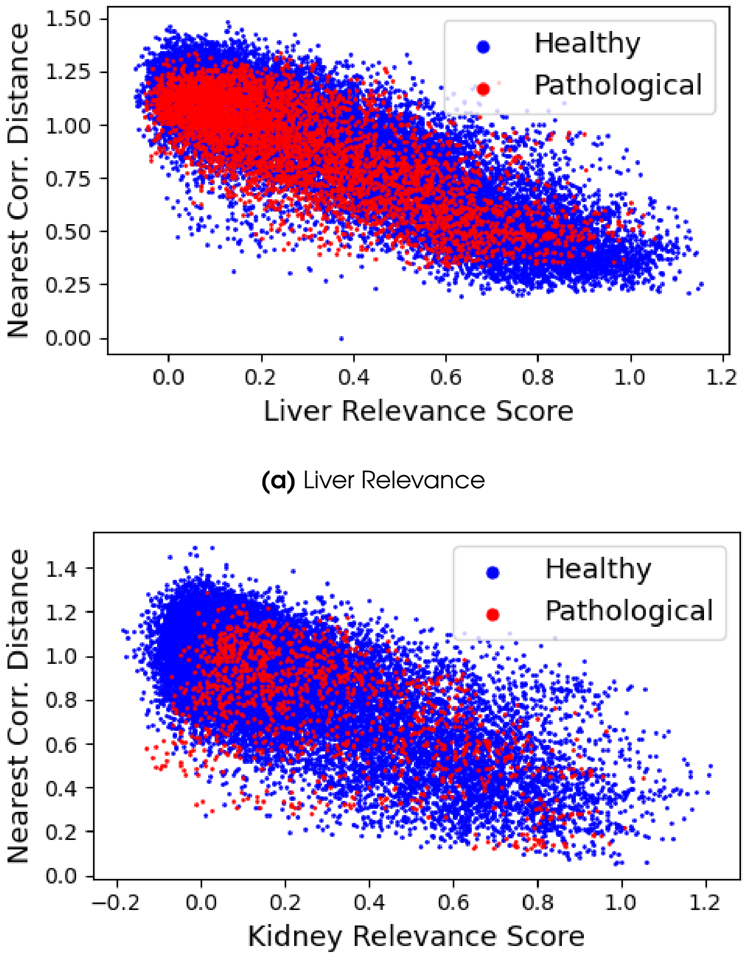
The Video Swin model supports variable clips size at inference time. Figure 4.13 shows the ROC-AUC scores for video classification, where we varied the length of the DR-Clips, passed to the Video Swin model, during inference. It generally shows superior performance with smaller clips. The results indicate that at inference time, the both tasks can be solved effectively with only a small number of relevant frames. Furthermore, increasing the size of the inference clips may harm performance, likely due to the increased proportion of irrelevant frames.

This effect is more pronounced for the Kidney Task, likely due to the localized nature of the lesions, which are adjacent to healthy-looking frames. The effect is less strong for the liver class due to the diffuse nature of the conditions studied, being visible in most of the liver.

4.5 Conclusion and Future Work

In this chapter, we addressed a critical question in ultrasound video computer-aided diagnosis (CAD) : Can video classifiers be effectively trained on untrimmed abdominal ultrasound videos by leveraging predictions from a model that assesses the general relevance of each frame ?

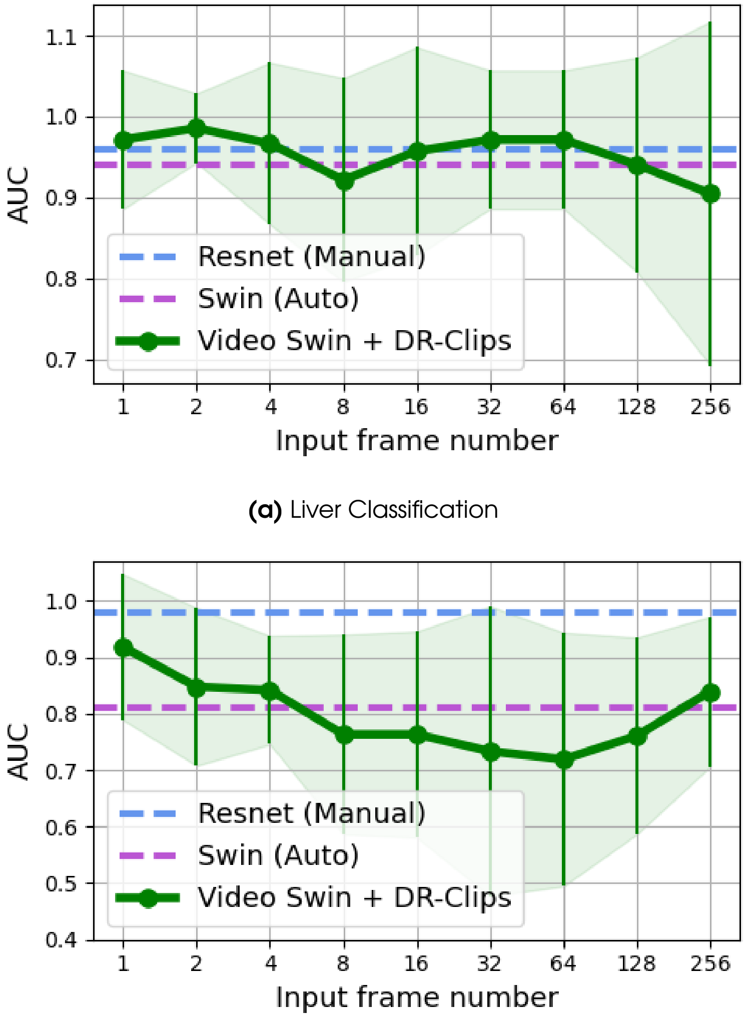
Our findings reveal a significant performance gap between image classifiers trained on manually selected keyframe images and state-of-the-art video classifiers trained on untrimmed videos that include these keyframes. This gap, previously not exposed due to the lack of abdominal ultraound video datasets, is primarily caused by the presence of non-relevant frames in untrimmed video data. This hampers the training and inference performance of video classifiers, especially in



(a) Liver Relevance

(b) Kidney Relevance

FIGURE 4.12 – Scatter plots showing the relationship between a frame's similarly with respect to the keyframe set, and its predicted relevance by the FRA. Two plots are shown, corresponding to the Liver and Kidney Tasks (above and below). The x-axis shows the FRA-predicted relevance score for the frame, and the y-axis shows its nearest-neighbor distance to the keyframe set in the feature space. Points are coloured according to whether the frame came from video labeled as healthy or pathological.



(a) Liver Classification

(b) Kidney Classification

FIGURE 4.13 – ROC-AUC of our trained model (**Video Swin + DR-Clips**) improves with smaller window sizes during inference, particularly with 4 or fewer frames.

the context of medical data constraints where uncurated ultrasound videos are not yet fully exploited for salable model training.

To address this challenge, we proposed DR-Clips, a novel approach that leverages a Frame Relevance Assessor (FRA) to guide the selection of diagnostically relevant frames. DR-Clips effectively narrows the performance gap between highly supervised single-image models, using image-level lables, and less supervised video-based approaches, using video-level labels. Our approach uses frame relevance scores to facilitate training and inference, making it especially effective for tasks involving localized or hard-to-detect pathologies.

Our results demonstrate the promise of DR-Clips and FRA in improving video classification performance : For the Liver Task (diffuse pathology), DR-Clips achieved ROC-AUC scores comparable to those obtained with image classifiers trained on manually selected keyframes. For the Kidney Task (localized pathology), DR-Clips significantly reduced the performance gap, although challenges remained when the size of DR-Clips for inference was enlarged, resulting in less relevant frames being passed to the viode classifier.

While DR-Clips substantially improves the performance of video classifiers, several areas require further exploration :

- *Generalizability* : A key question is how well the FRA, and DR-Clips can generalize especially to other pathologies, devices and centers. Future research is required to broaden the evaluation presented in this Chapter.
- *Application with other video classifiers* : We used DR-Clips with Video Swin, however, it is important to understand whether similar performance improvements are obtained across different competitive video classifiers.
- *Video classifier model improvement* : Our observation that inference performance could ultimately degrade if the clip length at inference time was too large. This suggests the model is not sufficiently robust to handle a large amount of low-relevance frames at inference time.
- *Supervised Learning Dependency* : The reliance on a supervised FRA may limit the applicability in scenarios with limited labeled data. To address this, follow-up research should explore weakly supervised learning approaches for frame relevance estimation, reducing the dependency on ultrasound data with frame-level relevance annotations. We explore this direction in the next chapter.

5. KeyFrameDiagFormer : Weakly-Supervised Keyframe Localization and Diagnosis Transformer Model for Untrimmed Ultrasound Videos

Chapter summary

In the preceding chapter, we introduced a video diagnostic method that uses external guidance to enable accurate automated diagnosis in untrimmed ultrasound videos. Despite its effectiveness, this reliance on an externsl, supervised relecance assessment methods poses a limitation for broader application, as extending the method to other pathologies with substantially different relevance criteria would necessitate a customized frame guidance system for each case.

In this section, we introduce an ultrasound video CAD method trained exclusively on video-level labels, which has not been done before with untrimmed abdominal ultrasound videos.This innovation has far-reaching implications : (1) it enhances scalability by enabling training with video-level labels, which can be sourced from other modalities like patient records, reducing the burden of detailed annotation ; (2) it improves explainability by identifying diagnostically relevant keyframes automatically ; (3) it mitigates selection bias, as the system independently identifies informative frames without explicit guidance ; and (4) it supports procedural video documentation through automated keyframe extraction, which could streamline clinical workflows and improve record-keeping.

The proposed architecture comprises four fundamental components : a **Frame Encoder**, a **Frame-Memory Module**, a **Video Self-Attention Module**, and a **Hierarchical Multi-Label Classification Module**.

The **Frame Encoder**, a 2D neural network, extracts representative frame features independently from the input ultrasound video. To efficiently process long videos while maintaining spatial resolution, we include a **Frame-Memory Module** that utilizes a memory bank to store frame features. During training, the Frame

Encoder is trained using only a subset of randomly selected frames, which update the respective frame features in the Frame-Memory Module, ensuring a balance between accuracy and computational efficiency. The frame features are then forwarded to the **Video Self-Attention Module**, which adds local and global organ-specific context into the feature representation. Finally, the **Hierarchical Multi-Label Classification Module** introduces a hierarchical structure to the prediction process.

For each frame, the system determines organ-specific relevance scores or background scores. The final video diagnosis detects one or more pathologies simultaneously, addressing a multi-label classification problem. This is achieved using an organ-specific global self-attention block, which enables features from frames containing a specific organ to interact with each other while ignoring background frames. For each pathology, a multi-label classification head aggregates the predictions, with the final predicted class obtained by averaging the top-N frame prediction scores for each class.

Compared to state-of-the-art, this approach offers several key advantages :

1. **Efficient Training and Inference :** The combination of the Frame Encoder and Frame-Memory Module enables frames to be processed in real-time during inference. It also maintains a favorable trade-off between spatial and temporal resolution for the video classification model during training.
2. **Keyframe Identification :** Inspired by weakly-supervised action localization, our system automatically identifies healthy and pathological frames in untrimmed videos, without requiring image-level labels. Additionally, it provides organ-specific relevance scores, explicitly distinguishing diagnostically relevant frames from background frames.
3. **Diagnostic Feasibility :** Leveraging a multi-label setup with specific output heads for healthy and diverse pathologies, our system can determine when a diagnosis for a particular organ is infeasible due to insufficient information.

The remainder of this chapter is organized as follows : Section 5.1 provides additional background that inspired our work. Section 5.2 details our methodology and database. Finally, Section 5.3 presents our results.

5.1 Additional background

5.1.1 Insights from Weakly-supervised Action Localization in the general computer vision literature

Weakly-supervised Temporal Action Localization (WTAL) involves identifying the temporal boundaries (start and end) of actions within a video and classifying those actions into predefined classes using weak annotations for training. These weak annotations typically consist of video-level action labels, significantly reducing the annotation effort compared to fully-supervised methods, which require frame-level supervision for training.

To produce a localization output, WTAL models typically generate an individual score for each frame and each class. These scores create class-temporal signals that are processed by post-processing modules to detect class-action boundaries, such as action start and end points.

WTAL shares similarities with the problem of diagnosing pathologies in untrimmed ultrasound videos. Specifically, a parallel can be made between action and background frames in WTAL to diagnostically relevant and non-relevant frames in US videos. Inspired by WTAL, an open research question that we answer is as follows : Can a neural network, that was originally designed to solve WTAL, be adapted to effectively identify diagnostically relevant frames within untrimmed ultrasound videos, without frame-level annotations or external guidance ?

A key mechanism employed by WTAL methods [158, 255, 141, 149, 75, 105, 81, 204, 196] is the explicit modeling of the background by distinguishing it from specified action classes. This concept is well explored in [158], where the background is treated as a separate action class present in all videos. Ground-truth labels for WTAL are thus composed of the actions present in the video combined with the background class. By using this approach, background scores are learned through the dissimilarities between videos containing different actions.

Another challenge in action localization is the high computational complexity associated with processing long video sequences. This arises from the need for most methods to maintain a temporally localized frame feature representation extracted by one (or more) Frame Encoder neural networks, which are themselves computationally demanding. Classical solutions to this problem typically constrain the models by reducing spatial-temporal resolution, employing simpler Frame Encoder backbones, or freezing the encoder weights.

A promising solution for efficient memory management is presented by [48], where the authors introduce a long-term memory module. This module first extracts frame representations from all clips in the input video using a clip encoder and stores them in a feature memory without retaining gradients. During training, randomly selected clips are fed to the encoder, and the extracted features are used to update the feature memory. The full feature memory, including the clip encoder, is trained, with gradients computed based on the randomly selected clips. This approach enables end-to-end training on long video sequences, making it particularly useful for analyzing untrimmed ultrasound videos.

However, the parallel between WTAL and the automatic diagnosis of pathologies in untrimmed ultrasound videos is not so straightforward. A key difference between the correspondence of WTAL action frames and diagnostically relevant frames in US videos lies in their temporal distribution. WTAL actions are typically temporally contiguous, whereas diagnostically relevant frames in US videos are sparsely distributed over time.

This is why many WTAL methods adopt a two-stream architecture, leveraging both RGB and optical flow inputs to model temporal dynamics effectively [158, 280, 141, 149, 105, 81, 204, 196, 175]. However, the temporal sparsity of diagnostically relevant frames in US videos, combined with the inherent challenges of generating reliable optical flow in ultrasound [183], significantly limits the applicability of this approach to US video diagnosis. Consequently, while WTAL methods provide a valuable foundation, their architectures require significant adaptation to address the unique temporal and imaging characteristics of US video diagnosis.

In summary, key insights from WTAL are incorporated into the development of our model. We draw inspiration from two core aspects : **Background Modeling [158]**, where the background is treated as a separate class to enhance discrimination between diagnostically relevant and non-relevant frames, and **Long-Term Memory [48]**, which enables efficient end-to-end training on long video sequences. These strategies are adapted to address the unique challenges posed by the temporal sparsity and imaging characteristics of untrimmed ultrasound videos, ensuring the relevance and effectiveness of our approach.

5.1.2 Multi-Label Learning with Missing Labels

In the previous chapter, we intentionally designed individual neural networks to analyze the kidney and liver pathologies, optimizing them under the condition that the null class was not considered. The null class, in this context, refers to scenarios where diagnosis is not feasible, either because the organ does not appear in the video or because its appearance is heavily obscured by noise and artifacts, making diagnosis unfeasible. However, to provide diagnostic feasibility feedback to the user, the inclusion of the null class is essential, as it is intrinsically linked to the Background Modeling discussed in the previous section.

One effective method for this could be Multi-Label Learning with Missing Labels [263, 276, 291, 290, 106, 147], which allows each video to simultaneously belong to multiple classes. For example, an abdominal ultrasound video may exhibit liver lesions, liver steatosis, and kidney lesions at the same time, and the objective is to identify all these conditions during a single forward pass.

Among the various methods available for handling Multi-Label problems, Problem Transformation Methods, which convert Multi-Label problems into multiple single-label problems, are the easiest to adapt to neural network architectures. One such method is Binary Relevance [287], which transforms a Multi-Label problem into multiple independent single-label classification problems, where each label indicates the presence (1) or absence (0) of a specific condition. Because these are treated as independent classification tasks, the Cross-Entropy Loss used in multiclass classification is replaced with Binary Cross-Entropy Loss.

The concept of Missing Labels refers to cases where an ultrasound video lacks a diagnosis for a given organ, corresponding to the null class. For instance, a video may contain liver labels but no kidney label, indicating missing information for the latter. A common approach to handle this challenge is to consider only observable labels during training [276].

5.1.3 Hierarchical Classification

The diagnosis of ultrasound videos can be framed as a Hierarchical Classification Problem [260, 286, 41, 7, 167]. Hierarchical classification is a subtype of multi-label classification where classes are organized in a structured hierarchy. For example, to determine whether a frame indicates a pathological or healthy liver, the presence of the liver in the frame must first be confirmed ; otherwise, the diagnosis is infeasible. In Figure 5.1 we show the hierarchy we use for the analysis of ultrasound videos in this chapter.

A common approach to solving Hierarchical Classification problems is to use flat classification, where only the classes at the leaf nodes are used as training labels (highlighted by the dotted lines in Figure 5.1). This method is well-suited to our case, given the nature of our data and the inclusion of the background class.

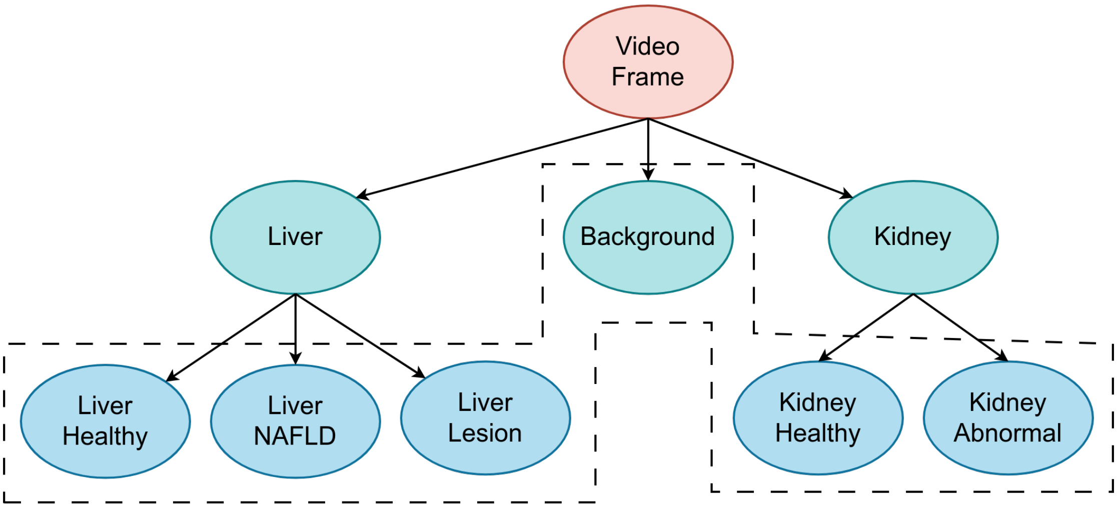


FIGURE 5.1 – The hierarchical structure used for the diagnosis of ultrasound videos in this chapter. Highlighted paths (dotted lines) indicate the labels used for training in the flat classification approach. During inference, only level 3 leaf nodes (blue ones) are used to provide an output.

5.2 Methodology

5.2.1 Section overview

Given an untrimmed ultrasound video $\mathbf{V}^i = \{\mathbf{I}_t\}_{t=1}^{T} \in \mathbb{R}^{C\times T\times H\times W}$, our goal is to predict one or more diagnoses attributed to the video. We denote these predictions as $\hat{\mathbf{Y}}^i = \{\hat{y}_k\}^{N_i}$, where N_i is the number of diagnoses detected in video i.

We state it as a multi-label classification problem as follows, where $\mathbf{O} = \{o_1, \cdots, o_n\}$ represents the set of organs being considered and each organ o has a healthy label h_o and a set of pathological labels $\left\{p_{\text{o}}^0, p_{\text{o}}^1, \cdots, p_{\text{o}}^k\right\}$.

$$\mathbf{Y}^i = \left\{h_{\text{o}}, p_{\text{o}}^0, p_{\text{o}}^1, \cdots, p_{\text{o}}^k \;\middle|\; \forall \text{o} \in \mathbf{O}\right\} \in \{0,1\}^{N_c} \tag{5.1}$$

The number of diagnosis classes is defined as N_c, comprising all h_o and p_{o}^k for all organs in $\mathbf{O}$. To maintain label consistency between healthy and pathological diagnoses for each organ o, we apply the following rules :

$$\begin{aligned}
&\text{If } h_{\text{o}} = 1, && \text{then } p_{\text{o}}^k = 0,\ \forall k.\\
&\text{If } p_{\text{o}}^k = 1 \text{ for any } k, && \text{then } h_{\text{o}} = 0.\\
&\text{Multiple } p_{\text{o}}^k \text{ can be 1 simultaneously.} &&
\end{aligned} \tag{5.2}$$

If no labels are estimated for a given organ o, this indicates that a diagnosis is not feasible for that organ based on the information available in $\mathbf{V}^i$.

Finally, for each diagnosis label present in $\hat{\mathbf{Y}}^i$, we want to output a list of R diagnostically relevant frames $\mathcal{K}_{\hat{y}_k}$ to support the diagnosis made.

$$\mathcal{K}_{\hat{y}_k} = \{\mathbf{I}t \mid t \in \mathcal{T}k, \mathbf{I}_t \text{ supports the diagnosis of } \hat{y}_k\}\,, \quad |\mathcal{T}k| = R \tag{5.3}$$

By doing this, we can perform we diagnosis of multiple pathologies at multiple organs, indicating which frames were used to perform the diagnosis. Our proposed architecture is illustrated in Figure 5.2.

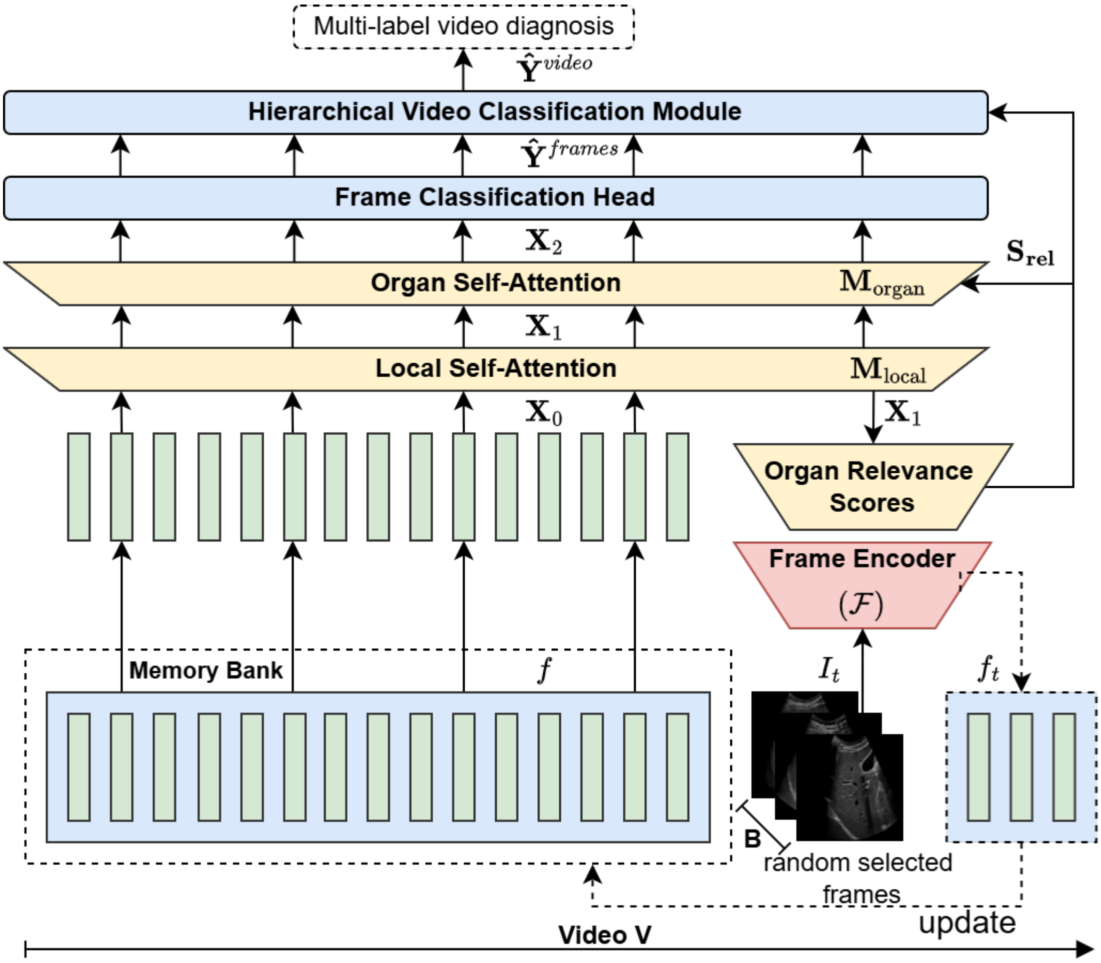

Multi-label video diagnosis
$\hat{\mathbf{Y}}^{video}$
Hierarchical Video Classification Module
$\hat{\mathbf{Y}}^{frames}$
Frame Classification Head
$\mathbf{X}_2$
$\mathbf{S}_{\mathbf{rel}}$
Organ Self-Attention
$\mathbf{M}_{\mathrm{organ}}$
$\mathbf{X}_1$
Local Self-Attention
$\mathbf{M}_{\mathrm{local}}$
$\mathbf{X}_0$
$\mathbf{X}_1$
Organ Relevance Scores
Frame Encoder
$(\mathcal{F})$
Memory Bank
f
I_t
f_t
B
random selected frames
update
Video V

Figure 5.2 – The hierarchical classification structure used for diagnosing ultrasound videos.

5.2.2 Proposed Video Classification Network

Our proposed neural network is composed of the following modules :

5.2.2.1 Long-memory Frame Embedding Bank

Considering a collection of video frames $\mathbf{V}^i = \{\mathbf{I}_t\}_{t=1}^{T} \in \mathbb{R}^{C\times T\times H\times W}$, we first utilize a still image neural network to extract image features, which are then stored in a memory bank. This 2D neural network is defined as $\mathcal{F} : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{C_1\times C_2\times f_{\text{dim}}}$, followed by the following operations : layer normalization, a max polling layer over C_1 and C_2 and flattening operation.The entire process can be formally expressed as :

$$\mathcal{F} : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{f_{\text{dim}}}, \quad I_t \mapsto f_t \tag{5.4}$$

Finally, the extracted features are used to construct the memory bank for the video $\mathbf{V}^i = \{\mathbf{I}_t\}_{t=1}^{T} \in \mathbb{R}^{C\times T\times H\times W}$, represented as $\boldsymbol{f}^i = \{f_t\}_{t=1}^{T} \in \mathbb{R}^{T\times f_{\text{dim}}}$.

During each training epoch, a random temporal window $T_{\text{w}} = [t_{\text{start}}, t_{\text{end}}]$, with a predefined size T_{epoch}, is uniformly sampled from each video $\mathbf{V}^i$. From this window, B frames I_t are randomly selected and processed by the 2D neural network $\mathcal{F}$, being used to update $\boldsymbol{f}^i$ and produce a feature representation for the temporal window T_{w}, denoted as $\boldsymbol{f}^i_{\text{epoch}} = \{f_t\}_{t_{\text{start}}}^{t_{\text{end}}} \in \mathbb{R}^{T_{\text{w}}\times f_{\text{dim}}}$.

It is important to note that only the B sampled frames are used to compute gradients for $\mathcal{F}$, the features from other frames in the memory bank are exclusively utilized for training subsequent blocks in the pipeline. Additionally, there is a trade-off between B, the video batch size, and the spatial resolution of the video. While a sufficiently large B is essential for effectively training the Frame Encoder $\mathcal{F}$, setting B too high can negatively impact either the spatial resolution or the video batch size used to train the entire network.

Our Frame Encoder $\mathcal{F}$ is implemented using the Swin Transformer v2 [161], specifically the swin_small_patch4_window7_224.ms_in22k variant from the `timm` library [261].

5.2.2.2 Video Self-Attention Module

Local Self-Attention

The Frame Encoder extracts frame features independently from one another. To combine information across these features, we first deploy an organ-specific diagnostic relevance module as defined by our hierarchical architecture.

This is achieved using features transformed by local self-attention multi-head transformer block, which incorporates a masking operation to restrict frame feature interactions to their immediate neighbors.

The operations performed are described below, where $\boldsymbol{f}^{i}_{\text{epoch}}$ is expressed as $\mathbf{X}_0 \in \mathbb{R}^{T_{\text{w}} \times f_{\text{dim}}}$ to simplify notation :

$$\mathbf{Q} = \mathbf{X}_0\mathbf{W}_Q, \quad \mathbf{K} = \mathbf{X}_0\mathbf{W}_K, \quad \mathbf{V} = \mathbf{X}_0\mathbf{W}_V \tag{5.5}$$

with dimensions adjusted for the h attention heads, where $d_k = \frac{f_{\text{dim}}}{h}$.

$$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{f_{\text{dim}} \times d_k}, \quad \mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{T_{\text{w}} \times d_k} \tag{5.6}$$

To constrain interactions between frames, a local attention matrix $\mathbf{M}_{\text{local}}$ is introduced. This mask assigns 0 to allowed positions, comprised of the l_w frames, and $-\infty$ (a very large negative value) to masked positions, restricting attention to neighboring frames. An example of this mask is shown in Figure 5.3(a).

$$\text{AttHead}_i = \text{softmax}\left(\frac{\mathbf{QK}^\top + \mathbf{M}_{\text{local}}}{\sqrt{d_k}}\right)\mathbf{V} \tag{5.7}$$

The outputs of the attention heads are concatenated, followed by the addition of a residual connection, layer normalization (omitted here for clarity), and projection through a linear layer $\mathbf{W}_O \in \mathbb{R}^{f_{\text{dim}} \times f_{\text{dim}}}$ to produce a new feature representation :

$$\mathbf{X_1} = \left[\text{Concat}\left(\text{AttHead}_0, \cdots, \text{AttHead}_h\right) + \mathbf{X_0}\right]\mathbf{W}_O \tag{5.8}$$

This new representation $\mathbf{X_1} \in \mathbb{R}^{T_{\text{w}} \times f_{\text{dim}}}$ encodes localized knowledge about each frame's surroundings while minimizing the differences between new features and features retrieved from the memory bank.

Organ Diagnostic Relevance Branch

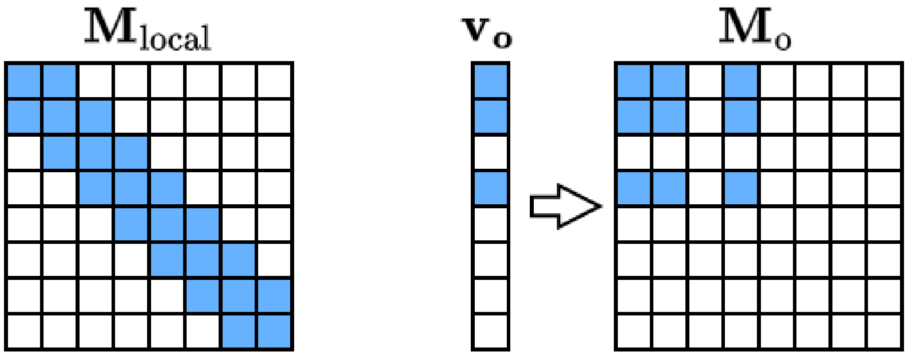
To compute diagnostic relevance scores for each frame and differentiate them from non-relevant frames (background), a linear projection followed by a softmax operation is applied :

$$\mathbf{S}_{\text{rel}} = \text{softmax}(\mathbf{X}_1\mathbf{W}_{\text{rel}} + \mathbf{b}_{\text{rel}}), \tag{5.9}$$

where O_n is the number of organs represented in $\mathbf{O}$, with dimensions defined as follows :

$$\mathbf{W}_{\text{rel}} \in \mathbb{R}^{f_{\text{dim}} \times (O_n+1)}, \quad \mathbf{b}_{\text{rel}} \in \mathbb{R}^{(O_n+1)}, \quad \mathbf{S}_{\text{rel}} \in \mathbb{R}^{T_{\text{w}} \times (O_n+1)} \tag{5.10}$$

The first O_n output channels in $\mathbf{S}_{\text{rel}}$ represent the relevance of each frame for diagnosing the corresponding organs, while the final channel represents the non-relevant (background) class. This configuration, combined with the softmax function, ensures that frames classified as non-relevant do not contribute to diagnosis estimation in subsequent steps.



(a) Local Attention Mask

(b) Organ Attention Mask

FIGURE 5.3 – Illustration of the various attention masks employed in this work. Blue cells represent permitted positions (assigned a value of zero), while white cells denote restricted positions (assigned a value of $-\infty$).

Organ-specific Self-Attention

To ensure the network focuses on all organ-specific frames while ignoring low-quality frames and frames belonging to other organs, we design an organ-specific attention mask using the organ diagnostic relevance scores $\mathbf{S}_{\text{rel}}$ obtained in the previous step.

For each organ o, excluding background dimension, we apply a threshold τ, resulting in a binary vector $\mathbf{v_o}$, where values 0 indicate the temporal position of organ-relevant frames for organ o, and 1 otherwise.

$$\mathbf{v_o} = \mathbf{S}_{\text{rel}}[:, o] > \tau, \quad \mathbf{v_o} \in \{0, 1\}^{T_{\text{w}}} \tag{5.11}$$

We compute an organ-specific attention mask, $\mathbf{M}$o, by calculating the dyadic product (outer product) of the vectors $\mathbf{vo}$, with ones inverted to zeros (not shown in the equation below).

$$\mathbf{M}_{\text{o}} = \mathbf{v_o} \otimes \mathbf{v_o} = \mathbf{v_o}\mathbf{v_o}^{\top} \tag{5.12}$$

The matrices $\mathbf{M}_{\text{o}}$ (one for each organ o) are then combined with element-wise multiplication ($\circ$ operator), and multiplied element-wise by a diagonal matrix $\mathbf{D}^{T_{\text{w}} \times T_{\text{w}}}$ with 0 in the principal diagonal and 1 elsewhere. finally, the resulting matrix $\mathbf{M}_{\text{organ}}$ is multiplied by $-\infty$ (a very large negative number). This operation is illustrated in the Figure 5.3(b).

$$\mathbf{M}_{\text{organ}} = -\infty \cdot \left(\mathbf{M}_{\text{0}} \circ \mathbf{M}_{\text{1}} \circ \cdots \circ \mathbf{M}_{\text{n}} \circ \mathbf{D}^{T_{\text{w}} \times T_{\text{w}}}\right) \tag{5.13}$$

In summary, $\mathbf{M}_{\text{organ}}$ forward individual frame features while allowing them to interact with all other frames belonging to the same organ (if any).

Finally, this masked is used in a new transformer block following the same equations describes in Equations 5.5 to 5.8.

$$\mathbf{X_2} = \text{TransformerBlock}\left(\mathbf{X_1}, \mathbf{M}_{\text{organ}}\right) \tag{5.14}$$

5.2.2.3 Hierarchical Classification Module

Finally, the frame features $\mathbf{X_2}$ are used to compute frame-level diagnosis for all N_c diagnosis defined in Equation 5.1. This is done by using a dedicated linear classifier for each combination of organ o diagnosis $\left\{h_{\text{o}}, p_{\text{o}}^{0}, p_{\text{o}}^{1}, \cdots, p_{\text{o}}^{k}\right\}$ weighted by the relevance scores $\mathbf{S}_{\text{rel}}$.

$$\hat{\mathbf{Y}}_{\mathbf{o}}^{frames} = \text{softmax}(\mathbf{X_2}\mathbf{W_o} + \mathbf{b_o}) \circ \mathbf{S}_{\text{rel}}\left[:, \text{o}\right] \tag{5.15}$$

The organ diagnosis scores for individual frames are concatenate, including the non-relevant/background scores $\mathbf{S}_{\text{rel}}\left[:, -1\right]$.

$$\hat{\mathbf{Y}}^{frames} = \text{Concat}\left(\hat{\mathbf{Y}}_{\mathbf{o}}^{frames} \mid \forall \text{o} \in \mathbf{O}, \mathbf{S}_{\text{rel}}\left[:, -1\right]\right) \tag{5.16}$$

Finally, the final video diagnosis score for each class c (including non-relevant or background) is computed as the mean of the top-R scores for that class.

$$\hat{\mathbf{Y}}^{i}\left(c\right) = \frac{1}{R} \sum_{t \in \text{top-}R\left(\hat{\mathbf{Y}}^{frames}\left[:, c\right]\right)} \hat{\mathbf{Y}}^{frames}\left[t, c\right] \tag{5.17}$$

This hierarchical classification strategy comprises several components that enable training the entire system in a weakly supervised manner :

- It explicitly uses the R highest contributing diagnosis scores to make a prediction for each class c, allowing these scores to be used as diagnostically relevant frames $\mathcal{K}\hat{y}k$ to support the diagnosis.
- All frame-level diagnosis estimations are weighted by the organ-relevance scores $\mathbf{S}_{\text{rel}}$, ensuring that frames not predicted as belonging to a specific organ are unlikely to contribute to the final diagnosis.

- It generates a video-level non-relevant (background) class estimation by averaging the $top - R$ values in $\mathbf{S}_{\text{rel}}[:, -1]$. This serves as guidance for training the non-relevant class, which is expected to be present in all videos.
- By relying solely on the top R frames from each video to make the final prediction, it is particularly well-suited for the classification of untrimmed videos, where only a limited number of frames are diagnostically useful.

5.2.3 Inference

To perform real-time inference from a video input, we extract frame features f_t using the Frame Encoder $\mathcal{F}$ and store them in a memory bank. Inference can be initiated at any time ; however, if the number of frames is fewer than the predefined T_w, zero-padding with temporal attention ($-\infty$ for padded frames) must be applied.

For sequences longer than T_w, and to maintain a fixed-size feature memory $\boldsymbol{f}$, when the $(T_w + 1)$-th frame arrives, it replaces the feature representation of the frame with the highest non-relevance score $\mathbf{S}_{\text{rel}}[:, -1]$. This allow us to keep the size of $\boldsymbol{f}$ fixed, while keeping the most relevant frames in the video.

Given that the Frame Encoder can operate in real-time—for example, using a RTX3090 TI, swin_small_patch4_window7_224.ms_in22k achieves inference in just 5 ms per image, and the transformer blocks and classification head require an average of 60 ms. This allows our system to perform near-real-time inference for live ultrasound CAD applications.

5.2.4 Dataset

Our dataset is an updated version of the MIM-US-107 Video Dataset described in Section 4.3.4.1, incorporating new videos acquired at our partner hospital. The updated distribution of video-level diagnoses is illustrated in Figure 5.4.

In this chapter, we grouped pathologies into the following classes to ensure an adequate number of samples in each class, given the availability of training data. This grouping is illustrated in Figure 5.1.

Class 0 (h_{L}) : Label for :

- liver_normal

Class 1 (p^0_{L}) : Label for :

- liver_steatosis
- liver_fibrosis

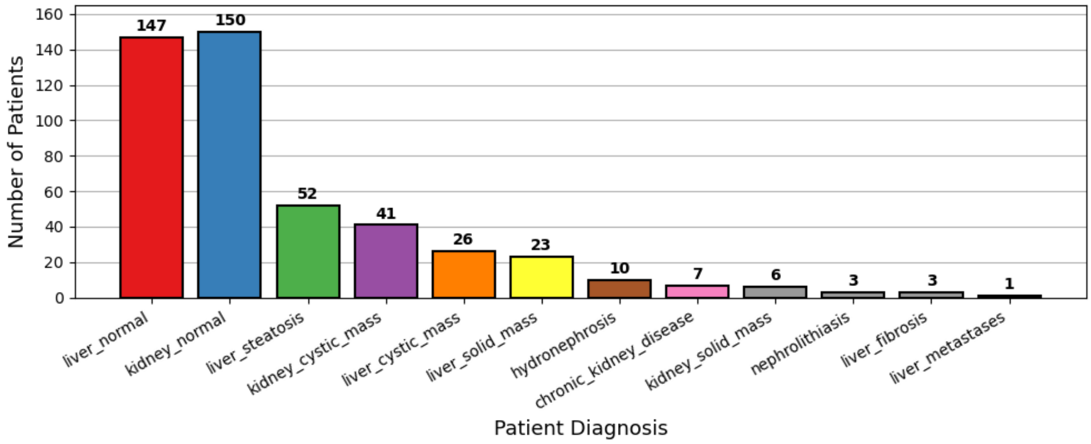


FIGURE 5.4 – Distribution of the video diagnosis dataset used in this study.

Class 2 (p_{L}^{1}) : Label for :

- liver_cystic_mass
- liver_solid_mass
- liver_metastases

Class 3 (h_{K}) : Label for :

- kidney_normal

Class 4 (p_{K}^{0}) : Label for :

- kidney_cystic_mass
- kidney_solid_mass
- chronic_kidney_disease
- nephrolithiasis
- hydronephrosis

The labels are grouped and assigned a value of 1 if the diagnosis is present in the video, and 0 otherwise. Additionally, for training purposes, we include the non-relevant class at the last position, which always has a label of 1 to indicate that all videos contain some non-relevant frames. Thus, the training labels for a video $\mathbf{V}^{i}$ are defined as follows :

$$\mathbf{Y}^{i} = \left\{h_{\mathsf{L}}, p_{\mathsf{L}}^{0}, p_{\mathsf{L}}^{1}, h_{\mathsf{K}}, p_{\mathsf{K}}^{0}, 1\right\} \tag{5.18}$$

We used 80% of the dataset for training and 20% for testing, ensuring an equal distribution of labels.

5.2.5 Training

5.2.5.1 Training Loss

The training loss used for our problem is the BCEWithLogitsLoss (Binary Cross Entropy with logits), which combines the binary cross-entropy loss with a sigmoid activation layer (σ). This loss is computed independently for each class and is defined as :

$$\begin{gathered}\ell(\mathbf{\hat{y}_c}, \mathbf{y_c}) = L_c = \{l_{1,c}, \dots, l_{N,c}\}^\top, \\ l_{n,c} = -\Big[p_c \cdot y_{n,c} \cdot \log \sigma(\hat{y}_{n,c}) + (1 - y_{n,c}) \cdot \log(1 - \sigma(\hat{y}_{n,c}))\Big]\end{gathered} \tag{5.19}$$

where $\mathbf{\hat{y}_c}$ and $\mathbf{y_c}$ represent the predicted and ground-truth values for class c, respectively. The weight value $\mathbf{p_c}$ is used to balance positive and negative samples, computed as the ratio of negative to positive instances.

The final loss is computed as the average of L_c across all classes, including the non-relevant class. Since lesions typically appear in only a few frames while the same video may also contain healthy-looking frames, we exclude the loss calculation for the healthy class in videos containing lesions for the respective organ.

5.2.5.2 Training parameters

Our proposed architecture includes some parameters specific to its design. The most important ones are listed below :

- $\mathbf{B}$: The number of frames sampled from each video to update the memory bank. In our experiments, we set this value to 24.
- $\mathbf{R}$: The predefined number of relevant frames per video. This value must balance two factors : it cannot be too large due to the limited number of diagnostically relevant frames per video, but if it is too small, it may affect training stability. In our experiments, we set $\mathbf{R}$ to 10 for all classes.
- $\mathbf{f}_{\text{dim}}$: The dimension of the frame-feature space, determined by the Frame Encoder used. In our case, it was set to 768.
- $\mathbf{T}_{\text{w}}$: The maximum number of frames stored in the memory bank. We set this parameter to 512.
- $\mathbf{l_w}$: The size of the local-attention window. We used a window size of 3 in our experiments.

- τ : The threshold used to compute the organ-attention mask. In our experiments, we set this value to 0.2, though it requires further investigation.

In addition, the image resolution was set to 224x224 pixels, and we applied the same image data augmentations described in the previous chapter. The network was trained with a learning rate of 10^{-5}, and early stopping was implemented after 200 epochs without improvement, with a minimum training duration of 300 epochs. Typically, training stops after approximately 1000 epochs, which takes around 48 hours on a RTX 3090 TI.

5.3 Results

5.3.1 Section overview

In this section, we present the results of our study. We focus on the following key aspects of our work :

1. **Video Diagnosis Classification :** Evaluating the performance of the proposed neural network in correctly estimating video-level diagnoses.
2. **Video Diagnosis Keyframe Localization :** Assessing the neural network's ability to identify keyframes that provide explainability for the diagnosis.
3. **Video Diagnosis Feasibility :** Determining the neural network's capability to assess whether the organ is visible with sufficient quality to enable reliable diagnosis estimation.

5.3.2 Video Diagnosis Performance and Keyframe Localization

Table 5.1 presents the numerical metrics for video diagnosis, while Figures 5.5 to 5.9 display the located keyframes supporting the diagnosis of each identified pathology. It is important to emphasize that, although the proposed model always outputs a set of supporting keyframes for each pathology, these keyframes are only meaningful when they correspond to pathologies with a high enough predicted score (denoted as "pred" in the figures). For instance, in Figure 5.5, the only valid keyframes are those for liver steatosis and a healthy kidney.

Additionally, the non-relevant class consistently has a predicted value of 1, which indicates low-quality frames, an inevitable occurrence during ultrasound acquisition. As shown in the figures, the model effectively detects ultrasound frames that either do not contain any organs or are heavily affected by shadows and artifacts.

Regarding the evaluation of video diagnosis performance, metrics such as Accuracy, Precision, Recall, and F1-Score require a threshold to convert continuous predicted scores into binary labels. These thresholds are computed by optimizing the F1-Score, as it provides a balanced trade-off between Precision and Recall.

Table 5.1 – Performance metrics for video diagnosis across different classes, including ROC-AUC, Accuracy, Precision, Recall, and F1-Score. The results highlight the model's strong performance for liver steatosis and reasonable outcomes for other liver conditions, with lower performance for kidney-related classes.

Class	**ROC-AUC**	**Acc.**	**Precision**	**Recall**	**F1**
Liver Healthy (h_L)	0.82	0.83	0.84	0.81	0.82
Liver steatosis (p_L^0)	0.97	0.94	0.92	0.85	0.88
Liver Lesion (p_L^1)	0.89	0.90	0.71	0.62	0.67
Kidney Healthy (h_K) :	0.66	0.69	0.80	0.57	0.67
Kidney Abnormal (p_K^0)	0.79	0.85	0.57	0.44	0.50

The results demonstrate that the network performs particularly well in detecting steatosis, achieving a ROC-AUC of 0.97 and an accuracy of 0.94. It is important to note that these values differ fundamentally from those reported in the previous chapter, as the evaluation in this case distinguishes between liver steatosis and non-steatosis cases, rather than liver steatosis versus healthy liver.

In the second row of Figures 5.5 and 5.9, we present the automatically selected keyframes used to support the diagnosis of liver steatosis from untrimmed ultrasound videos. As observed, both sets of selected images display a bright liver parenchyma and attenuation of the ultrasound wave, which are common findings associated with the diagnosis of steatosis.

The network demonstrated promising performance for the liver healthy and liver lesion classes, achieving ROC-AUC and accuracy values between 0.82 and 0.90. Figure 5.6 illustrates the automatically detected support keyframes for identifying a liver mass (third row), while Figure 5.7 presents the keyframes supporting the diagnosis of a healthy liver. Although there remains significant room for improvement, the approach produces reasonable results and effectively localizes lesions.

However, the model underperformed for the kidney classes, achieving AUC values of 0.79 and 0.66 for the kidney abnormal and kidney healthy classes, respectively. A well-performing example is shown in Figure 5.7.

We hypothesize that the unsatisfactory performance for the kidney classes is due to three main reasons. First, as shown in Figure 5.8, there is confusion between shadowing artifacts and kidney lesions. Dense lesions, such as solid masses and kidney stones (nephrolithiasis), can produce shadowing similar to that caused by artifacts. Second, some kidney frames were not identified by the radiologist during the examination, as illustrated in Figure 5.9. Finally, in an effort to increase the number of samples in the pathological classes, we grouped a wide range of kidney pathologies with different anatomical findings into the kidney abnormal class. This may have introduced excessive variability within the class, negatively impacting performance.

5.3.3 Diagnostic Feasibility Performance

To compute the feasibility of a diagnosis for a given organ, we combine the ground-truth labels and predicted scores for all diagnoses associated with that organ. Specifically, if the radiologists were able to make a diagnosis for the organ, the feasibility ground-truth label is assigned a value of 1. The predicted feasibility score is then calculated as the average of all diagnosis predicted scores for that organ.

Table 5.2 – Diagnostic feasibility results for liver and kidney, demonstrating satisfactory performance and supporting the hierarchical classification structure.

Class	**ROC-AUC**	**Acc.**	**Precision**	**Recall**	**F1**
Liver Diagnostic Feasibility	0.90	0.94	0.98	0.96	0.97
Kidney Diagnostic Feasibility	0.91	0.85	0.89	0.89	0.89

The results for liver and kidney diagnostic feasibility are presented in Table 5.2. Both provide satisfactory outcomes, with liver feasibility scores excelling. These results indirectly indicate that our hierarchical classification structure for predicting liver and kidney diagnostically relevance scores functions as intended.

Figure 5.8 illustrates a case where feasibility scores are particularly useful. In this example, the video does not contain liver images, resulting in a liver diagnostic feasibility score of 0. This informs the operator that additional images are required to produce a precise diagnosis.

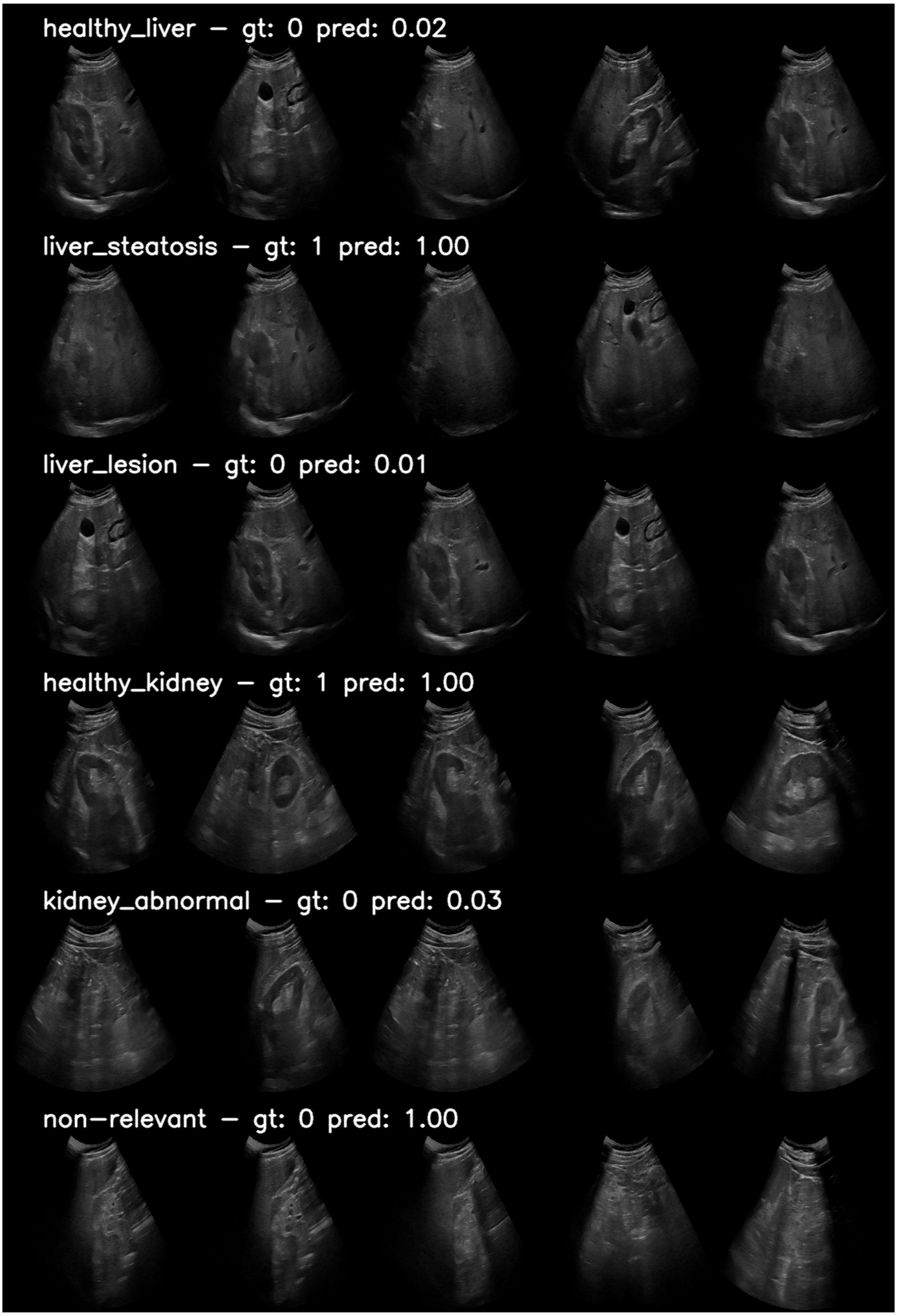

healthy_liver – gt: 0 pred: 0.02
liver_steatosis – gt: 1 pred: 1.00
liver_lesion – gt: 0 pred: 0.01
healthy_kidney – gt: 1 pred: 1.00
kidney_abnormal – gt: 0 pred: 0.03
non–relevant – gt: 0 pred: 1.00

FIGURE 5.5 – Diagnostic keyframes localized in a weakly-supervised manner by the proposed network, highlighting a patient with liver steatosis and a healthy kidney.

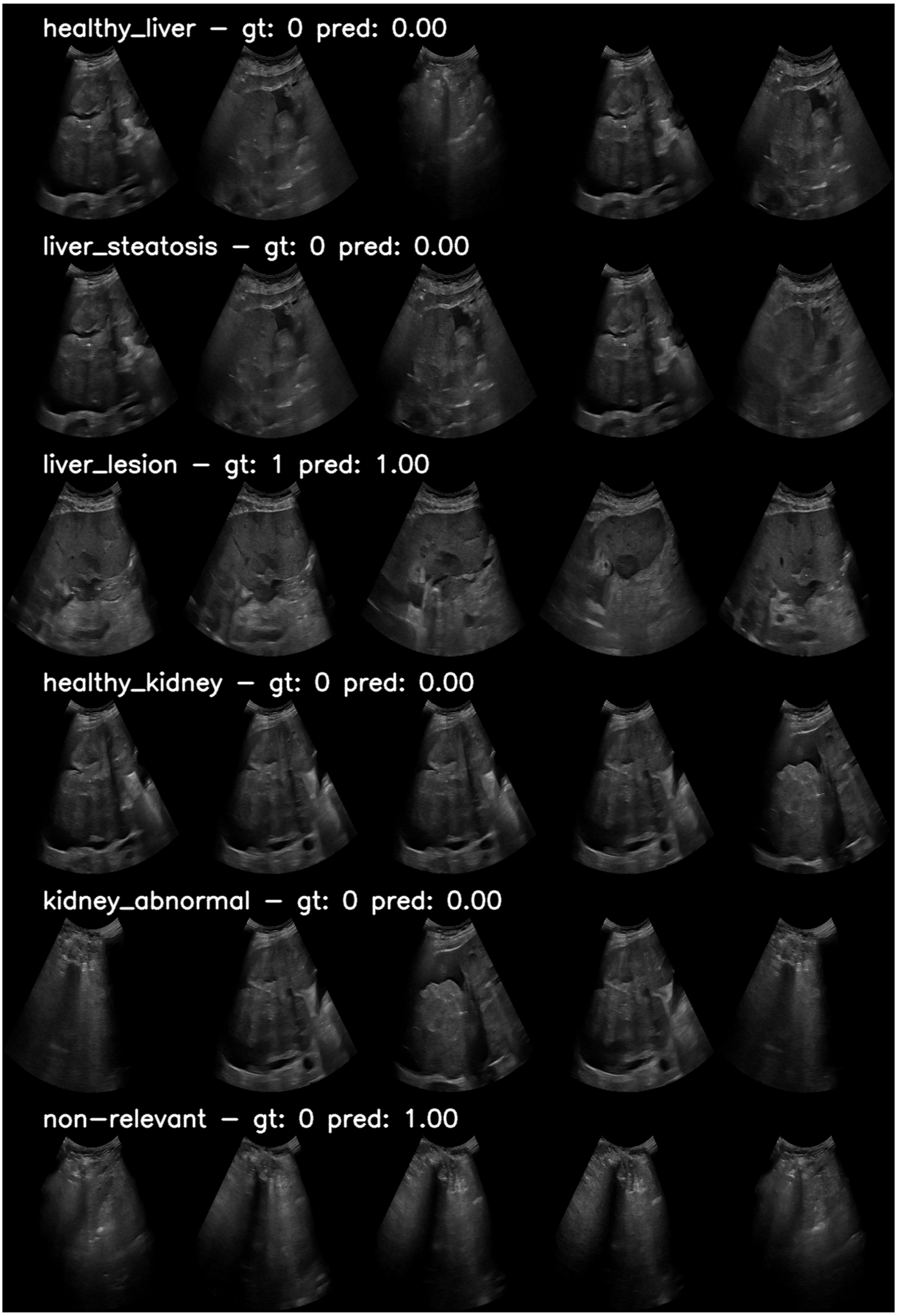

healthy_liver – gt: 0 pred: 0.00
liver_steatosis – gt: 0 pred: 0.00
liver_lesion – gt: 1 pred: 1.00
healthy_kidney – gt: 0 pred: 0.00
kidney_abnormal – gt: 0 pred: 0.00
non–relevant – gt: 0 pred: 1.00

Figure 5.6 – Diagnostic keyframes localized in a weakly-supervised manner by the proposed network, focusing on a patient with a liver lesion and insufficient information for a kidney diagnosis.

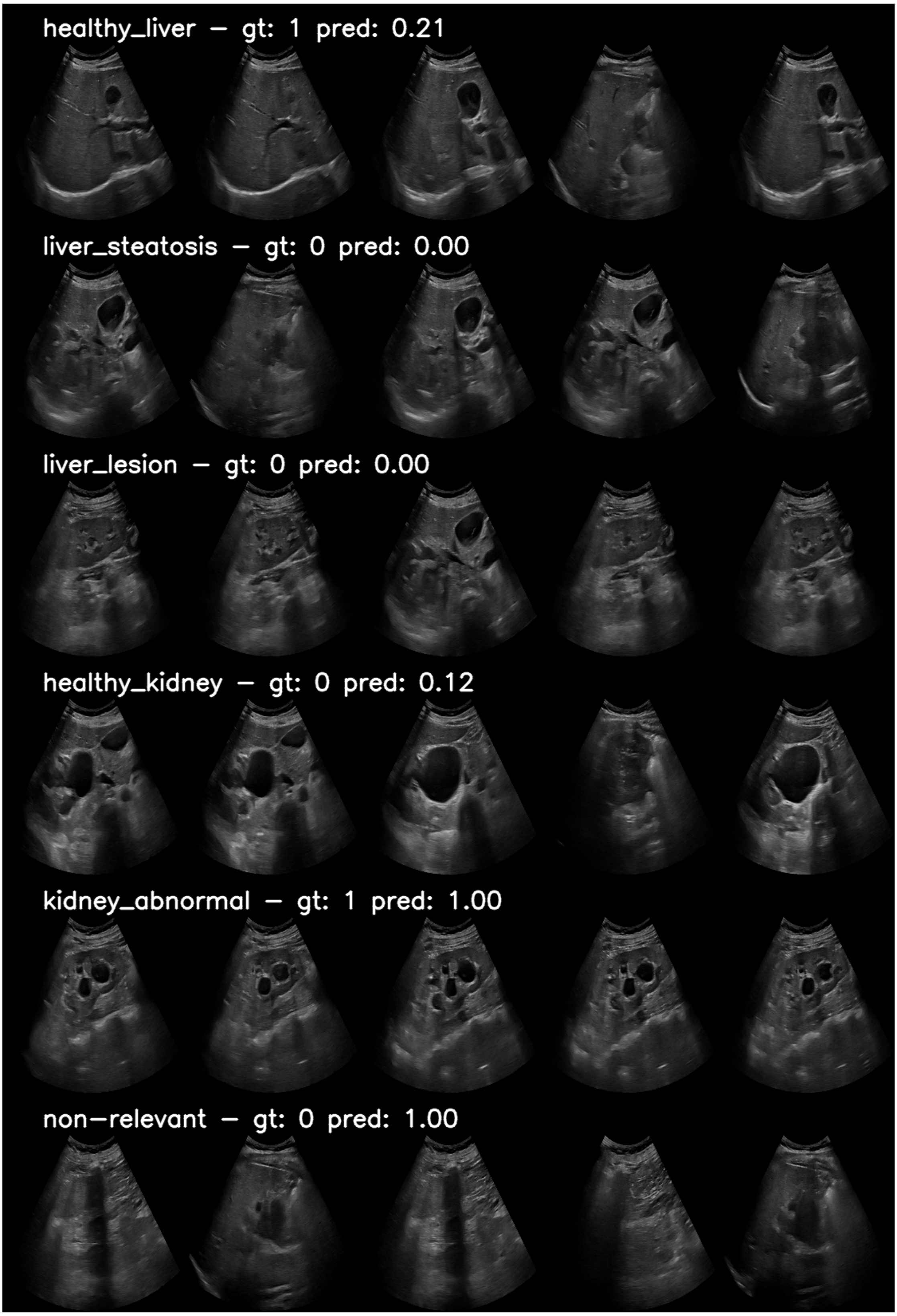

healthy_liver – gt: 1 pred: 0.21
liver_steatosis – gt: 0 pred: 0.00
liver_lesion – gt: 0 pred: 0.00
healthy_kidney – gt: 0 pred: 0.12
kidney_abnormal – gt: 1 pred: 1.00
non–relevant – gt: 0 pred: 1.00

FIGURE 5.7 – Diagnostic keyframes localized in a weakly-supervised manner by the proposed network, focusing on a patient with kidney abnormalities and a healthy liver detected with low confidence.

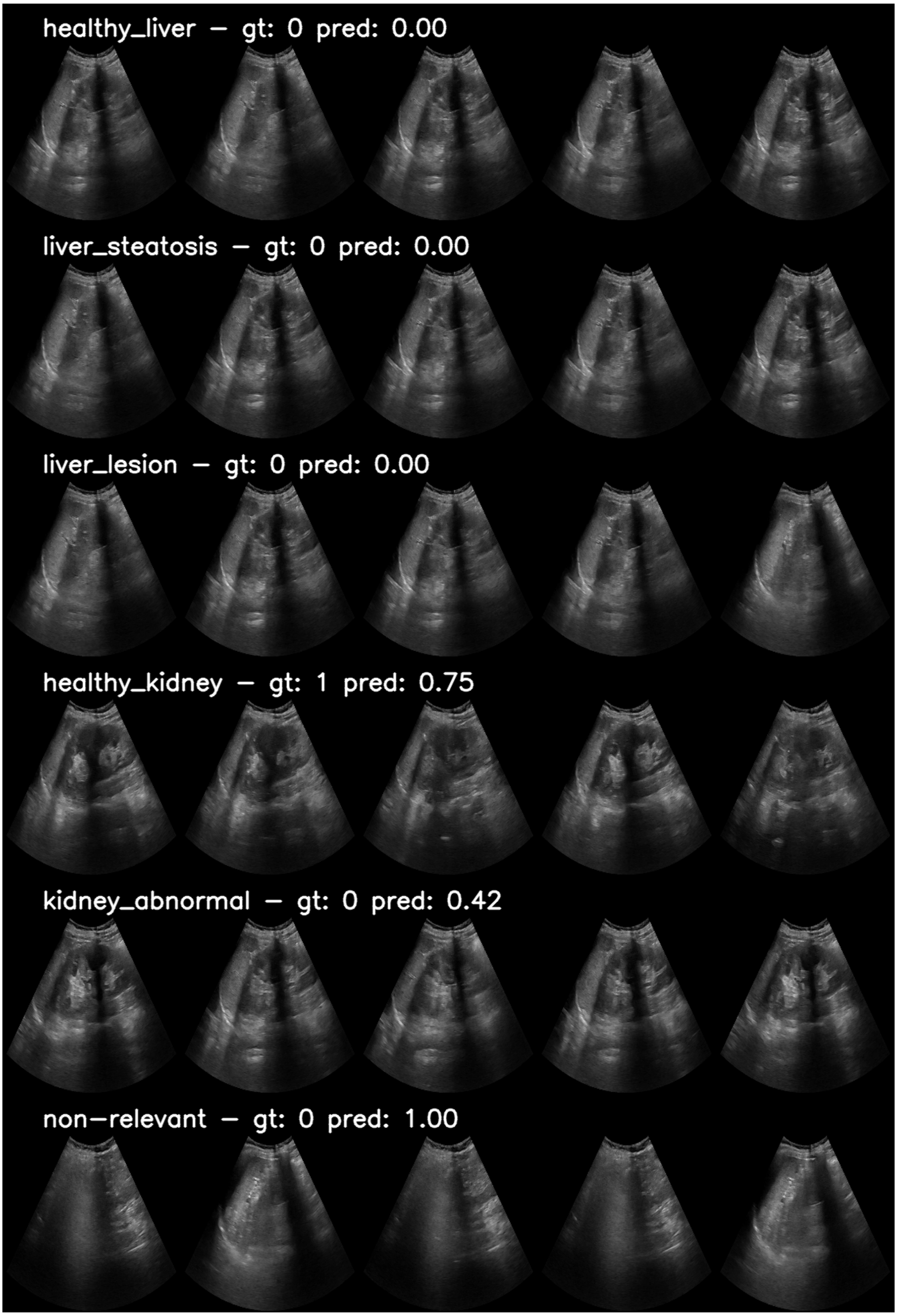

healthy_liver – gt: 0 pred: 0.00
liver_steatosis – gt: 0 pred: 0.00
liver_lesion – gt: 0 pred: 0.00
healthy_kidney – gt: 1 pred: 0.75
kidney_abnormal – gt: 0 pred: 0.42
non–relevant – gt: 0 pred: 1.00

FIGURE 5.8 – Diagnostic keyframes localized in a weakly-supervised manner by the proposed network, highlighting the negative effect of shadowing, which is confused with kidney abnormalities.

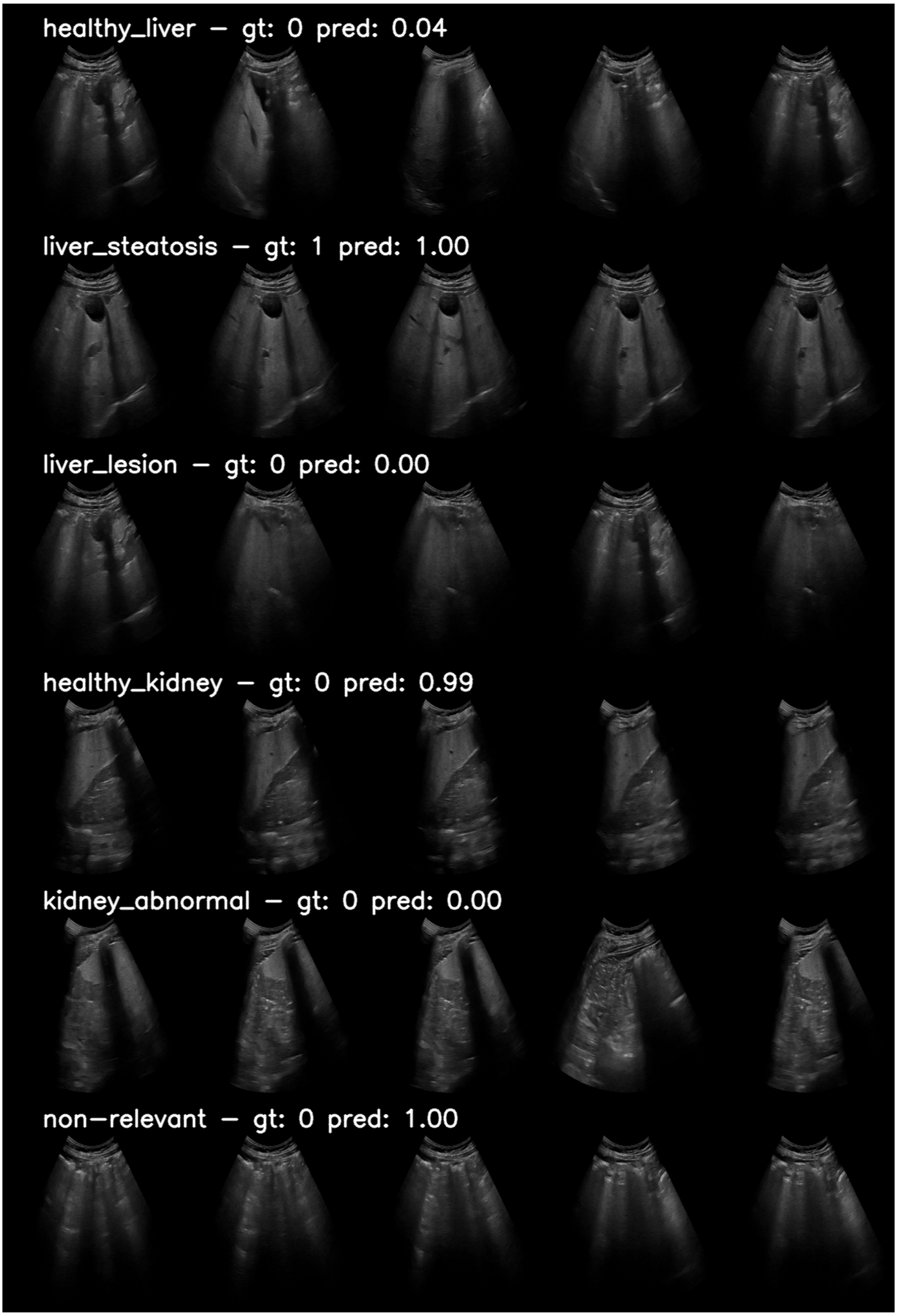

healthy_liver – gt: 0 pred: 0.04
liver_steatosis – gt: 1 pred: 1.00
liver_lesion – gt: 0 pred: 0.00
healthy_kidney – gt: 0 pred: 0.99
kidney_abnormal – gt: 0 pred: 0.00
non–relevant – gt: 0 pred: 1.00

FIGURE 5.9 – Diagnostic keyframes localized in a weakly-supervised manner by the proposed network, highlighting kidney frames which were not identified by the radiologist during annotation.

5.4 Conclusion and Future Work

In this chapter, we introduced KeyFrameDiagFormer, a novel weakly-supervised keyframe localization and diagnosis transformer model designed for untrimmed ultrasound videos. Trained exclusively on video-level labels, the model autonomously identifies diagnostically relevant keyframes to support diagnosis, aligning seamlessly with standard clinical practices. This approach could also potentially uncover new pathological visual findings in ultrasound by training the model with labels derived from other diagnostic modalities, such as biopsy, CT, or MRI.

Through its hierarchical multi-label classification framework and the integration of background modeling, organ-specific attention mechanisms, and long-memory frame embedding, KeyFrameDiagFormer demonstrates robust performance in diagnosing pathologies and identifying diagnostically relevant frames. The model excels in detecting liver steatosis and other liver pathologies, showcasing its effectiveness in complex diagnostic tasks. Although its performance on kidney-related conditions is less consistent, it highlights opportunities for refinement and further optimization.

While our approach showed promising results, there are multiple areas which required further investigation :

- **Optimization of KeyFrameDiagFormer Parameters :** Time constraints limited our ability to comprehensively explore the parameter space of KeyFrameDiagFormer, including variations in the Frame Encoder design, the number and configuration of Transformer blocks, and the optimal number of frames for video diagnosis. Determining the best configuration is essential, especially for scaling the model to include more organs or pathologies. Future work should prioritize systematic tuning of these parameters and assess their scalability for broader diagnostic tasks.
- **Incorporating Loss Functions from WTAL :** Weakly supervised action localization commonly utilizes diverse loss functions to enhance the distinction between background and action frames. Exploring how these loss functions interact with our model on ultrasound data could yield valuable insights and performance improvements.
- **Leveraging Multimodal Training Data :** The independent frame feature representations in KeyFrameDiagFormer enable training with single-image data, expanding its flexibility. Furthermore, the temporal feature representations in the transformer blocks allow the model to incorporate labels from other modalities, such as text or patient clinical records. Future research could investigate training or pre-training with multimodal data to improve access to

labeled datasets and enhance the model's generalization across diagnostic tasks.

6. Conclusion

In this work we addressed key challenges associated with developing **deep learning models for US-based CAD targeting liver and kidney pathologies in untrimmed b-mode ultrasound videos**. This effort is crucial given the global shortage of experienced radiologists and the inherent difficulties in acquiring and diagnosing ultrasound data. To address these issues, we focused on creating fully automated solutions that are user-friendly for novices and non-expert healthcare providers, thereby alleviating critical bottlenecks in the medical care pipeline.

In the literature, most automated diagnosis methods rely on single-image classification approaches. While these methods demonstrate promising performance, they fail to represent the true data distribution of real-world ultrasound videos. As a result, such systems require operators to manually select high-quality frames compatible with the model, a task that demands significant expertise in ultrasound examination. This limitation restricts the democratization of ultrasound diagnosis, making these methods inaccessible to non-experts. Furthermore, the few studies that utilize video classification models for diagnosing abdominal pathologies often depend on trimmed videos, which suffer from similar limitations. Some approaches attempt to overcome these issues by employing strong supervision, such as segmentation masks or bounding boxes. However, creating these annotations is time-consuming and resource-intensive, further hindering scalability and widespread adoption.

Another problem addressed in this work is the evaluation of the **reliability of standard visual annotations provided by annotators**, which are commonly used for training and developing CAD systems in ultrasound. Due to the inherent challenges of ultrasound diagnosis, these annotations often involve a degree of subjectivity, which can negatively affect the performance and robustness of deep learning models trained on such data.

We have proposed several methods to overcome these challenges, which we summarize below :

Key Contributions

In our first contribution, presented in Chapter 3 and titled **CVL+RankNet**, we introduced a novel methodology for labeling diagnoses in medical images. This approach improves upon traditional single-image visual labeling (SVL) by incorporating comparative annotations, where annotators indicate which of two images exhibits a higher perceived degree of pathology. This comparative process reduces subjectivity in labeling, as it eliminates the need to establish explicit boundaries between healthy and pathological diagnoses, a task that often varies significantly among annotators. Furthermore, by employing a Learn-to-Rank formulation, we can rank images from the most to the least pathological, generating real-valued pathological scores that show strong correlation with histopathological findings.

In our experiments, presented in Table 3.7, we demonstrate that our proposed method significantly improves annotation reliability for all annotators compared to the standard single-image visual labeling (SVL) approach when using histopathology labels as ground-truth. When evaluating fused labels, our method achieves a 10% superiority over SVL, highlighting its effectiveness in enhancing consistency and accuracy in diagnostic labeling in ultrasound. This work addresses a critical issue regarding the accuracy of visual labels, which are typically considered the gold standard and used as ground truth in training and evaluation. Developing methods to reduce labeling errors is essential for advancing AI-based CAD systems in ultrasound, as reliable annotations are foundational to their effectiveness.

In our second contribution, presented in chapter 4 and named **DR-Clips**, we proposed a method for classifying untrimmed ultrasound videos using a pragmatic external guidance agent. This was achieved by developing an external frame-guidance neural network capable of assigning diagnostic relevance scores to individual ultrasound frames extracted from input videos. These relevance scores were then utilized to create DR-Clips, which are clips composed of sequential, highly diagnostically relevant frames. These DR-Clips are used both for inference and as an effective data augmentation tool, enhancing the performance and robustness of the classification model.

Our approach outperformed traditional end-to-end video classification models and guided single-image classification models, achieving performance comparable to models trained on manually selected images curated by expert radiologists. This highlights the limitations of these models when applied to untrimmed ultrasound data, an issue that may go unnoticed in video diagnosis research. Developing methods to enable the deployment of such models on untrimmed ultrasound data is a crucial area of study, as it bridges the gap between experimental settings and real-world applications, ensuring these video classification networks can be effectively utilized in clinical practice.

In our final contribution, presented in Chapter 5 and titled **KeyFrameDiagFormer**, we introduced a model capable of performing diagnosis on untrimmed ultrasound videos while simultaneously identifying diagnostically relevant keyframes in a weakly supervised manner. Our approach incorporated several novel insights for the diagnosis of ultrasound videos, including the design as multi-label problem, the adoption of background class, a concept commonly used in weakly supervised action recognition, a memory bank for processing long sequences efficiently, and a hierarchical structure. These designs choices allow the model to indicate when there is not enough frames with sufficient image quality to realize the diagnosis of a given organ by providing organ-specific frame relevance scores which are trained indirect from pathological labels.

The proposed approach achieved impressive results in diagnosing liver steatosis and satisfactory performance in diagnosing liver lesions, successfully identifying keyframes where pathologies are clearly visible and offering valuable support for clinical diagnoses. While there is still room for refinement, the method provides several significant advantages that establish it as a strong candidate for general-purpose video diagnosis in ultrasound. First, the model is trained exclusively on video-level labels, which can often be derived automatically from patient clinical records, reducing the need for extensive manual annotation. Second, it effectively highlights keyframes containing anatomical features that align closely with those used by radiologists in standard screening practices. Third, the model can assess diagnosis feasibility, providing real-time feedback when a video lacks sufficient high-quality frames to make a reliable diagnosis. Lastly, the integration of a memory bank ensures efficient real-time inference, a crucial feature for deploying video diagnosis models in practical clinical settings.

The contributions presented in this work are designed to serve as a foundation for the diagnosis of abdominal pathologies in real-world conditions. Our proposed methods address critical challenges in real-world data, either by reducing the inherent errors in visual annotations or by enabling diagnosis in untrimmed ultrasound videos, which often contain a high proportion of non-relevant frames. We hope that this work inspires further research focused on developing robust solutions adapted to real-world ultrasound data, ultimately advancing the field of AI-based CAD.

Implications

Our contributions support the democratization of ultrasound screening by reducing the expertise required to perform and interpret examinations, thereby increasing the frequency with which patients can be screened. This enables expert radiologists to focus on suspicious or complex cases, optimizing healthcare workflows and enhancing the overall efficiency of diagnostic processes.

This is the proposition of the Disrumpere Project, developed by the R&D teams of IRCAD France and IRCAD Rwanda. The goal is to leverage affordable portable ultrasound devices and artificial intelligence to expand access to diagnostic imaging, particularly in underserved regions. By automating key aspects of image acquisition, interpretation, and diagnosis, the project aims to empower non-experts and healthcare providers with limited training to perform reliable screenings. This approach not only addresses the scarcity of skilled radiologists but also ensures timely detection and monitoring of common pathologies, ultimately reducing healthcare disparities and improving outcomes globally.

In the context of the Disrumpere Project, we have integrated a simplified version of our models for the automatic diagnosis of liver steatosis into the IRCAD Therapus software, utilizing a low-cost ultrasound probe. Figures 6.1 and 6.2 illustrate examples of AI-powered steatosis screening performed by non-experts, showcasing a pathological and a healthy case, respectively.

At the beginning of the screening (Figures 6.1(a) and 6.2(a)), the system determines that there are not yet enough diagnostically relevant images of sufficient quality to provide a diagnosis. Once the first diagnostically relevant frames are automatically identified, the system generates an initial low-confidence diagnosis (Figures 6.1(b) and 6.2(b)). As additional high-quality, diagnostically relevant images are detected, the system's confidence in the diagnosis increases, resulting in a final, reliable diagnosis (Figures 6.1(c) and 6.2(c)).

These experiments highlight the potential implications of our contributions , which can reduce healthcare costs through early pathology detection and expand access to screening for individuals in underserved or low-resource settings.

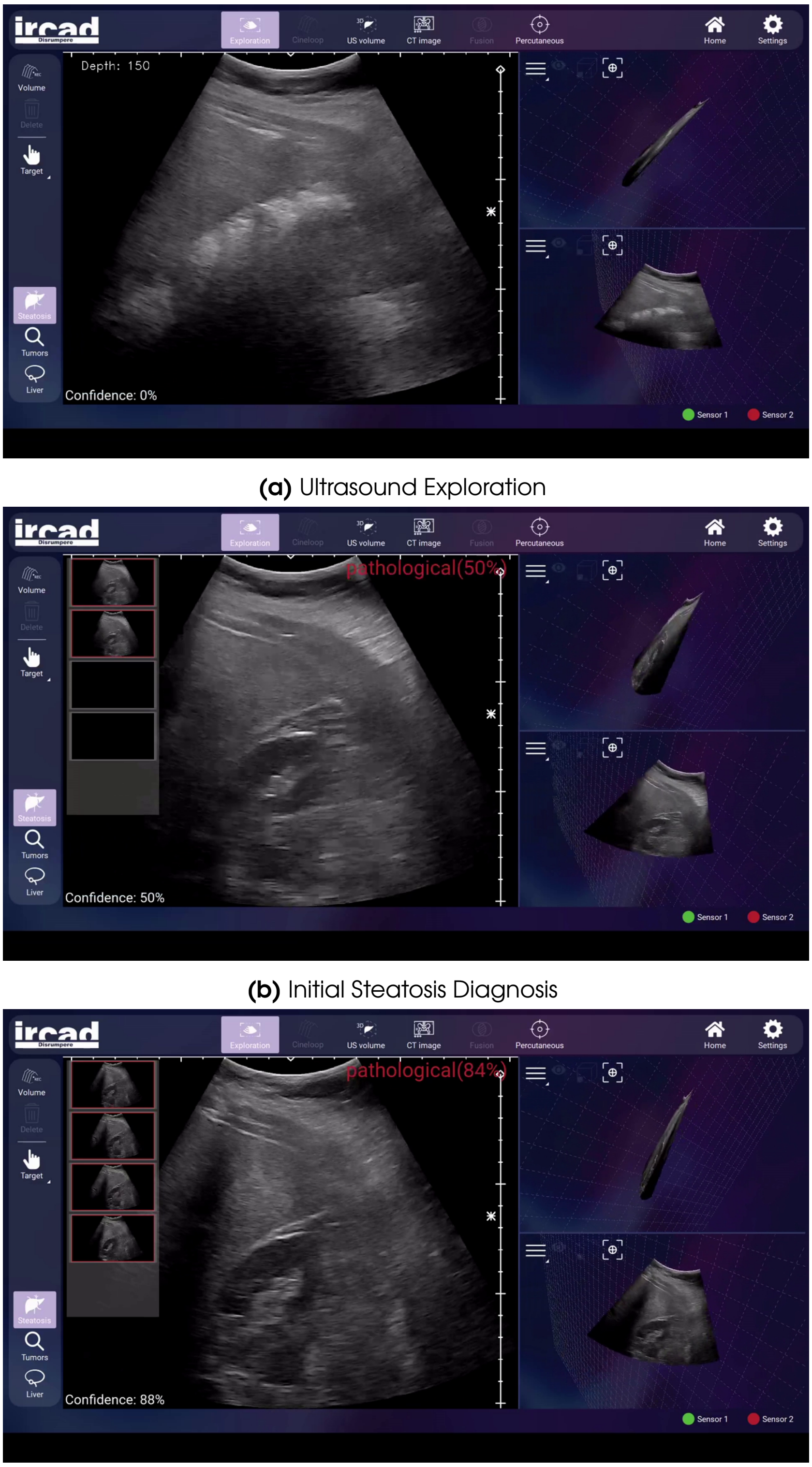
Limitations and Future Research Directions

While our research contributions efficiently address the challenges of developing deep learning models for diagnosis of ultrasound data, some limitations persist, highlighting areas for refinement and potential directions for future research.

A primary limitation across our methods is the need for validation on a broader range of pathologies, datasets, and devices. Expanding this scope would enhance the generalizability and robustness of our findings. Below, we outline specific limitations and future research directions related to the use of CVL annotations and the diagnosis of untrimmed ultrasound videos :

Concerning the use of **Comparative Visual Labeling** described in Chapter 3 :

- **Annotation Effort and Active Learning :** CVL requires a greater number of annotations compared to standard single-image visual labeling (SVL). Leveraging active learning techniques [176, 122] could help reduce this burden by prioritizing the most informative samples for labeling.
- **Siamese Network Potential :** Our research work focused on using CVL to



Depth: 150

Confidence: 0%

(a) Ultrasound Exploration

pathological(50%)

Confidence: 50%

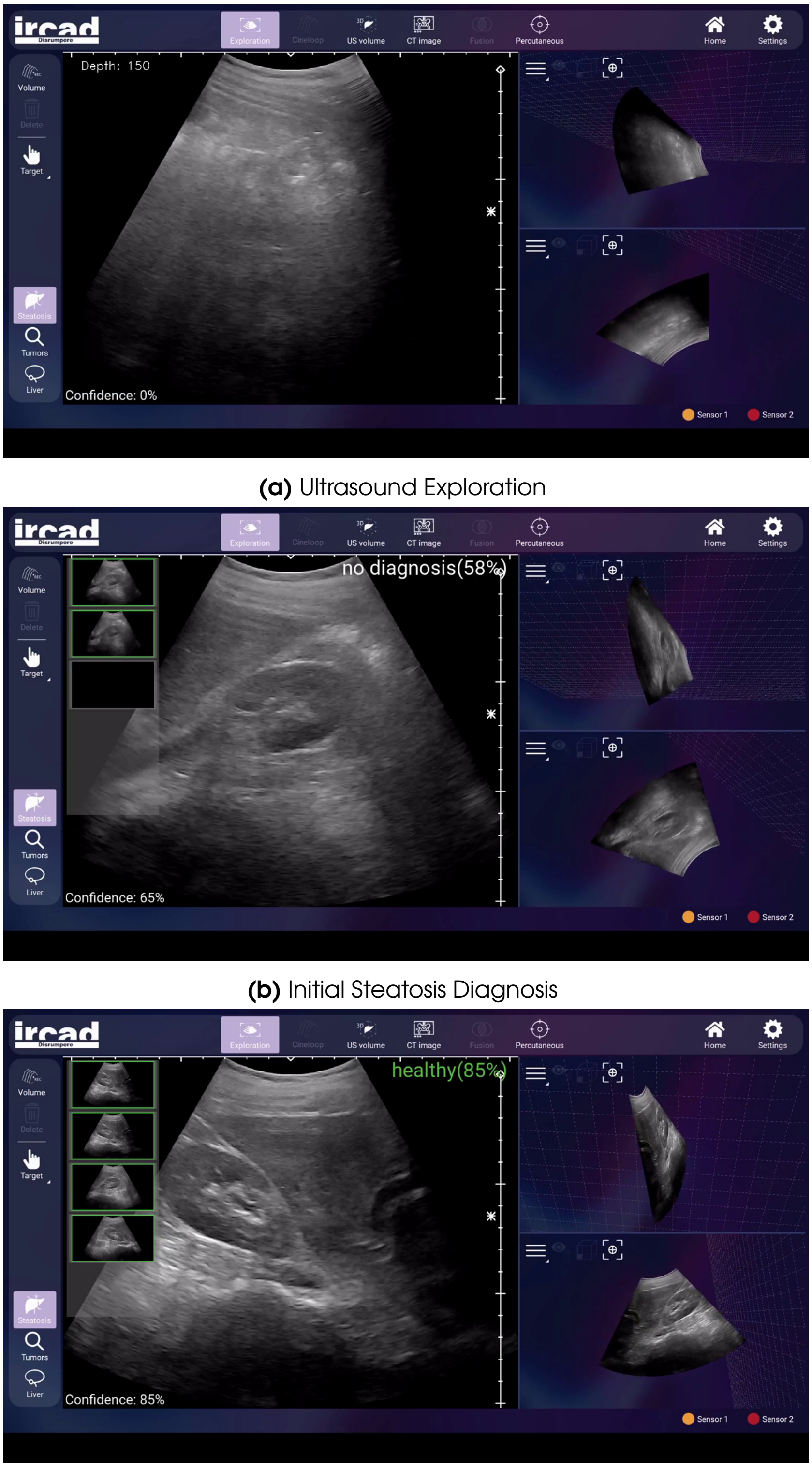
(b) Initial Steatosis Diagnosis

pathological(84%)

Confidence: 88%

(c) Final Steatosis Diagnosis

FIGURE 6.1 – AI-powered liver steatosis screening in a pathological case, showcasing the progression from (a) initial ultrasound exploration, (b) low-confidence diagnosis, to (c) final diagnosis as diagnostically relevant frames are identified. Images are generated using IRCAD's proprietary software, Therapus.



(a) Ultrasound Exploration

(b) Initial Steatosis Diagnosis

(c) Final Steatosis Diagnosis

FIGURE 6.2 – AI-powered liver steatosis screening in a healthy case, illustrating the process from (a) initial ultrasound exploration, (b) low-confidence diagnosis, to (c) final confirmation as diagnostically relevant frames are analyzed. Images are generated using IRCAD's proprietary software, Therapus.

enhance single-image annotations, but CVL could also be applied directly to train siamese neural networks with contrastive loss. This could enable continuous-valued pathological scoring, offering fine-grained information, allowing to predicted the evolution of the condition.

— **Video-Level CVL :** In our work, CVL was not applied to video-level data due to the lack of histopathological results for validation. Future research could explore inter-video and intra-video CVL strategies to enhance video model training.

Concerning the **diagnosis of untrimmed ultrasound videos** discussed in Chapters 4 and 5 :

— **Video Classifier in DR-Clips and Inference :** While DR-Clips were tested with the Video Swin Transformer, evaluating their impact on other video classifiers remains important. Additionally, we observed performance degradation with large DR-Clips during inference, suggesting difficulties in handling low-relevance clips. Addressing these challenges and proposing solutions are crucial areas for future research.

— **Parameters in KeyFrameDiagFormer :** Due to time constraints, we were unable to fully explore the design space of KeyFrameDiagFormer, including variations in the Frame Encoder architecture, the number and configuration of Transformer blocks, and the optimal number of frames used for final video diagnosis. Identifying the ideal configuration is crucial, particularly when scaling the model to accommodate additional organs or pathologies. Future research should focus on systematically tuning these parameters and evaluating their adaptability to broader diagnostic tasks.

— **Pre-training of Ultrasound Video Transformer Models :** Self-supervised pre-training is crucial for maximizing the potential of Transformer models [191]. Incorporating this step into our framework could enhance performance and enable the use of deeper Transformer architectures.

— **Exploration of other losses from WTAL :** Weakly supervised action localization often employs multiple loss functions to better distinguish between background and action frames. Investigating how these loss functions interact with our proposed model in ultrasound data could provide valuable insights.

— **Training with Multimodal data :** KeyFrameDiagFormer's independent frame feature representations allow it to be trained using single-image data, increasing its versatility. Additionally, the temporal feature representations generated in the final transformer blocks allow the model to use other modalities of labels, such as text data or patient clinical records. This capability could be explored for model training or pre-training, potentially improving access to labeled data and enhancing the model´s generalizability.

Final Thoughts

This thesis has addressed key challenges in the development of deep learning-based CAD systems for ultrasound data, emphasizing the development of models capable to be deployed in real-world untrimmed video data. By reducing reliance on extensive annotations and external guidance, and enabling diagnosis in untrimmed videos, our contributions aim to bridge the gap between experimental research and clinical practice. We hope this work not only enhances current methodologies but also motivates future research to focus on fully automated solutions for ultrasound diagnostics. Ultimately, these advancements have the potential to democratize healthcare, reduce disparities, and significantly improve patient outcomes, particularly in developing and underserved regions.

A. Appendix A

Table A.1 – Training hyperparameters of our proposed model and others baselines.

Model name	Hyperparameter	Value
Video Swin + DR-Clips (ours)	pretrained_model	swin_small_patch4_window7_224_22k.pth
	patch_size	[4, 4, 4]
	embed_dim	96
	depths	[2, 2, 18, 2]
	num_heads	[3, 6, 12, 24]
	mlp_ratio	4
	qkv_bias	True
	drop_rate	0
	attn_drop_rate	0.1
	batch_size	12
	learning_rate	10^{-6}
Resnet (Manual)	pretrained	ImageNet
	dropout	0.5
	batch_size	96
	learning_rate	10^{-5}
Resnet (Auto)	pretrained	ImageNet
	dropout	0.5
	batch_size	48
	learning_rate	10^{-4}
Swin (Auto)	pretrained_model	swin_small_patch4_window7_224_22k.pth
	batch_size	80
	learning_rate	10^{-6}
BabyNet	Resnet (Manual)	0.96 (0.11)
	msha3D	True
	batch_size	8
	learning_rate	10^{-5}
EchoGNN	video_encoder out_channels	[16, 32, 64, 128, 256]
	video_encoder kernel_sizes	3
	video_encoder pool_sizes	2
	video_encoder output_dim	256
	video_encoder cnn_dropout_p	0.1
	video_encoder fc_dropout_p	0.5
	attention_encoder fc_dropout_p	0.5
	attention_encoder hidden_dim	128
	graph_regressor gnn_hidden_dims	[128, 64, 32]
	graph_regressor fc_hidden_dim	16
	graph_regressor dropout_p	0.5
	batch_size	32

	learning_rate	10^{-5}
UVT	latent_dim	1024
	intermediate_size	1024
	num_hidden_layers	1
	attention_heads	4
	batch_size	8
	learning_rate	10^{-6}
Video Swin	pretrained_model	swin_small_patch4_window7_224_22k.pth
	patch_size	[4, 4, 4]
	embed_dim	96
	depths	[2, 2, 18, 2]
	num_heads	[3, 6, 12, 24]
	mlp_ratio	4
	qkv_bias	True
	drop_rate	0
	attn_drop_rate	0.1
	batch_size	12
	learning_rate	10^{-6}
VideoMAE V2	model_name	vit_small_patch16_224
	pretrained_model	vit_s_k710_dl_from_giant.pth
	batch_size	8
	learning_rate	$2 * 10^{-6}$
MViTv2	mvit_version	mvit_v2
	batch_size	8
	learning_rate	10^{-5}
X3D	mvit_version	x3d_m
	batch_size	18
	learning_rate	$4 * 10^{-6}$

B. Résume Long Français

Introduction

L'imagerie médicale joue un rôle primordial dans la détection et la surveillance de nombreuses pathologies, qu'il s'agisse d'infections, de lésions tumorales ou de maladies chroniques. Parmi les différentes modalités existantes, l'échographie (en particulier l'échographie b-mode) se démarque par plusieurs avantages qui en font un outil de choix pour un dépistage à large échelle :

- **Coût réduit et accessibilité :** L'équipement échographique est généralement bien moins onéreux que la tomodensitométrie (CT) ou l'IRM, et la portabilité croissante des dispositifs d'échographie permet de réaliser des examens hors des centres hospitaliers traditionnels.
- **Visualisation en temps réel :** L'observation directe et instantanée des structures internes est particulièrement précieuse pour guider certaines interventions (biopsies, pose d'aiguille, etc.).
- **Sécurité pour le patient :** Contrairement à d'autres techniques, l'échographie n'expose pas à des rayonnements ionisants et ne requiert qu'exceptionnellement l'injection d'un produit de contraste, minimisant ainsi les risques et effets secondaires.
- **Efficacité pour les tissus mous :** Dans le cadre de pathologies hépatiques ou rénales, l'échographie offre souvent une sensibilité élevée pour détecter des anomalies, masses ou fluides.

Malgré ces avantages, l'échographie présente également des limites et des défis qui restreignent son adoption à grande échelle :

- **Qualité d'image limitée :** L'omniprésence d'artefacts (ombrage, bruit, flou de mouvement, etc.) complique l'interprétation et réduit la fiabilité diagnostique.
- **Forte dépendance à l'opérateur :** Le résultat de l'examen dépend largement de l'habileté et de l'expérience de la personne manipulant la sonde.
- **Subjectivité du diagnostic :** Les radiologues, même expérimentés, peuvent interpréter différemment les mêmes images selon leur formation et leur expérience clinique.

- **Variabilité anatomique :** Les différences morphologiques d'un patient à l'autre complexifient la standardisation de l'examen et rendent parfois le diagnostic plus difficile.
- **Pénurie de spécialistes :** Dans de nombreuses régions, le manque de radiologues formés restreint considérablement l'accès à l'échographie.

Face à ces limites, le développement de méthodes d'aide au diagnostic (CAD) se révèle essentiel pour renforcer la détection et l'analyse automatique de pathologies abdominales, en particulier dans le foie et les reins. Bien que plusieurs approches se soient révélées prometteuses, elles reposent souvent sur des ensembles de données méticuleusement sélectionnés ou des annotations spécialisées coûteuses et parfois approximatives (dû à la subjectivité des experts ou à la qualité variable des images). Par ailleurs, un grand nombre de ces travaux se concentrent uniquement sur des images fixes ou de courtes séquences vidéo dites « trimmed », alors que les données échographiques cliniques sont fréquemment constituées de longues vidéos non recadrées, dont une grande partie peut ne pas contenir d'informations pertinentes pour le diagnostic.

Pour répondre à ces enjeux, cette thèse propose trois contributions majeures :

- **CVL+RankNet (Amélioration de la fiabilité de l'annotation) :** Une méthode d'annotation comparative (Comparative Visual Labeling, CVL), associée à un cadre d'apprentissage par classement (RankNet). Ce protocole fait appel à des comparaisons visuelles entre paires d'images plutôt qu'à l'assignation d'étiquettes absolues, réduisant ainsi la subjectivité et améliorant la cohérence des labels.
- **DR-Clips (Approche guidée pour l'analyse de vidéos non recadrées) :** Un modèle de classification vidéo intégrant un *Frame Relevance Assessor* (FRA), capable de détecter et de sélectionner automatiquement les images ou segments essentiels au diagnostic dans de longues séquences. Cette approche exploite ces sous-clips à la fois comme forme d'augmentation de données et comme guide à l'inférence, tout en épargnant un fastidieux recadrage manuel.
- **KeyFrameDiagFormer (Diagnostic sans module externe) :** Un réseau inspiré de la localisation d'actions en contexte faiblement supervisé, reposant sur un encodeur à mémoire et un mécanisme d'attention hiérarchique. Il permet d'identifier automatiquement les images pertinentes pour le diagnostic sans supervision spécifique sur leur position temporelle, tout en se basant uniquement sur des étiquettes globales (niveau vidéo) pour la classification et la localisation.

Ensemble, ces approches visent à accroître l'efficacité de l'échographie b-mode pour le dépistage précoce de pathologies hépatiques et rénales, tout en limitant la dépendance à l'opérateur, l'exigence en annotations exhaustives et la subjectivité inhérente à l'interprétation humaine. Les chapitres qui suivent détaillent la conception de ces méthodes, leur validation expérimentale et leurs perspectives d'évolution.

Revue de la littérature

Cette section propose un panorama des recherches existantes sur les systèmes d'aide au diagnostic (CAD) en échographie. Nous commençons par présenter les approches qui analysent des images isolées, puis passons aux méthodes tirant parti de séquences vidéo, avant de discuter l'utilisation de modules externes (détection, segmentation, etc.) pour affiner la qualité des données. Enfin, nous mettons en évidence les principaux verrous scientifiques et opportunités de recherche liés à l'imagerie échographique.

Analyse d'images isolées

Les méthodes à image unique consistent à sélectionner (manuellement ou automatiquement) une vue considérée comme pertinente, puis à recourir à un réseau de neurones convolutifs (souvent appelé CNN) pour en extraire un score diagnostique. Elles se distinguent essentiellement par :

- **Zone d'intérêt (ROI) ou image entière :** Certaines approches se focalisent sur une région spécifique du parenchyme (lorsqu'elle est identifiable), ce qui diminue la variabilité et peut accroître la précision. D'autres exploitent l'intégralité de l'image, mais exigent alors davantage de données pour apprendre à ignorer les zones non pertinentes.
- **Multiplicité des vues (multivue) :** Afin de mieux cerner la pathologie, certaines études combinent diverses projections anatomiques (par exemple, plusieurs plans du foie ou du rein).
- **Utilisation de modalités complémentaires :** L'élastographie, le Doppler ou les signaux radiofréquence (RF) peuvent être incorporés dans une architecture multi-branches, afin d'enrichir la représentation de l'organe.

Bien que ces méthodes puissent atteindre des performances très élevées (souvent AUC $>$ 0,95), elles reposent presque toujours sur un tri manuel des images pour ne retenir que les vues de qualité. Cela limite l'automatisation et rend délicate l'extension à d'autres contextes cliniques, où la disponibilité d'images parfaitement cadrées n'est pas garantie.

Extension à l'analyse vidéo

Pour dépasser les limites de l'approche à image unique, de nombreuses recherches s'intéressent à la dimension temporelle et considèrent la vidéo échographique dans son ensemble (ou de longs segments). L'objectif est d'exploiter non seulement le contenu spatial de chaque image, mais aussi la dynamique de la sonde et du patient (respiration, déplacement de l'organe, etc.). Parmi les stratégies les plus courantes :

- **Réseaux convolutifs avec modélisation temporelle (RNN, LSTM, attention) :** On commence par extraire des descripteurs dans chaque image (à l'aide d'un réseau convolutif), puis on capture la dimension séquentielle grâce à des blocs récurrents (LSTM) ou un mécanisme d'attention (souvent désigné par "transformer").
- **Réseaux 3D (3D-CNN, ViViT, Swin Transformer vidéo) :** Les images sont combinés pour former un volume spatio-temporel. Les filtres convolutifs ou les mécanismes d'auto-attention s'appliquent directement à cette structure 3D, au prix d'une charge de calcul plus élevée.
- **Modèles à deux flux :** Un flux dédie l'analyse à la structure spatiale (b-mode), tandis que l'autre se consacre à la composante mouvement (p. ex. calcul du flux optique). Les sorties sont ensuite fusionnées pour produire un diagnostic tenant compte à la fois de la texture et de la dynamique.

L'adoption de labels au niveau vidéo (ex. "présence ou absence de pathologie") simplifie la création de bases d'apprentissage, mais la plupart des séquences cliniques demeurent longues, hétérogènes, et renferment de nombreux images n'apportant aucune information diagnostique.

Approches guidées par modules externes

Plusieurs auteurs proposent des modules complémentaires pour filtrer ou annoter automatiquement la séquence avant le diagnostic :

- **Réseaux de segmentation ou de détection :** On isole d'abord la structure cible (foie, rein, lésion), ce qui permet au module diagnostique de se concentrer sur la zone identifiée. Les masques ou bounding boxes exigent cependant des annotations détaillées.
- **Évaluation de la qualité :** Un réseau peut estimer la netteté ou la lisibilité de chaque image, ne conservant que ceux qui respectent un seuil qualitatif. Cette étape améliore la pertinence des données mais suppose un modèle entraîné à cet effet (et donc un effort de labellisation supplémentaire).

Ces approches guidées offrent des gains de précision non négligeables mais accroissent la complexité et la dépendance à des annotations plus riches.

Fiabilité et défis de la vidéo échographique non recadrée

L'échographie est sujette à une forte variabilité, tant en termes de qualité d'image que d'expertise humaine :

- **Subjectivité et hétérogénéité des annotations :** L'interprétation dépend du spécialiste, et l'on constate des désaccords fréquents sur l'existence ou la sévérité d'une pathologie. Les techniques d'apprentissage doivent donc composer avec des labels parfois incohérents ou imprécis.
- **Vidéos non recadrées :** Dans la pratique, on collecte de longues séquences, dont une fraction significative est inexploitée (bruit, artéfacts, parties hors-champ). Les modèles actuels sont souvent évalués sur des extraits déjà nettoyés ou de courte durée, ce qui limite leur généralisation.
- **Accès restreint aux références cliniques solides :** Les étiquettes basées sur la biopsie ou l'IRM constituent l'étalon-or diagnostique, mais sont coûteuses et invasives, freinant l'acquisition de grandes bases annotées.

Bien que certaines recherches proposent des stratégies pour détecter les labels douteux ou intégrer la variabilité entre annotateurs, la question de la fiabilité demeure entière, en particulier pour l'échographie où l'opérateur conditionne grandement la qualité et le contenu des données acquises.

Verrous actuels et pistes d'avenir

Plusieurs points demeurent problématiques :

- **Gestion des annotations incertaines :** Réduire la subjectivité et définir des protocoles robustes pour la labellisation échographique sont des priorités pour fiabiliser les modèles.
- **Automatisation et temps réel :** Les systèmes destinés à un large dépistage doivent être capables de traiter des vidéos entières, de façon rapide et sans intervention humaine pour éliminer les images non pertinents.
- **Généralisation et robustesse :** L'implémentation clinique nécessite des algorithmes assez souples pour supporter une large variété de machines, de réglages et de situations anatomiques.

— **Identification automatique de images utiles :** La sélection ou la pondération automatique des segments clés, dans une séquence pouvant compter plusieurs milliers de images, reste largement ouverte.

En définitive, les solutions d'apprentissage profond ont fait leurs preuves sur des échantillons soigneusement triés, mais il reste à concevoir des systèmes capables de traiter l'échographie dans toute son hétérogénéité, afin d'offrir un réel appui au dépistage précoce et à la standardisation du diagnostic.

CVL+RankNet : une nouvelle approche d'annotation des images pour l'aide au diagnostic

Résumé

Le succès des systèmes d'aide au diagnostic (CAD) basés sur l'intelligence artificielle dépend en grande partie de la disponibilité de données annotées à grande échelle, et de manière fiable. En échographie abdominale, l'obtention de tels jeux de données reste particulièrement difficile : d'une part, les méthodes de référence (biopsies hépatiques, IRM) sont coûteuses et peu accessibles, et d'autre part, l'annotation visuelle par des experts s'avère subjective et souvent imprécise, notamment lors de l'évaluation de la sévérité d'une pathologie comme la stéatose hépatique.

Afin de pallier ces limites, nous proposons **Comparative Visual Labeling (CVL)**, une méthode d'annotation par comparaisons relatives. Plutôt que d'attribuer un label binaire (« pathologique » ou « sain ») à chaque image, on compare deux images à la fois : l'annotateur juge laquelle présente le degré pathologique le plus prononcé. Nous convertissons ensuite ces comparaisons en scores numériques continus à l'aide de **RankNet**, un algorithme d'apprentissage par classement (*Learning-to-Rank*). De fait, chaque image reçoit un *score de sévérité* cohérent avec l'ensemble des comparaisons.

Contexte de la stéatose hépatique en échographie

La stéatose hépatique se caractérise par une accumulation de graisse dans le parenchyme du foie, fréquemment liée à des facteurs métaboliques ou à la consommation d'alcool. À un stade avancé, elle peut évoluer en cirrhose ou en carcinome hépatocellulaire. L'échographie b-mode, bien qu'elle soit un moyen de dépistage non invasif et abordable, souffre d'une forte variabilité liée à l'opérateur et à la qualité des images, rendant le diagnostic parfois incertain.

Dans ce cadre, **CVL+RankNet** propose une solution pour générer un grand nombre de labels exploitables (scores de sévérité) à partir d'annotations *relatives*. L'objectif est de surmonter la subjectivité de l'annotation directe (Single-Image Visual Labeling, SVL), où chaque image est jugée isolément.

Méthodologie de CVL+RankNet

Comparative Visual Labeling (CVL).
Dans CVL, l'annotateur reçoit des paires d'images échographiques

$$\left\{(I_i, I_j) \mid i \neq j\right\}$$

et doit décider laquelle des deux images apparaît la plus pathologique. Plus formellement, on définit quatre types de labels :

$$D_{i,j} \in \{+1, -1, 0^+, 0^-\},$$

avec :

- $D_{i,j} = +1$: l'image I_i semble plus pathologique que l'image I_j,
- $D_{i,j} = -1$: l'image I_j semble plus pathologique que l'image I_i,
- $D_{i,j} = 0^+$: les deux images sont perçues comme pathologiques mais indiscernables en termes de sévérité,
- $D_{i,j} = 0^-$: les deux images sont jugées saines et indiscernables.

Cette approche s'avère souvent plus robuste que l'annotation binaire classique, car l'être humain discerne plus aisément une différence relative de sévérité qu'un seuil absolu. La Figure B.1 illustre ce principe.

Transformation en scores continus (RankNet).
La deuxième étape convertit ces labels de paires en un *score de sévérité* réel

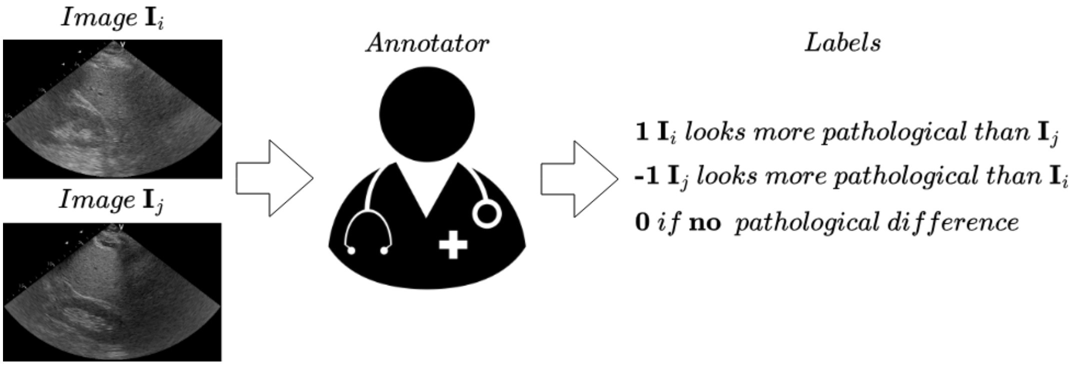


Image $\mathbf{I}_i$
Image $\mathbf{I}_j$
Annotator
Labels
1 $\mathbf{I}_i$ looks more pathological than $\mathbf{I}_j$
-1 $\mathbf{I}_j$ looks more pathological than $\mathbf{I}_i$
0 if no pathological difference

FIGURE B.1 – Processus d'annotation en Comparative Visual Labeling (CVL). Chaque paire (I_i, I_j) reçoit un label dans $\{1, -1, 0^+, 0^-\}$ d'après l'impression visuelle de sévérité pathologique.

pour chaque image. On formalise cela comme un problème d'apprentissage par classement (*Learning-to-Rank*). Soit $\Phi(\mathbf{x}_i; \boldsymbol{\theta}, x_q)$ la fonction (paramétrée par $\boldsymbol{\theta}$) qui assigne un score $s_i \in \mathbb{R}$ à chaque image I_i. Ici, $\mathbf{x}_i$ désigne le vecteur de caractéristiques associé à l'image I_i et x_q peut représenter la pathologie d'intérêt (ici, la stéatose).

RankNet [24] apprend Φ à partir des comparaisons sous forme probabiliste : la probabilité que I_i soit plus pathologique que I_j est

$$P(i \succ j) = \sigma\Big(\Phi(\mathbf{x}_i; \boldsymbol{\theta}, x_q) \;-\; \Phi(\mathbf{x}_j; \boldsymbol{\theta}, x_q)\Big), \tag{B.1}$$

où σ est la fonction sigmoïde. Pour chaque paire annotée (i, j), on associe un label binaire y_{ij} (par exemple, $y_{ij} = 1$ si I_i est jugé plus pathologique que I_j, et 0 dans le cas contraire). Le réseau cherche alors à minimiser la perte :

$$\mathcal{L}'_{PW}(\boldsymbol{\theta}) = -\sum_{(i,j)\in C} y_{ij} \log\Big(P(i \succ j)\Big), \tag{B.2}$$

où C est l'ensemble des paires comparées. Au terme de l'entraînement, on obtient un *score continu* s_i par image. La Figure B.2 donne un aperçu du fonctionnement global de RankNet.

Seuils et calibration.

Une fois les scores obtenus, on peut définir un seuil τ afin de distinguer par exemple « sain » ($s_i < \tau$) versus « pathologique » ($s_i \geq \tau$). L'avantage est de pouvoir ajuster τ pour privilégier la sensibilité ou la spécificité, sans recommencer l'annotation.

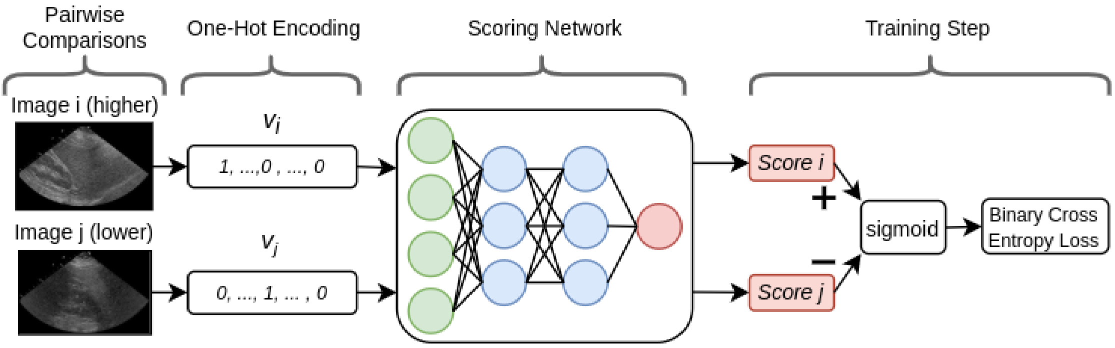


Figure B.2 – Schéma d'implémentation de RankNet, un réseau de neurones (apprentissage par classement) entraîné sur des comparaisons de paires (+1 ou -1). Chaque image reçoit à la fin un score réel s_i, reflétant son degré de pathologie.

Principaux résultats et observations

Réduction du taux d'erreur.
Nous avons évalué CVL+RankNet sur un jeu de données échographiques contenant 55 patients, pour lesquels des analyses histopathologiques (pourcentage de cellules graisseuses, PFH) servent de référence. Comme le montre le Tableau B.1, les taux d'erreur (F1 et AUC) s'améliorent sensiblement par rapport à l'annotation image-par-image (SVL), surtout pour les formes modérées ("mild steatosis") où le risque de sous-estimation est élevé.

Table B.1 – Qualité des labels (F1 et ROC-AUC) sur le Dataset 1, comparant SVL et CVL+RankNet. Les intervalles de confiance (2,5 % et 97,5 %) sont indiqués entre crochets.

Annotateur / Méthode	A	B	C	Fusion (*majority vote*)
SVL (F1)	0.92 [0.85, 0.98]	0.83 [0.72, 0.92]	0.85 [0.75, 0.93]	0.87 [0.77, 0.94]
CVL+RankNet (F1)	0.99 [0.96, 1.00]	0.93 [0.86, 0.98]	0.93 [0.86, 0.98]	0.97 [0.93, 1.00]
CVL+RankNet (AUC)	0.99 [0.90, 1.00]	0.97 [0.88, 0.99]	0.95 [0.88, 0.99]	0.99 [0.89, 1.00]

Meilleure cohérence inter-annotateurs.
L'indice de Fleiss' Kappa augmente quand on passe de SVL (0.75, « accord substantiel ») à CVL+RankNet (0.84, « accord quasi parfait »). Autrement dit, les observateurs sont plus en phase lorsqu'ils comparent deux images plutôt que quand ils jugent chaque image isolément.

Corrélation avec la sévérité réelle.
La Figure B.3 illustre la concordance entre les scores CVL+RankNet (axe y) et le PFH issu de l'histopathologie (axe x). La corrélation de Spearman atteint 0,87, preuve que CVL+RankNet reflète efficacement la progression de la stéatose, notamment dans les stades légers et modérés.

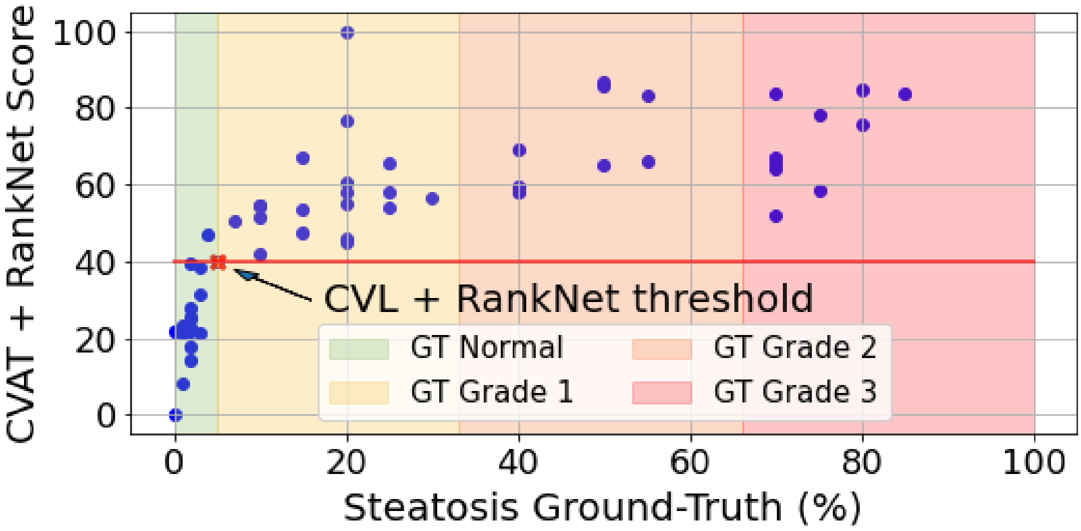


FIGURE B.3 – Relation entre les scores CVL+RankNet obtenus par fusion d'annotations (y) et la valeur de PFH (x) mesurée par histologie. Les zones colorées indiquent la sévérité : sain (vert), Grade 1 (jaune), Grade 2 (orange) et Grade 3 (rose). La ligne rouge marque le seuil choisi pour séparer « sain » et « pathologique ».

Annotation flexible et extensible.
Même avec un nombre de comparaisons modéré (5 à 6 paires par image), CVL+RankNet garde un net avantage sur SVL. De plus, ce principe peut s'appliquer à d'autres maladies (myopie sévère, colite ulcéreuse) et potentiellement à d'autres modalités d'imagerie (IRM, endoscopie, etc.), comme le montrent des travaux récents dans la littérature.

Perspectives et travaux futurs

— **Réduction de l'effort d'annotation** : Des stratégies d'apprentissage actif (*active learning*) permettraient de sélectionner de manière plus optimale les paires à annoter, diminuant la charge pour l'expert.

— **Extension à d'autres maladies et modalités** : CVL a déjà montré son intérêt pour la myopie sévère et la colite ulcéreuse. Les mêmes principes (comparaison, classement) peuvent s'étendre à d'autres applications médicales.

— **Intégration dans des CAD vidéo** : À ce jour, CVL+RankNet concerne surtout des images fixes. Son adaptation aux séquences échographiques longues et non recadrées (avec repérage automatique de images pertinents) constituerait un prolongement naturel.

Conclusion

En associant **Comparative Visual Labeling** (CVL) et un apprentissage par classement (**RankNet**), il est possible d'augmenter sensiblement la fiabilité des annotations visuelles en échographie. L'approche est particulièrement adaptée à la détection précoce de la stéatose hépatique, où l'expertise humaine tend à sous-estimer les cas modérés. Elle représente un compromis prometteur entre la facilité d'utilisation (comparaison par paires) et la production de scores continus permettant un diagnostic plus fin. Cette démarche ouvre la voie à des systèmes d'aide au diagnostic plus robustes et mieux adaptés à la pratique clinique, avec une meilleure sensibilité dans les premiers stades et un potentiel d'extension vers d'autres contextes (multimodalité, séquences vidéo, etc.).

DR-Clips : une approche guidée pour la classification vidéo en échographie non recadrée

Contexte et motivation

Dans le chapitre précédent, nous nous sommes concentrés sur l'amélioration de la qualité des annotations visuelles en échographie. Nous abordons ici un second défi majeur : entraîner des modèles de classification vidéo à partir de séquences échographiques non recadrées (*untrimmed videos*), pour lesquelles seuls des labels globaux (diagnostics au niveau vidéo) sont disponibles. Contrairement aux images statiques ou aux clips strictement recadrés sur des moments clés, les vidéos non recadrées comportent de nombreux segments peu ou pas utiles pour le diagnostic. Cela complique l'apprentissage et accroît le risque de surapprentissage (*overfitting*), notamment lorsque la base de données est de taille modeste.

Méthodologie DR-Clips

Synthèse de la méthode Pour relever ce défi, nous proposons **DR-Clips**, une méthodologie visant à améliorer la performance et la robustesse de classifieurs vidéo en échographie. L'originalité tient dans l'utilisation d'un **Frame Relevance Assessor (FRA)**, réseau de neurones de régression qui attribue un score de pertinence à chaque image. Les images estimées pertinentes (selon la qualité visuelle, l'angle de la sonde, l'absence d'artéfacts, la présence éventuelle d'indices pathologiques, etc.) sont alors ordonnées pour constituer des *« clips diagnostiquement pertinents »* (**DR-Clips**).

Un réseau de neurones vidéo est ensuite entraîné sur ces DR-Clips (*avec labels au niveau vidéo*), apprenant à ignorer les images non pertinentes. Pour éviter de dépendre entièrement d'un FRA « parfait », nous introduisons un échantillonnage aléatoire au sein des clips, ce qui renforce la robustesse du classifieur même en présence de images peu informatives.

Dataset Pour démontrer l'efficacité de DR-Clips, nous considérons deux tâches binaires en échographie abdominale :

- **Tâche Foie** : détection de dommages hépatiques (stéatose ou fibrose) vs. foie sain,
- **Tâche Rein** : détection d'anomalies structurelles (kystes, hydronéphrose, etc.) vs. rein sain.

Nous avons construit un nouveau jeu de données, *MIM-US-107 Video Dataset*, constitué de 107 vidéos échographiques non recadrées (une par patient), labellisées globalement (sain vs. pathologique) par un radiologue. Chaque vidéo peut contenir de nombreux segments inutiles et seulement quelques images pertinentes. En complément, nous disposons du *MIM-US-473 Still Image Dataset* (7 924 images fixes, annotées par un score de pertinence).

Frame Relevance Assessor (FRA) Le FRA est entraîné via une fonction de coût de type Smooth L1 Loss :

$$\text{Loss}_{FRA} = \frac{1}{N}\sum_{i=1}^{N} \textit{SmoothL1}\Big(\Phi_{\text{FRA}}(I_i),\, r_i\Big),$$

où I_i est l'image d'entraînement, $\Phi_{\text{FRA}}(I_i)$ la valeur prédite par le FRA et $r_i \in [0,1]$ la note de pertinence normalisée (du moins pertinent au plus pertinent).

Principe d'apprentissage de DR-Clips Une fois le FRA entraîné, on segmente chaque vidéo en images associées à un score $\hat{r}_k$. Puis on génère plusieurs *DR-Clips* :

1. **Extraction de clips aléatoires** : Pour chaque vidéo, on échantillonne *L=500* sous-séquences (1 à 32 images), chacune étant triée par ordre décroissant de pertinence.
2. **Pondération par la pertinence** : Chaque DR-Clip hérite d'un poids w_k égal à la moyenne de ses scores (passés dans une sigmoïde décalée).

3. **Pruning** : On ne conserve que les *top-K* DR-Clips les plus pertinents (ici $K = 10$).
4. **Entraînement du classifieur vidéo** : on applique une loss de type DR-Clip (inspirée du *Curriculum Learning*) qui pénalise davantage les clips les mieux notés :

$$\mathcal{L}_{\text{DR-Clip}}(\Theta) \;=\; \sum_{v=1}^{V}\sum_{k=1}^{K} w_k^v\,\mathcal{L}\Big(\Theta(C_k^v),\, y_v\Big),$$

où y_v est le label de la vidéo (*foie sain* vs. *foie pathologique*, etc.).

À l'inférence, on retient uniquement les N images les plus pertinentes ($N \leq 32$) pour constituer un DR-Clip unique, que le réseau vidéo transforme en prédiction diagnostique.

Résultats expérimentaux

Nous comparons **DR-Clips** à plusieurs approches :

- *Classifieurs vidéo non guidés* (ex. Video Swin [68], MViTv2, X3D...), entraînés sans filtrage particulier et exploitant une fusion (moyenne ou max) sur des clips de taille fixe (32 images).
- *Classifieurs single-image guidés manuellement* (*ResNet* (*Manual*)), utilisant des images-clés choisies par l'expert.
- *Classifieurs single-image guidés automatiquement* (*ResNet* (*Auto*), etc.), utilisant des images-clés proposées par le FRA.

TABLE B.2 – Comparaison DR-Clips vs. baselines (ROC-AUC moyen sur 10 folds, écart-type entre parenthèses).

		Foie (AUC)		**Rein (AUC)**	
Cat. de Modèle	**Nom**	*moy. fusion*	*max fusion*	*moy. fusion*	*max fusion*
Single-image	ResNet (Manual)	0.96 (0.11)	0.95 (0.15)	0.91 (0.25)	0.98 (0.05)
	ResNet (Auto)	0.94 (0.17)	0.94 (0.13)	0.71 (0.26)	0.53 (0.24)
	Swin (Auto)	0.94 (0.17)	0.89 (0.16)	0.81 (0.21)	0.69 (0.22)
Vidéo	BabyNet	0.89 (0.25)	0.87 (0.30)	0.61 (0.21)	0.59 (0.31)
	EchoGNN	0.82 (0.29)	0.77 (0.30)	0.55 (0.19)	0.60 (0.24)
	UVT	0.80 (0.29)	0.74 (0.24)	0.68 (0.17)	0.67 (0.16)
	Video Swin	0.89 (0.30)	0.91 (0.17)	0.70 (0.29)	0.72 (0.22)
	VideoMAE V2	0.80 (0.37)	0.87 (0.26)	0.56 (0.25)	0.60 (0.20)
	MViTv2	0.88 (0.30)	0.90 (0.26)	0.74 (0.19)	0.64 (0.26)
	X3D	0.85 (0.24)	0.94 (0.13)	0.73 (0.23)	0.69 (0.26)
DR-Clips	Video Swin + DR-Clips	0.97 (0.09)†		0.92 (0.13)†	

† Le modèle produit une seule valeur par vidéo, indépendamment de la fusion mean ou max.

Comme le montre le tableau B.2 :

- **Foie (NAFLD/fibrose, pathologie diffuse)** : Les classifieurs vidéo non guidés (Video Swin, MViTv2, etc.) atteignent 0.94. ResNet (Manual) avoisine 0.95–0.96. DR-Clips dépasse légèrement (0.97).
- **Rein (kyste/hydronéphrose, pathologie localisée)** : Les méthodes vidéo non guidées chutent (0.70–0.74). Les approches single-image guidées manuellement plafonnent jusqu'à 0.98, et DR-Clips réduit fortement l'écart (0.92).
- **Interprétation** : Pour le foie, la pathologie s'observe dans presque toutes les images, facilitant l'apprentissage. Au contraire, pour le rein, la lésion n'apparaît que dans un nombre restreint de images. Les modèles non guidés noyent alors la classe pathologique parmi trop de images saines.
- **DR-Clips vs. single-image auto (Swin/ResNet Auto)** : Sur la Tâche Rein, l'étiquetage image-par-image est plus bruyant : des images saines reçoivent un label pathologique. Les DR-Clips vidéo s'avèrent plus robustes, car il est peu probable qu'un lot entier de images soit faux.

Analyses complémentaires

Qualité du FRA : Nous avons mesuré la corrélation entre le score $\hat{r}$ du FRA et la proximité visuelle d'une image avec les images-clés manuelles (dans un espace PCA). On obtient une forte corrélation négative (-0.84 sur la Tâche Foie), indiquant que le FRA valorise davantage les images proches des images-clés expertes.

Longueur du DR-Clip : À l'inférence, conserver un grand nombre de images (au-delà de 4) peut dégrader la performance, surtout pour le rein, en raison des images non pertinentes qui diluent l'information utile.

Conclusion et perspectives

Notre méthode **DR-Clips** concilie deux points :

- un *Frame Relevance Assessor*, formé sur des scores de visibilité/artéfacts,
- un classifieur vidéo (ex. Video Swin) traitant des clips de images jugées pertinentes, avec échantillonnage aléatoire et pondération par la pertinence.

Les résultats montrent qu'en échographie abdominale non recadrée, DR-Clips surpasse les classifieurs vidéo classiques pour la Tâche Rein et fait jeu égal (ou légèrement mieux) sur la Tâche Foie. Plusieurs améliorations sont envisagées :

- **Généralisation** : Vérifier la robustesse du FRA et de DR-Clips à d'autres pathologies et d'autres machines d'échographie.

— **Autres architectures** : Évaluer DR-Clips avec différents classifieurs vidéo (3D-CNN, Transformers, etc.).

— **Robustesse aux clips longs** : Ajuster la gestion des images peu pertinentes lors d'inférences longues.

— **Réduction de la supervision** : Se passer d'annotations explicites de pertinence (approche auto-supervisée ou faiblement supervisée).

En somme, DR-Clips apporte une réponse au problème du « bruit » dans les vidéos non recadrées, atteignant un niveau de performance proche de l'extraction manuelle d'images-clés. Son intégration à large échelle promet des diagnostics plus automatisés et plus fiables en pratique clinique.

KeyFrameDiagFormer : Transformer pour la Localisation de images Clés et le Diagnostic Faiblement Supervisé dans les Vidéos Échographiques Non-Découpées

Dans ce dernier chapitre, nous proposons **KeyFrameDiagFormer**, un modèle de classification vidéo faiblement supervisé, dédié à l'analyse d'échographies abdominales non découpées (*untrimmed videos*). Contrairement au chapitre précédent, où l'identification de images pertinentes s'appuyait sur un module externe pré-entraîné (FRA), **KeyFrameDiagFormer** apprend *intégralement* à partir de simples labels uniquement au niveau de la vidéo, sans besoin d'annotations fines (par images ou via un module externe).

L'idée directrice est d'adapter certaines techniques de *Weakly-supervised Temporal Action Localization* (*WTAL*) issues de la vision par ordinateur classique, afin de repérer automatiquement, dans la séquence échographique, les images présentant un réel intérêt diagnostique (détection de pathologie ou confirmation du statut sain). À partir de ces images-clés (*keyframes*), le système délivre un diagnostic multi-label au niveau de la vidéo complète, tout en fournissant un *score de faisabilité* (dans le cas où l'organe est mal visible).

Contexte et motivation

Les échos vidéos non recadrées (p. ex. un enregistrement complet d'examen abdominal) contiennent inévitablement de nombreuses images non pertinentes : organes hors champ, artéfacts, passages de recherche du point d'intérêt, etc. Les approches précédentes (DR-Clip) ont montré que l'ajout d'un guidage (FRA) améliorait nettement la performance, mais nécessitait un entraînement supervisé de ce module. Dès lors, pour généraliser à d'autres pathologies (par ex. pancréas, rate, prostate. . .), il faudrait *personnaliser* ce guidage.

Ici, on vise à s'affranchir de ce guidage externe en recourant à un schéma **faiblement supervisé** : seuls des labels globaux (pathologies présentes ou non dans la vidéo) sont utilisés, ce qui ouvre la voie à un apprentissage à partir de données plus nombreuses (y compris labels issus de dossiers médicaux).

Méthodologie KeyFrameDiagFormer

Architecture globale

Le réseau se décompose en quatre blocs :

1. **Frame Encoder** : un réseau 2D (Swin Transformer v2) encode chaque image indépendamment, générant une représentation compacte (*feature*).
2. **Frame-Memory Module** : une "banque de mémoire" stocke les features des images précédentes. À chaque itération d'entraînement, seuls les features de quelques images sont recalculés, ce qui allège les contraintes mémoire et permet d'entraîner le image encoder efficacement sur de longues séquences.
3. **Video Self-Attention Module** : des blocs Transformers intègrent un *local self-attention* (pour la cohérence temporelle de voisinage) et un *global organ-specific attention* (pour agréger les images d'un organe tout en excluant les autres ou le bruit).
4. **Hierarchical Multi-Label Classification** : enfin, un module de classification multi-label hiérarchique produit (1) un score de "fond" (background class), (2) un score d'organe sain (*healthy*), et (3) les scores de pathologies (lésions, stéatose, etc.). Les images les plus discriminantes (*top-R*) deviennent de facto les *keyframes* identifiées pour justifier le diagnostic.

Entraînement faiblement supervisé

Le modèle exploite uniquement des étiquettes au niveau de la vidéo : un vecteur multi-label (ex. : {*foie sain*=0, stéatose=1, lésion hépatique=2, . . . }) et applique une *Binary Cross Entropy* par classe, en ignorant les classes "saines" lorsqu'une pathologie est confirmée pour l'organe correspondant. Par ailleurs, la classe "non pertinente" est supposée toujours présente, pour refléter la majorité des images inexploitables.

Diagnostic en temps réel

Lors de l'inférence, on encode chaque nouvelle image pour mettre à jour la mémoire, puis on applique le Transformer avec un nombre maximal de images (T_{w}). Les images estimées comme peu pertinentes sont évincées si on dépasse T_{w}. Ce procédé autorise une utilisation quasi temps-réel en contexte clinique.

Principaux résultats et observations

Performances de classification vidéo

Sur un jeu de données élargi (> 107 vidéos), **KeyFrameDiagFormer** obtient d'excellents scores pour la stéatose hépatique (ROC-AUC à 0,97), et des résultats satisfaisants sur les lésions du foie (ROC-AUC $\approx$ 0,89). Les tâches rénales se révèlent plus complexes (AUC 0,66–0,79). La Table B.3 illustre les performances globales de classification vidéo. Les illustrations montrent que le réseau localise correctement, sans supervision fine, les images présentant des indices pathologiques ou confirmant le caractère sain, tout en écartant la majorité des segments non pertinents.

TABLE B.3 – Performance metrics for video diagnosis across different classes, including ROC-AUC, Accuracy, Precision, Recall, and F1-Score. The results highlight the model's strong performance for liver steatosis and reasonable outcomes for other liver conditions, with lower performance for kidney-related classes.

Class	**ROC-AUC**	**Acc.**	**Precision**	**Recall**	**F1**
Liver Healthy (h_{L})	0.82	0.83	0.84	0.81	0.82
Liver steatosis (p_{L}^{0})	0.97	0.94	0.92	0.85	0.88
Liver Lesion (p_{L}^{1})	0.89	0.90	0.71	0.62	0.67
Kidney Healthy (h_{K})	0.66	0.69	0.80	0.57	0.67
Kidney Abnormal (p_{K}^{0})	0.79	0.85	0.57	0.44	0.50

Évaluation de la faisabilité diagnostique

Grâce à sa structure multi-label hiérarchique, le réseau est capable d'indiquer un score de *faisabilité* pour un organe donné (ex. : *aucune image de bonne* *qualité pour le rein*). Cette fonctionnalité prévient l'utilisateur qu'un examen complémentaire ou des images supplémentaires sont requis pour poser un diagnostic certain.

Limites pour le rein
Les pathologies rénales sont variées (kystes, lithiases, hydronephrose...) et parfois difficiles à distinguer d'artéfacts (zones d'ombre). On note une performance moindre sur ces classes, accentuée par le fait que plusieurs lésions rénales sont regroupées en une seule étiquette *kidney abnormal*, augmentant l'hétérogénéité de la classe.

Conclusion et perspectives

KeyFrameDiagFormer illustre la possibilité de **former un classifieur vidéo échographique sans guidage externe**, tout en localisant de façon autonome les images déterminantes (*keyframes*). La démarche s'appuie sur :

- Un *masquage du bruit* (background) inspiré de l'action localization,
- Une *Structure multiclass hiérarchique* pour distinguer les images de différents organes et le *background*,
- Une *banque mémoire* de images pour un entraînement scalable.

Les résultats sont particulièrement solides pour la stéatose hépatique et encouragent l'intégration de cette méthode sur d'autres organes/pathologies. Les travaux futurs portent sur :

- **Affinage et tuning de l'architecture** : nombre de blocs Transformers, taille de batch, hyperparamètres comme τ, top-R, etc.
- **Exploration de nouvelles fonctions de pertes (WTAL)** : des fonctions de coût spécifiques à la localisation d'actions pourraient améliorer la séparation images utiles/*background*.
- **Apprentissage multimodal** : la structure indépendante des images permet d'intégrer d'autres données (texte, IRM, CT), y compris pour un pré-entraînement à large échelle.

En conclusion, **KeyFrameDiagFormer** propose un cadre faiblement supervisé prometteur, permettant d'analyser des vidéos échographiques non recadrées, d'identifier automatiquement des images-clés et de gérer la diversité anatomique via une classification hiérarchique.

Conclusion

Dans ce travail, nous nous sommes concentrés sur le développement de **modèles de deep learning dédiés au diagnostic assisté par ordinateur (CAD) pour l'échographie abdominale**, ciblant spécifiquement les pathologies hépatiques et rénales dans des vidéos non découpées (*untrimmed*). L'objectif était d'élaborer des solutions plus complètes, capables de traiter la variabilité et la complexité de l'imagerie échographique, tout en réduisant la dépendance à des annotations coûteuses ou à l'expertise d'opérateurs confirmés. Nous présentons ci-dessous un résumé de nos principales contributions.

Réduire l'erreur d'annotation avec CVL+RankNet

Dans un premier temps, nous avons introduit **CVL+RankNet** pour améliorer la *fiabilité des annotations* en échographie. Plutôt que de s'appuyer sur des labels visuels classiques, plus sensibles à la subjectivité, nous avons adopté une méthode de *Comparative Visual Labeling* (CVL). Celle-ci invite les annotateurs à comparer deux images pour déterminer celle présentant un degré plus marqué de pathologie, ce qui réduit la variabilité inter-annotateurs. La formulation en *Learn-to-Rank* génère des scores pathologiques continus, validés empiriquement, et montre une amélioration notable de la fiabilité de l'annotation. Cette approche contribue à la création de jeux de données de haute qualité, condition essentielle pour des systèmes de diagnostic assisté efficaces.

Diagnostiquer des vidéos échographiques non découpées avec DR-Clips

Ensuite, nous avons proposé **DR-Clips**, une méthode de classification vidéo qui repose sur un module de guidage externe évaluant la pertinence diagnostique de chaque image. Les images les plus utiles sont regroupées pour former des séquences courtes, servant à la fois pour l'apprentissage et l'inférence. Cette approche a démontré de bonnes performances, comparables à celles de modèles entraînés sur des images sélectionnées par des experts, tout en offrant la possibilité d'exploiter des vidéos non découpées. Elle met en évidence l'importance d'un filtrage des images non pertinentes pour traiter efficacement des données échographiques en conditions réelles.

Une approche faiblement supervisée avec KeyFrameDiagFormer

Enfin, nous avons introduit **KeyFrameDiagFormer**, un modèle capable de diagnostiquer des vidéos échographiques non découpées *sans guidance externe*. Il s'appuie uniquement sur des labels de haut niveau (présence ou absence de pathologie), en s'inspirant des méthodes de *weakly-supervised action localization*.

Le modèle apprend à extraire des *keyframes* pertinentes pour la pathologie, grâce à une classification multi-label hiérarchique. Les résultats sont particulièrement prometteurs, notamment pour la détection de la stéatose hépatique. Le modèle peut également indiquer un *score de faisabilité* lorsqu'un organe est mal représenté, fournissant un retour en temps réel sur la qualité diagnostique de la séquence. Cette approche faiblement supervisée présente un fort potentiel d'extensibilité et un coût d'annotation réduit.

Intégration et implications pratiques

Les travaux s'inscrivent dans la logique du *Disrumpere Project*, développé par l'IRCAD, qui vise à démocratiser l'échographie via des sondes portables et des algorithmes d'IA. Une version simplifiée de nos modèles a été intégrée dans un logiciel clinique pour le dépistage automatisé de la stéatose hépatique. Les premiers retours indiquent que des opérateurs peu expérimentés peuvent effectuer un dépistage fiable, optimisant l'emploi des spécialistes et rendant l'échographie plus accessible, en particulier dans les régions sous-équipées.

Limites et perspectives

Plusieurs pistes d'amélioration restent à explorer. Premièrement, la *généralisation* doit être étendue à un éventail plus large de pathologies et de dispositifs échographiques. Deuxièmement, la méthode CVL pourrait être appliquée à des annotations au niveau vidéo ou servir directement à l'entraînement de réseaux siamois. Troisièmement, KeyFrameDiagFormer bénéficierait de techniques de *self-supervised pre-training* pour améliorer encore ses performances. De même, il serait judicieux d'envisager des fonctions de coût plus complexes issues du *weakly-supervised temporal action localization*. Enfin, l'intégration de données multimodales (texte, IRM, dossiers cliniques) offre une opportunité de renforcer la robustesse et la précision du diagnostic.

Conclusion générale

Les méthodes présentées dans ce travail contribuent à **améliorer la fiabilité des annotations** (grâce à CVL+RankNet) et à **développer des réseaux capables de gérer des données échographiques non découpées**, que ce soit en recourant à un module de guidage (DR-Clips) ou via un apprentissage faiblement supervisé (KeyFrameDiagFormer). Ces avancées ouvrent la voie à des systèmes de diagnostic échographique plus *scalables*, limitant la dépendance à l'expertise et aux annotations minutieuses. À terme, elles participent à une démocratisation de l'échographie, susceptible de réduire les inégalités d'accès aux soins, et de répondre aux besoins toujours croissants en imagerie médicale.

Bibliographie

[1] U Rajendra Acharya, Hamido Fujita, Shreya Bhat, U Raghavendra, Anjan Gudigar, Filippo Molinari, Anushya Vijayananthan et Kwan Hoong Ng. « Decision support system for fatty liver disease using GIST descriptors extracted from ultrasound images ». In : *Information Fusion* 29 (2016), pages 32-39.

[2] U Rajendra Acharya, S Vinitha Sree, Ricardo Ribeiro, Ganapathy Krishnamurthi, Rui Tato Marinho, João Sanches et Jasjit S Suri. « Data mining framework for fatty liver disease classification in ultrasound : a hybrid feature extraction paradigm ». In : *Medical physics* 39.7Part1 (2012), pages 4255-4264.

[3] Rehan Ahmad et Basant K. Mohanty. « Chronic kidney disease stage identification using texture analysis of ultrasound images ». In : *Biomedical Signal Processing and Control* 69 (2021), page 102695. issn : 1746-8094. doi : `https://doi.org/10.1016/j.bspc.2021.102695`. url : `https://www.sciencedirect.com/science/article/pii/S1746809421002925`.

[4] N Ahmadi, Michael Y Tsang, Ang Nan Gu, Teresa SM Tsang et Purang Abolmaesumi. « Transformer-based spatio-temporal analysis for classification of aortic stenosis severity from echocardiography cine series ». In : *IEEE Transactions on Medical Imaging* (2023).

[5] Taymaz Akan, Sait Alp, Md Shenuarin Bhuiyan, Tarek Helmy, A Wayne Orr, Md Mostafizur Bhuiyan, Steve Conrad, John Vanchiere, Christopher G Kevil et Mohammad Alfrad Nobel Bhuiyan. « ViViEchoformer : Deep Video Regressor Predicting Ejection Fraction ». In : *medRxiv* (2024), pages 2024-06.

[6] Alaleh Alivar, Habibollah Danyali et Mohammad Sadegh Helfroush. « Hierarchical classification of normal, fatty and heterogeneous liver diseases from ultrasound images using serial and parallel feature fusion ». In : *Biocybernetics and Biomedical Engineering* 36.4 (2016), pages 697-707.

[7] Rami Aly, Steffen Remus et Chris Biemann. « Hierarchical multi-label classification of text with capsule networks ». In : *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics : Student Research Workshop*. 2019, pages 323-330.

[8] Mohammed Yusuf Ansari, Yin Yang, Pramod Kumar Meher et Sarada Prasad Dakua. « Dense-PSP-UNet : A neural network for fast inference liver ultrasound segmentation ». In : *Computers in Biology and Medicine* 153 (2023), page 106478.

[9] Bruno Antonio, Davide Moroni et Massimo Martinelli. « Efficient adaptive ensembling for image classification ». In : *Expert Systems* n/a.n/a (). doi : https://doi.org/10.1111/exsy.13424. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1111/exsy.13424. url : https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13424.

[10] Ana Ruth Araújo, Natalia Rosso, Giorgio Bedogni, Claudio Tiribelli et Stefano Bellentani. « Global epidemiology of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis : What we need in the future ». In : *Liver International* (2018). doi : https://doi.org/10.1111/liv.13643.

[11] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić et Cordelia Schmid. « Vivit : A video vision transformer ». In : *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pages 6836-6846.

[12] Alireza Askarzadeh. « A Novel Metaheuristic Method for Solving Constrained Engineering Optimization Problems : Crow Search Algorithm ». In : *Computers & Structures* 169 (2016), pages 1-12.

[13] Ambarish M Athavale, Peter D Hart, Mathew Itteera, David Cimbaluk, Tushar Patel, Anas Alabkaa, Jose Arruda, Ashok Singh, Avi Rosenberg et Hemant Kulkarni. « Development and validation of a deep learning model to quantify interstitial fibrosis and tubular atrophy from kidney ultrasonography images ». In : *JAMA Network Open* 4.5 (2021), e2111176-e2111176.

[14] Gelan Ayana et Se-woon Choe. « BUViTNet : Breast Ultrasound Detection via Vision Transformers ». In : *Diagnostics* 12.11 (2022). issn : 2075-4418. doi : 10.3390/diagnostics12112654. url : https://www.mdpi.com/2075-4418/12/11/2654.

[15] Hilda Azimi, Ashkan Ebadi, Jessy Song, Pengcheng Xi et Alexander Wong. « COVID-Net UV : An End-to-End Spatio-Temporal Deep Neural Network Architecture for Automated Diagnosis of COVID-19 Infection from Ultrasound Videos ». In : *arXiv preprint arXiv :2205.08932* (2022).

[16] Ki Bae Bang et Yong Kyun Cho. « Comorbidities and metabolic derangement of NAFLD ». In : *Journal of lifestyle medicine* 5.1 (2015), pages 7-3. doi : https://doi.org/10.15280/jlm.2015.5.1.7.

[17] Bruno Barros, Paulo Lacerda, Celio Albuquerque et Aura Conci. « Pulmonary COVID-19 : learning spatiotemporal features combining CNN and LSTM networks for lung ultrasound video classification ». In : *Sensors* 21 (2021), page 5486. doi : 10.3390/s21165486.

[18] Pierre Bedossa. « Pathology of non-alcoholic fatty liver disease ». In : *Liver International* 37 (2017), pages 85-89. doi : https://doi.org/10.1111/liv.13301.

[19] Ashiq Hussain Bhat et MAK Baig. « Some Coding Theorems on Generalized Reyni's Entropy of Order α and Type β ». In : *International Journal of Applied Mathematics and Information Sciences Letters* 5.1 (2016), pages 13-19.

[20] Mainak Biswas, Venkatanareshbabu Kuppili, Damodar Reddy Edla, Harman S. Suri, Luca Saba, Rui Tato Marinhoe, J. Miguel Sanches et Jasjit S. Suri. « Symtosis : A liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm ». In : *Computer Methods and Programs in Biomedicine* 155 (2018), pages 165-177. issn : 0169-2607. doi : https://doi.org/10.1016/j.cmpb.2017.12.016. url : https://www.sciencedirect.com/science/article/pii/S0169260717308416.

[21] Jannis Born, Nina Wiedemann, Manuel Cossio, Charlotte Buhre, Gabriel Brändle, Konstantin Leidermann, Julie Goulet, Avinash Aujayeb, Michael Moor, Bastian Rieck et Karsten Borgwardt. « Accelerating detection of lung pathologies with explainable ultrasound image analysis ». In : *Applied Sciences* 11 (2021), page 672. doi : 10.3390/app11020672.

[22] G. Bradski. « The OpenCV Library ». In : *Dr. Dobb's Journal of Software Tools* (2000).

[23] A. Buades, B. Coll et J.-M. Morel. « A non-local algorithm for image denoising ». In : *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Tome 2. 2005, 60-65 vol. 2. doi : 10.1109/CVPR.2005.38.

[24] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton et Greg Hullender. « Learning to rank using gradient descent ». In : *Proceedings of the 22nd International Conference on Machine Learning*. ICML '05. Bonn, Germany : Association for Computing Machinery, 2005, pages 89-96. isbn : 1595931805. doi : 10.1145/1102351.1102363. url : https://doi.org/10.1145/1102351.1102363.

[25] Christopher Burges, Robert Ragno et Quoc Le. « Learning to Rank with Nonsmooth Cost Functions ». In : *Advances in Neural Information Processing Systems*. Sous la direction de B. Schölkopf, J. Platt et T. Hoffman. Tome 19. MIT Press, 2006. doi : https://dl.acm.org/doi/10.5555/2976456.2976481. url : https://proceedings.neurips.cc/paper_files/paper/2006/file/af44c4c56f385c43f2529f9b1b018f6a-Paper.pdf.

[26] Christopher JC Burges. « From ranknet to lambdarank to lambdamart : An overview ». In : *Learning* 11.23-581 (2010), page 81.

[27] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin et Alexandr A. Kalinin. « Albumentations : Fast and Flexible Image Augmentations ». In : *Information* 11.2 (2020). issn : 2078-2489. doi :

10.3390/info11020125. URL : https://www.mdpi.com/2078-2489/11/2/125.

[28] Michal Byra, Aiguo Han, Andrew S. Boehringer, Yingzhen N. Zhang, William D. O'Brien Jr, John W. Erdman Jr, Rohit Loomba, Claude B. Sirlin et Michael Andre. « Liver Fat Assessment in Multiview Sonography Using Transfer Learning With Convolutional Neural Networks ». In : *Journal of Ultrasound in Medicine* 41.1 (2022), pages 175-184. DOI : https://doi.org/10.1002/jum.15693. eprint : https://onlinelibrary.wiley.com/doi/pdf/10.1002/jum.15693. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/jum.15693.

[29] Michał Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, Łukasz Michałowski, Rafał Paluszkiewicz, Bogna Ziarkiewicz-Wróblewska, Krzysztof Zieniewicz, Piotr Sobieraj et Andrzej Nowicki. « Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images ». In : *International journal of computer assisted radiology and surgery* 13 (2018), pages 1895-1903. DOI : https://doi.org/10.1007/s11548-018-1843-2.

[30] Simone Cammarasana, Paolo Nicolardi et Giuseppe Patanè. « Real-time denoising of ultrasound images based on deep learning ». In : *Medical & Biological Engineering & Computing* 60.8 (2022), pages 2229-2244.

[31] SM Camps, Tim Houben, Gustavo Carneiro, Christopher Edwards, Maria Antico, Matteo Dunnhofer, EGHJ Martens, JA Baeza, BGL Vanneste, EJ van Limbergen et al. « Automatic quality assessment of transperineal ultrasound images of the male pelvic region, using deep learning ». In : *Ultrasound in medicine & biology* 46.2 (2020), pages 445-454.

[32] Jiuwen Cao, Zhiping Lin, Guang-Bin Huang et Nan Liu. « Voting based extreme learning machine ». In : *Information Sciences* 185.1 (2012), pages 66-77. ISSN : 0020-0255. DOI : https://doi.org/10.1016/j.ins.2011.09.015. URL : https://www.sciencedirect.com/science/article/pii/S0020025511004725.

[33] Wen Cao, Xing An, Longfei Cong, Chaoyang Lyu, Qian Zhou et Ruijun Guo. « Application of deep learning in quantitative analysis of 2-dimensional ultrasound imaging of nonalcoholic fatty liver disease ». In : *Journal of Ultrasound in Medicine* 39.1 (2020), pages 51-59.

[34] Zhantao Cao, Guowu Yang, Qin Chen, Xiaolong Chen et Fengmao Lv. « Breast tumor classification through learning from noisy labeled ultrasound images ». In : *Medical Physics* 47.3 (2020), pages 1048-1057.

[35] Holistic Primary Care. *Confronting the Hidden Epidemic of Fatty Liver Disease* (*First Image*). https://holisticprimarycare.net/topics/chronic-disease/confronting-the-hidden-epidemic-of-fatty-liver-disease/. Accessed : 2023-12-09.

[36] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov et Sergey Zagoruyko. « End-to-End Object Detection with Transformers ». In : *Computer Vision – ECCV 2020 : 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*. Glasgow, United Kingdom : Springer-Verlag, 2020, pages 213-229. isbn : 978-3-030-58451-1. doi : 10.1007/978-3-030-58452-8_13. url : https://doi.org/10.1007/978-3-030-58452-8_13.

[37] Juan J. Cerrolaza, Craig A. Peters, Aaron D. Martin, Emmarie Myers, Nabile Safdar et Marius George Linguraru. « Quantitative Ultrasound for Measuring Obstructive Severity in Children with Hydronephrosis ». In : *The Journal of Urology* 195.4, Part 1 (2016), pages 1093-1099. issn : 0022-5347. doi : https://doi.org/10.1016/j.juro.2015.10.173. url : https://www.sciencedirect.com/science/article/pii/S0022534715052003.

[38] Naga Chalasani, Zobair Younossi, Joel E Lavine, Anna Mae Diehl, Elizabeth M Brunt, Kenneth Cusi, Michael Charlton et Arun J Sanyal. « The diagnosis and management of non-alcoholic fatty liver disease : Practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association ». In : *Hepatology* 55.6 (2012), pages 2005-2023. doi : https://doi.org/10.1002/hep.25762.

[39] Naga Chalasani, Zobair Younossi, Joel E. Lavine, Anna Mae Diehl, Elizabeth M. Brunt, Kenneth Cusi, Michael Charlton et Arun J. Sanyal. « The diagnosis and management of non-alcoholic fatty liver disease : Practice Guideline by the American Association for the Study of Liver Diseases, American College of Gastroenterology, and the American Gastroenterological Association ». In : *Hepatology* 55.6 (2012), pages 2005-2023. doi : https://doi.org/10.1002/hep.25762.

[40] Hui Che. « Improved nonalcoholic fatty liver disease diagnosis from ultrasound data based on deep learning ». Mémoire de master. Rutgers The State University of New Jersey, School of Graduate Studies, 2021.

[41] Boli Chen, Xin Huang, Lin Xiao, Zixin Cai et Liping Jing. « Hyperbolic interaction model for hierarchical multi-label classification ». In : *Proceedings of the AAAI conference on artificial intelligence*. Tome 34. 05. 2020, pages 7496-7503.

[42] Chi-Jim CHEN, Tun-Wen PAI, Hui-Huang HSU, Chien-Hung LEE, Kuo-Su CHEN et Yung-Chih CHEN. « Prediction of chronic kidney disease stages by renal ultrasound imaging ». In : *Enterprise Information Systems* 14.2 (2020), pages 178-195. DOI : 10.1080/17517575.2019.1597386.

[43] Gongping CHEN, Yu DAI, Jianxun ZHANG, Xiaotao YIN et Liang CUI. « MBA-Net : Multi-branch aware network for kidney ultrasound images segmentation ». In : *Computers in Biology and Medicine* 141 (2022), page 105140. ISSN : 0010-4825. DOI : https://doi.org/10.1016/j.compbiomed.2021.105140. URL : https://www.sciencedirect.com/science/article/pii/S0010482521009343.

[44] Jheng-Ru CHEN, Yi-Ping CHAO, Yu-Wei TSAI, Hsien-Jung CHAN, Yung-Liang WAN, Dar-In TAI et Po-Hsiang TSUI. « Clinical Value of Information Entropy Compared with Deep Learning for Ultrasound Grading of Hepatic Steatosis ». In : *Entropy* 22.9 (2020). ISSN : 1099-4300. DOI : 10.3390/e22091006. URL : https://www.mdpi.com/1099-4300/22/9/1006.

[45] Liang-Chieh CHEN, Yukun ZHU, George PAPANDREOU, Florian SCHROFF et Hartwig ADAM. « Encoder-decoder with atrous separable convolution for semantic image segmentation ». In : *Proceedings of the European conference on computer vision (ECCV)*. 2018, pages 801-818.

[46] Qi CHEN, Xiongkuo MIN, Huiyu DUAN, Yucheng ZHU et Guangtao ZHAI. « Muiqa : Image quality assessment database and algorithm for medical ultrasound images ». In : *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2021, pages 2958-2962.

[47] Tianqi CHEN et Carlos GUESTRIN. « XGBoost : A Scalable Tree Boosting System ». In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA : Association for Computing Machinery, 2016, pages 785-794. ISBN : 9781450342322. DOI : 10.1145/2939672.2939785. URL : https://doi.org/10.1145/2939672.2939785.

[48] Feng CHENG et Gedas BERTASIUS. « Tallformer : Temporal action localization with a long-memory transformer ». In : *European Conference on Computer Vision*. Springer. 2022, pages 503-521.

[49] François CHOLLET. « Xception : Deep learning with depthwise separable convolutions ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 1251-1258. DOI : 10.1109/CVPR.2017.195.

[50] Tsung-Hsien Chou, Hsing-Jung Yeh, Chun-Chao Chang, Jui-Hsiang Tang, Wei-Yu Kao, I-Chia Su, Chien-Hung Li, Wei-Hao Chang, Chun-Kai Huang, Herdiantri Sufriyana et Emily Chia-Yu Su. « Deep learning for abdominal ultrasound : A computer-aided diagnostic system for the severity of fatty liver ». In : *Journal of the Chinese Medical Association* 84.9 (2021), pages 842-850. doi : https://doi.org/10.1097/JCMA.0000000000000585.

[51] Sara Colantonio, Antonio Salvati, Claudia Caudai, Ferruccio Bonino, Laura De Rosa, Maria Antonietta Pascali, Danila Germanese, Maurizia Rossana Brunetto et Francesco Faita. « A Deep Learning Approach for Hepatic Steatosis Estimation from Ultrasound Imaging ». In : *Advances in Computational Collective Intelligence*. Sous la direction de Krystian Wojtkiewicz, Jan Treur, Elias Pimenidis et Marcin Maleszka. Cham : Springer International Publishing, 2021, pages 703-714. isbn : 978-3-030-88113-9.

[52] Elena Codruta Constantinescu, Anca-Loredana Udriștoiu, Ștefan Cristinel Udriștoiu, Andreea Valentina Iacob, Lucian Gheorghe Gruionu, Gabriel Gruionu, Larisa Săndulescu et Adrian Săftoiu. « Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images ». In : *Medical Ultrasonography* 23.2 (2021), pages 135-139.

[53] Hind Dadoun, Anne-Laure Rousseau, Eric de Kerviler, Jean-Michel Correas, Anne-Marie Tissier, Fanny Joujou, Sylvain Bodard, Kemel Khezzane, Constance de Margerie-Mellon, Hervé Delingette et Nicholas Ayache. « Deep learning for the detection, localization, and characterization of focal liver lesions on abdominal US images ». In : *Radiology : Artificial Intelligence* 4 (2022), e210110. doi : 10.1148/ryai.210110.

[54] W Dai, X Li, X Ding et KT Cheng. *Cyclical Self-Supervision for Semi-Supervised Ejection Fraction Prediction from Echocardiogram Videos* (*2022*).

[55] N. Dalal et B. Triggs. « Histograms of oriented gradients for human detection ». In : *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR'05*). Tome 1. 2005, 886-893 vol. 1. doi : 10.1109/CVPR.2005.177.

[56] Amit Das, Mary Connell et Shailesh Khetarpal. « Digital image analysis of ultrasound images using machine learning to diagnose pediatric nonalcoholic fatty liver disease ». In : *Clinical Imaging* 77 (2021), pages 62-68. issn : 0899-7071. doi : https://doi.org/10.1016/j.clinimag.2021.02.038. url : https://www.sciencedirect.com/science/article/pii/S0899707121000942.

[57] Ankan Ghosh Dastider, Farhan Sadik et Shaikh Anowarul Fattah. « An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound ». In : *Computers in Biology and Medicine* 132 (2021), page 104296.

[58] John G Daugman. « Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters ». In : *JOSA A* 2.7 (1985), pages 1160-1169.

[59] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li et Li Fei-Fei. « Imagenet : A large-scale hierarchical image database ». In : *2009 IEEE CVPR*. 2009, pages 248-255. doi : 10.1109/CVPR.2009.5206848.

[60] François Destrempes, Marc Gesnik, Boris Chayer, Marie-Hélène Roy-Cardinal, Damien Olivié, Jeanne-Marie Giard, Giada Sebastiani, Bich N Nguyen, Guy Cloutier et An Tang. « Quantitative ultrasound, elastography, and machine learning for assessment of steatosis, inflammation, and fibrosis in chronic liver disease ». In : *PLoS One* 17.1 (2022), e0262291.

[61] Jacob Devlin. « Bert : Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv :1810.04805* (2018).

[62] Michele Di Martino, Lucia Pacifico, Mario Bezzi, Rossella Di Miscio, Beatrice Sacconi, Claudio Chiesa et Carlo Catalano. « Comparison of magnetic resonance spectroscopy, proton density fat fraction and histological analysis in the quantification of liver steatosis in children and adolescents ». In : *World journal of gastroenterology* 22.39 (2016), page 8812.

[63] Marco Di Serafino, Francesca Iacobellis, Maria Laura Schillirò, Divina D'auria, Francesco Verde, Dario Grimaldi, Giuseppina Dell'Aversano Orabona, Martina Caruso, Vittorio Sabatino, Chiara Rinaldo, Pasquale Guerriero, Vito Cantisani, Gianfranco Vallone et Luigia Romano. « Common and uncommon errors in emergency ultrasound ». In : *Diagnostics* 12.3 (2022), page 631. doi : https://doi.org/10.3390/diagnostics12030631.

[64] Alexey Dosovitskiy. « An image is worth 16x16 words : Transformers for image recognition at scale ». In : *arXiv preprint arXiv :2010.11929* (2020).

[65] Abhishek Dutta et Andrew Zisserman. « The VIA annotation software for images, audio and video ». In : *Proceedings of the 27th ACM international conference on multimedia*. 2019, pages 2276-2279.

[66] Murtada K Elbashir, Alshimaa Mahmoud, Ayman Mohamed Mostafa, Eslam Hamouda, Meshrif Alruily, Sadeem M Alotaibi, Hosameldeen Shabana et Mohamed Ezz. « A Transfer Learning Approach Based on Ultrasound Images for Liver Cancer Detection. » In : *Computers, Materials & Continua* 75.3 (2023).

[67] Salehe Erfanian Ebadi, Deepa Krishnaswamy, Seyed Ehsan Seyed Bolouri, Dornoosh Zonoobi, Russell Greiner, Nathaniel Meuser-Herr, Jacob L. Jaremko, Jeevesh Kapur, Michelle Noga et Kumaradevan Punithakumar. « Automated detection of pneumonia in lung ultrasound using deep video classification for COVID-19 ». In : *Informatics in Medicine Unlocked* 25 (2021), page 100687. issn : 2352-9148. doi : `https://doi.org/10.1016/j.imu.2021.100687`. url : `https://www.sciencedirect.com/science/article/pii/S2352914821001714`.

[68] Lhuqita Fazry, Asep Haryono, Nuzulul Khairu Nissa, Sunarno, Naufal Muhammad Hirzi, Muhammad Febrian Rachmadi et Wisnu Jatmiko. « Hierarchical Vision Transformers for Cardiac Ejection Fraction Estimation ». In : *IWBIS 2022*. 2022, pages 39-44. doi : `10.1109/IWBIS56557.2022.9924664`.

[69] Christoph Feichtenhofer. « X3d : Expanding architectures for efficient video recognition ». In : *Proceedings of the IEEE/CVF CVPR*. 2020, pages 200-210. doi : `10.1109/CVPR42600.2020.00028`.

[70] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik et Kaiming He. « Slowfast networks for video recognition ». In : *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pages 6202-6211.

[71] Christoph Feichtenhofer, Axel Pinz et Richard P Wildes. « Spatiotemporal residual networks for video action recognition. corr abs/1611.02155 (2016) ». In : *arXiv preprint arXiv :1611.02155* 54 (2016), page 55.

[72] Xiangfei Feng, Wenjia Cai, Rongqin Zheng, Lina Tang, Jianhua Zhou, Hui Wang, Jintang Liao, Baoming Luo, Wen Cheng, An Wei et al. « Diagnosis of hepatocellular carcinoma using deep network with multi-view enhanced patterns mined in contrast-enhanced ultrasound data ». In : *Engineering Applications of Artificial Intelligence* 118 (2023), page 105635.

[73] Zishun Feng, Joseph A Sivak et Ashok K Krishnamurthy. « Two-stream attention spatio-temporal network for classification of echocardiography videos ». In : *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pages 1461-1465.

[74] JL Fleiss. « Measuring nominal scale agreement among many raters ». In : *Psychological bulletin* 76.5 (1971), pages 378-382. issn : 0033-2909. doi : `https://doi.org/10.1037/h0031619`.

[75] Jie Fu, Junyu Gao et Changsheng Xu. « Semantic and temporal contextual correlation learning for weakly-supervised temporal action localization ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.10 (2023), pages 12427-12443.

[76] Ahmed Gaber, Hassan A Youness, Alaa Hamdy, Hammam M Abdelaal et Ammar M Hassan. « Automatic classification of fatty liver disease based on supervised learning and genetic algorithm ». In : *Applied Sciences* 12.1 (2022), page 521.

[77] Marco Gazzoni, Marco La Salvia, Emanuele Torti, Gianmarco Secco, Stefano Perlini et Francesco Leporati. « Perceptive SARS-CoV-2 End-To-End Ultrasound Video Classification through X3D and Key-Frames Selection ». In : *Bioengineering* 10 (2023), page 282. doi : 10.3390/bioengineering10030282.

[78] Madhumala Ghosh, Chandan Chakraborty et Ajoy K Ray. « Yager's measure based fuzzy divergence for microscopic color image segmentation ». In : *2013 Indian Conference on Medical Informatics and Telemedicine (ICMIT)*. 2013, pages 13-16. doi : 10.1109/IndianCMIT.2013.6529400.

[79] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals et George E Dahl. « Neural message passing for quantum chemistry ». In : *International conference on machine learning*. PMLR. 2017, pages 1263-1272.

[80] Ross Girshick. « Fast R-CNN ». In : *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pages 1440-1448. doi : 10.1109/ICCV.2015.169.

[81] Guoqiang Gong, Liangfeng Zheng, Wenhao Jiang et Yadong Mu. « Self-Supervised Video Action Localization with Adversarial Temporal Transforms. » In : *IJCAI*. 2021, pages 693-699.

[82] Jiulian Gu, Shousheng Liu, Shuixian Du, Qing Zhang, Jianhan Xiao, Quanjiang Dong et Yongning Xin. « Diagnostic value of MRI-PDFF for hepatic steatosis in patients with non-alcoholic fatty liver disease : a meta-analysis ». In : *European radiology* 29 (2019), pages 3564-3573. doi : https://doi.org/10.1007/s00330-019-06072-4.

[83] Qing Gu, Li Cen, Jiawei Lai, Zhongchen Zhang, Jiaqi Pan, Feng Zhao, Chaohui Yu, Youming Li, Chunxiao Chen, Weixing Chen et Zhe Shen. « A meta-analysis on the diagnostic performance of magnetic resonance imaging and transient elastography in nonalcoholic fatty liver disease ». In : *European Journal of Clinical Investigation* 51.2 (2021), e13446. doi : https://doi.org/10.1111/eci.13446.

[84] Shuhang Gu, Lei Zhang, Wangmeng Zuo et Xiangchu Feng. « Weighted Nuclear Norm Minimization with Application to Image Denoising ». In : *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pages 2862-2869. doi : 10.1109/CVPR.2014.366.

[85] Anjan Gudigar, Shreesha Chokkadi, U Raghavendra et U Rajendra Acharya. « An efficient traffic sign recognition based on graph embedding features ». In : *Neural Computing and Applications* 31 (2019), pages 395-407.

[86] Anjan Gudigar, Raghavendra U, Jyothi Samanth, Mokshagna Rohit Gangavarapu, Abhilash Kudva, Ganesh Paramasivam, Krishnananda Nayak, Ru-San Tan, Filippo Molinari, Edward J. Ciaccio et U. Rajendra Acharya. « Automated detection of chronic kidney disease using image fusion and graph embedding techniques with ultrasound images ». In : *Biomedical Signal Processing and Control* 68 (2021), page 102733. issn : 1746-8094. doi : https://doi.org/10.1016/j.bspc.2021.102733. url : https://www.sciencedirect.com/science/article/pii/S174680942100330X.

[87] Sriharsha Gummadi, Nirmal Patel, Haresh Naringrekar, Laurence Needleman, Andrej Lyshchik, Patrick O'Kane, Jesse Civan et John R Eisenbrey. « Automated machine learning in the sonographic diagnosis of non-alcoholic fatty liver disease ». In : *Advanced Ultrasound in Diagnosis and Therapy* 4.3 (2020), pages 176-182.

[88] Dinghao Guo, Chunyu Lu, Dali Chen, Jizhong Yuan, Qimu Duan, Zheng Xue, Shixin Liu et Ying Huang. « A multimodal breast cancer diagnosis method based on Knowledge-Augmented Deep Learning ». In : *Biomedical Signal Processing and Control* 90 (2024), page 105843.

[89] Juncheng Guo, Jianxin Lin, Guanghua Tan, Yuhuan Lu, Zhan Gao, Shengli Li et Kenli Li. « Unsupervised Ultrasound Image Quality Assessment with Score Consistency and Relativity Co-learning ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pages 734-743.

[90] Aiguo Han, Michal Byra, Elhamy Heba, Michael P. Andre, John W. Erdman, Rohit Loomba, Claude B. Sirlin et William D. O'Brien. « Noninvasive Diagnosis of Nonalcoholic Fatty Liver Disease and Quantification of Liver Fat with Radiofrequency Ultrasound Data Using One-dimensional Convolutional Neural Networks ». In : *Radiology* 295.2 (2020). PMID : 32096706, pages 342-350. doi : 10.1148/radiol.2020191160. eprint : https://doi.org/10.1148/radiol.2020191160. url : https://doi.org/10.1148/radiol.2020191160.

[91] Xiangmin Han, Bangming Gong, Lehang Guo, Jun Wang, Shihui Ying, Shuo Li et Jun Shi. « B-mode ultrasound based CAD for liver cancers via multi-view privileged information learning ». In : *Neural Networks* 164 (2023), pages 369-381.

[92] Kensho Hara, Hirokatsu Kataoka et Yutaka Satoh. « Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? » In : *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pages 6546-6555.

[93] R.M. Haralick, K. Shanmugam et I. Dinstein. « Textural Features for Image Classification ». In : *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (1973), pages 610-621.

[94] Adam P. Harrison, Bowen Li, Tse-Hwa Hsu, Cheng-Jen Chen, Wan-Ting Yu, Jennifer Tai, Le Lu et Dar-In Tai. « Steatosis Quantification on Ultrasound Images by a Deep Learning Algorithm on Patients Undergoing Weight Changes ». In : *Diagnostics* 13.20 (2023). issn : 2075-4418. doi : 10.3390/diagnostics13203225. url : https://www.mdpi.com/2075-4418/13/20/3225.

[95] Donya Hassan et Ali Obied. « 3DCNN Model for Left Ventricular Ejection Fraction Evaluation in Echocardiography ». In : *2023 Al-Sadiq International Conference on Communication and Information Technology* (*AICCIT*). 2023, pages 1-6. doi : 10.1109/AICCIT57614.2023.10218223.

[96] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. « Deep Residual Learning for Image Recognition ». In : *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 2016, pages 770-778. doi : 10.1109/CVPR.2016.90.

[97] Kaiming He, Xiangyu Zhang, Shaoqing Ren et Jian Sun. « Deep residual learning for image recognition ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pages 770-778. doi : 10.1109/CVPR.2016.90.

[98] Xiaofei He, Deng Cai, Shuicheng Yan et Hong-Jiang Zhang. « Neighborhood preserving embedding ». In : *Tenth IEEE International Conference on Computer Vision* (*ICCV'05*) *Volume 1*. Tome 2. IEEE. 2005, pages 1208-1213.

[99] Ming-Chih Ho, Yu-Hsin Lee, Yung-Ming Jeng, Chiung-Nien Chen, King-Jen Chang et Po-Hsiang Tsui. « Relationship between ultrasound backscattered statistics and the concentration of fatty droplets in livers : an animal study ». In : *PLoS One* 8.5 (2013), e63543.

[100] S Hochreiter. « Long Short-term Memory ». In : *Neural Computation MIT-Press* (1997).

[101] Gregory Holste, Evangelos K Oikonomou, Bobak J Mortazavi, Zhangyang Wang et Rohan Khera. « Efficient deep learning-based automated diagnosis from echocardiography with contrastive self-supervised learning ». In : *Communications Medicine* 4.1 (2024), page 133.

[102] James P Howard, Jeremy Tan, Matthew J Shun-Shin, Dina Mahdi, Alexandra N Nowbar, Ahran D Arnold, Yousif Ahmad, Peter McCartney, Massoud Zolgharni, Nick WF Linton et al. « Improving ultrasound video classification : an evaluation of novel deep learning methods in echocardiography ». In : *Journal of medical artificial intelligence* 3 (2020).

[103] Gao Huang, Zhuang Liu, Laurens Van Der Maaten et Kilian Q. Weinberger. « Densely Connected Convolutional Networks ». In : *2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). 2017, pages 2261-2269. doi : 10.1109/CVPR.2017.243.

[104] Jing Huang, Tianyu Chen, Wen Jiang, Hewei Zhang et Ruoqi Wang. « Thyroid Nodule Classification in Ultrasound Videos by Combining 3D CNN and Video Transformer ». In : *2023 IEEE International Conference on Systems, Man, and Cybernetics* (*SMC*). IEEE. 2023, pages 5273-5278.

[105] Jing Huang, Ming Kong, Luyuan Chen, Tian Liang et Qiang Zhu. « Temporal RPN Learning for Weakly-Supervised Temporal Action Localization ». In : *Asian Conference on Machine Learning*. PMLR. 2024, pages 470-485.

[106] Jun Huang, Feng Qin, Xiao Zheng, Zekai Cheng, Zhixiang Yuan, Weigang Zhang et Qingming Huang. « Improving multi-label classification with missing labels by learning label-specific features ». In : *Information Sciences* 492 (2019), pages 124-146.

[107] Libing Huang, Yingying Lin, Peng Cao, Xia Zou, Qian Qin, Zhanye Lin, Fengting Liang et Zhengyi Li. « Automated detection and segmentation of pleural effusion on ultrasound images using an Attention U-net ». In : *Journal of Applied Clinical Medical Physics* 25.1 (2024), e14231.

[108] Yunwen Huang, Hongyu Hu, Ying Zhu et Yi Xu 0001. « Breast Lesion Diagnosis Using Static Images and Dynamic Video ». In : *20th IEEE International Symposium on Biomedical Imaging, ISBI 2023, Cartagena, Colombia, April 18-21, 2023*. IEEE, 2023, pages 1-5. isbn : 978-1-6654-7358-3. doi : 10.1109/ISBI53787.2023.10230620. url : https://doi.org/10.1109/ISBI53787.2023.10230620.

[109] Jaeyoung Huh, Shujaat Khan et Jong Chul Ye. « Unsupervised Learning for Acoustic Shadowing Artifact Removal in Ultrasound Imaging ». In : *2021 IEEE International Ultrasonics Symposium* (*IUS*). 2021, pages 1-4. doi : 10.1109/IUS52206.2021.9593451.

[110] In-Chang Hwang, Dongjun Choi, You-Jung Choi, Lia Ju, Myeongju Kim, Ji-Eun Hong, Hyun-Jung Lee, Yeonyee E Yoon, Jun-Bean Park, Seung-Pyo Lee et al. « Differential diagnosis of common etiologies of left ventricular hypertrophy using a hybrid CNN-LSTM model ». In : *Scientific Reports* 12.1 (2022), page 20998.

[111] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy et Thomas Brox. « Flownet 2.0 : Evolution of optical flow estimation with deep networks ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 2462-2470.

[112] Sergey Ioffe et Christian Szegedy. « Batch normalization : Accelerating deep network training by reducing internal covariate shift ». In : *Proceedings of the 32nd International Conference on Machine Learning* (*ICML-15*). 2015, pages 448-456. doi : `10.48550/arXiv.1502.03167`.

[113] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou et Alexei A Efros. « Image-to-image translation with conditional adversarial networks ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 1125-1134.

[114] Golara Javadi, Samareh Samadi, Sharareh Bayat, Samira Sojoudi, Antonio Hurtado, Silvia Chang, Peter Black, Parvin Mousavi et Purang Abolmaesumi. « Characterizing the uncertainty of label noise in systematic ultrasound-guided prostate biopsy ». In : *2021 IEEE 18th International symposium on biomedical imaging* (*ISBI*). IEEE. 2021, pages 424-428.

[115] Sun Kyung Jeon, Jeong Min Lee, Ijin Joo, Jeong Hee Yoon et Gunwoo Lee. « Two-dimensional convolutional neural network using quantitative US for noninvasive assessment of hepatic steatosis in NAFLD ». In : *Radiology* 307.1 (2023), e221510. doi : `10.1148/radiol.221510`.

[116] Younbeom Jeong, Jung Hoon Kim, Hee-Dong Chae, Sae-Jin Park, Jae Seok Bae, Ijin Joo et Joon Koo Han. « Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography : preliminary results ». In : *Scientific reports* 10 (2020), page 7700. doi : `10.1038/s41598-020-64205-y`.

[117] Xue Jiang, Yukun Luo, Xuelei He, Kun Wang, Wenjing Song, Qinggui Ye, Lei Feng, Wei Wang, Xiaojuan Hu et Hua Li. « Development and validation of the diagnostic accuracy of artificial intelligence-assisted ultrasound in the classification of splenic trauma ». In : *Annals of Translational Medicine* 10 (2022), page 1060. doi : `10.21037/atm-22-3767`.

[118] Jianbo Jiao, Richard Droste, Lior Drukker, Aris T Papageorghiou et J Alison Noble. « Self-supervised representation learning for ultrasound video ». In : *2020 IEEE 17th ISBI*. 2020, pages 1847-1850. doi : `10.1109/ISBI45749.2020.9098666`.

[119] Glenn Jocher, Ankush Chaurasia, Jiu Qiu et Alex Stoken. *YOLOv5*. `https://github.com/ultralytics/yolov5`. 2020.

[120] Yunsang Joo, Hyun-Cheol Park, O-Joun Lee, Changhan Yoon, Moon Hyung Choi et Chang Choi. « Classification of liver fibrosis from heterogeneous ultrasound image ». In : *IEEE Access* 11 (2023), pages 9920-9930.

[121] Yasser M Kadah, Amal A Farag, Jozef M Zurada, Ahmed M Badawi et A-BM Youssef. « Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images ». In : *IEEE transactions on Medical Imaging* 15.4 (1996), pages 466-478.

[122] Takeaki Kadota, Hideaki Hayashi, Ryoma Bise, Kiyohito Tanaka et Seiichi Uchida. « Deep Bayesian active learning-to-rank with relative annotation for estimation of ulcerative colitis severity ». In : *Medical Image Analysis* 97 (2024), page 103262. issn : 1361-8415. doi : `https://doi.org/10.1016/j.media.2024.103262`. url : `https://www.sciencedirect.com/science/article/pii/S1361841524001877`.

[123] Piotr Kalinowski, Rafal Paluszkiewicz, Bogna Ziarkiewicz-Wroblewska, Tadeusz Wroblewski, Piotr Remiszewski, Mariusz Grodzicki et Marek Krawczyk. « Liver function in patients with nonalcoholic fatty liver disease randomized to Roux-en-Y gastric bypass versus sleeve gastrectomy : a secondary analysis of a randomized clinical trial ». In : *Annals of surgery* 266.5 (2017), pages 738-745. doi : `https://doi.org/10.1097/SLA.0000000000002397`.

[124] Yurie Kanauchi, Masahiro Hashimoto, Naoki Toda, Saori Okamoto, Hasnine Haque, Masahiro Jinzaki et Yasubumi Sakakibara. « Automatic Detection and Measurement of Renal Cysts in Ultrasound Images : A Deep Learning Approach ». In : *Healthcare*. Tome 11. 4. MDPI. 2023, page 484.

[125] Michael Kemmler, Erik Rodner, Esther-Sabrina Wacker et Joachim Denzler. « One-class classification with Gaussian processes ». In : *Pattern Recognition* 46.12 (2013), pages 3507-3518. issn : 0031-3203. doi : `https://doi.org/10.1016/j.patcog.2013.06.005`. url : `https://www.sciencedirect.com/science/article/pii/S0031320313002574`.

[126] Shujaat Khan, Jaeyoung Huh et Jong Chul Ye. « Pushing the limit of unsupervised learning for ultrasound image artifact removal ». In : *arXiv preprint arXiv :2006.14773* (2020).

[127] Nancy Khov, Amol Sharma et Thomas R Riley. « Bedside ultrasound in the diagnosis of nonalcoholic fatty liver disease ». In : *World journal of gastroenterology* 20.22 (2014), pages 6821-6825. doi : `https://doi.org/10.3748/wjg.v20.i22.6821`.

[128] Dong-Hyun Kim et Soo-Young Ye. « Classification of chronic kidney disease in sonography using the GLCM and artificial neural network ». In : *Diagnostics* 11.5 (2021), page 864.

[129] Kyuseok Kim, Nuri Chon, Hyun-Woo Jeong et Youngjin Lee. « Improvement of ultrasound image quality using non-local means noise-reduction approach for precise quality control and accurate diagnosis of thyroid nodules ». In :

International Journal of Environmental Research and Public Health 19.21 (2022), page 13743.

[130] Taewoo Kim, Dong Hyun Lee, Eun-Kee Park, Sanghun Choi et al. « Deep learning techniques for fatty liver using multi-view ultrasound images scanned by different scanners : Development and validation study ». In : *JMIR Medical Informatics* 9.11 (2021), e30066.

[131] Diederik P Kingma. « Adam : A method for stochastic optimization ». In : *arXiv preprint arXiv :1412.6980* (2014).

[132] Thomas N Kipf et Max Welling. « Semi-supervised classification with graph convolutional networks ». In : *arXiv preprint arXiv :1609.02907* (2016).

[133] Masaaki Komatsu, Akira Sakai, Reina Komatsu, Ryu Matsuoka, Suguru Yasutomi, Kanto Shozu, Ai Dozen, Hidenori Machino, Hirokazu Hidaka, Tatsuya Arakaki et al. « Detection of cardiac structural abnormalities in fetal ultrasound videos using deep learning ». In : *Applied Sciences* 11.1 (2021), page 371.

[134] Dezhuang Kong, Shunbo Hu et Guojia Zhao. « MV-STCNet : Breast cancer diagnosis using spatial and temporal dual-attention guided classification network based on multi-view ultrasound videos ». In : *Biomedical Signal Processing and Control* 87 (2024), page 105541.

[135] Marius Köppel, Alexander Segner, Martin Wagener, Lukas Pensel, Andreas Karwath et Stefan Kramer. « Pairwise learning to rank by neural networks revisited : Reconstruction, theoretical analysis and practical performance ». In : *Machine Learning and Knowledge Discovery in Databases : European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part III*. Springer. 2020, pages 237-252. doi : `https://doi.org/10.1007/978-3-030-46133-1_15`.

[136] Yukinori Koyama et David A Brenner. « Liver inflammation and fibrosis ». In : *The Journal of clinical investigation* 127.1 (2017), pages 55-64. doi : `https://doi.org/10.1172/JCI88881.`.

[137] Alex Krizhevsky, Ilya Sutskever et Geoffrey E Hinton. « Imagenet classification with deep convolutional neural networks ». In : *Advances in neural information processing systems*. Tome 25. 2012, pages 1097-1105.

[138] Chin-Chi Kuo, Chun-Min Chang, Kuan-Ting Liu, Wei-Kai Lin, Hsiu-Yin Chiang, Chih-Wei Chung, Meng-Ru Ho, Pei-Ran Sun, Rong-Lin Yang et Kuan-Ta Chen. « Automation of the kidney function prediction and classification through ultrasound-based kidney imaging using deep learning ». In : *npj Digit. Med.* 2 (2019), page 29. doi : `10.1038/s41746-019-0104-2`.

[139] Venkatanareshbabu Kuppili, Mainak Biswas, Aswini Sreekumar, Harman S Suri, Luca Saba, Damodar Reddy Edla, Rui Tato Marinhoe, J Miguel Sanches et Jasjit S Suri. « Extreme learning machine framework for risk stratification of fatty liver disease using ultrasound tissue characterization ». In : *Journal of medical systems* 41 (2017), pages 1-20.

[140] Chul-min Lee, Mimi Kim, Bo-Kyeong Kang, Dae Won Jun et Eileen L Yoon. « Discordance diagnosis between B-mode ultrasonography and MRI proton density fat fraction for fatty liver ». In : *Scientific Reports* 13.1 (2023), page 15557. doi : https://doi.org/10.1038/s41598-023-42422-5.

[141] Pilhyeon Lee, Youngjung Uh et Hyeran Byun. « Background suppression network for weakly-supervised temporal action localization ». In : *Proceedings of the AAAI conference on artificial intelligence*. Tome 34. 07. 2020, pages 11320-11327.

[142] Seung Soo Lee et Seong Ho Park. « Radiologic evaluation of nonalcoholic fatty liver disease ». In : *World journal of gastroenterology : WJG* 20.23 (2014), pages 7392-7402. doi : https://doi.org/10.3748/wjg.v20.i23.7392.

[143] Bowen Li, Xinping Ren, Ke Yan, Le Lu, Lingyun Huang, Guotong Xie, Jing Xiao, Dar-In Tai et Adam P Harrison. « Learning from subjective ratings using auto-decoded deep latent embeddings ». In : *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021 : 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer. 2021, pages 270-280.

[144] Bowen Li, Dar-In Tai, Ke Yan, Yi-Cheng Chen, Cheng-Jen Chen, Shiu-Feng Huang, Tse-Hwa Hsu, Wan-Ting Yu, Jing Xiao, Lu Le et al. « Accurate and generalizable quantitative scoring of liver steatosis from ultrasound images via scalable deep learning ». In : *World Journal of Gastroenterology* 28.22 (2022), page 2494.

[145] Junyu Li, Han Huang, Dong Ni, Wufeng Xue, Dongmei Zhu et Jun Cheng. « MUVF-YOLOX : A Multi-modal Ultrasound Video Fusion Network for Renal Tumor Diagnosis ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pages 642-651.

[146] Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li et Yu Qiao. « Uniformer : Unified transformer for efficient spatiotemporal representation learning ». In : *arXiv preprint arXiv :2201.04676* (2022).

[147] Xuewei Li, Hongjun Wu, Mengzhu Li et Hongzhe Liu. « Multi-label video classification via coupling attentional multiple instance learning with label relation graph ». In : *Pattern Recognition Letters* 156 (2022), pages 53-59.

[148] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik et Christoph Feichtenhofer. « Mvitv2 : Improved multiscale vision transformers for classification and detection ». In : *Proceedings of the IEEE/CVF CVPR*. 2022, pages 4794-4804. doi : 10.1109/CVPR52688.2022.00476.

[149] Zhilin Li, Zilei Wang et Qinying Liu. « Weakly supervised temporal action localization with actionness-guided false positive suppression ». In : *Neural Networks* 175 (2024), page 106307.

[150] Yin-Yin Liao, Kuen-Cheh Yang, Ming-Ju Lee, Kuo-Chin Huang, Jin-De Chen et Chih-Kuang Yeh. « Multifeature analysis of an ultrasound quantitative diagnostic index for classifying nonalcoholic fatty liver disease ». In : *Scientific Reports* 6.1 (2016), page 35083. issn : 2045-2322. doi : 10.1038/srep35083. url : https://doi.org/10.1038/srep35083.

[151] Ching-Kai Lin, Chin-Wen Chen et Yun-Chien Cheng. « Using Spatio-Temporal Dual-Stream Network with Self-Supervised Learning for Lung Tumor Classification on Radial Probe Endobronchial Ultrasound Video ». In : *arXiv preprint arXiv :2305.02719* (2023). doi : https://doi.org/10.48550/arXiv.2305.02719.

[152] Ching-Kai Lin, Shao-Hua Wu, Jerry Chang et Yun-Chien Cheng. « The interpretation of endobronchial ultrasound image using 3D convolutional neural network for differentiating malignant and benign mediastinal lesions ». In : *arXiv preprint arXiv :2107.13820* (2021).

[153] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He et Piotr Dollar. « Focal Loss for Dense Object Detection ». In : *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pages 2980-2988. doi : 10.1109/ICCV.2017.324. url : https://arxiv.org/abs/1708.02002.

[154] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár et C Lawrence Zitnick. « Microsoft coco : Common objects in context ». In : *Computer Vision–ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pages 740-755.

[155] Tsung-Yu Lin, Aruni RoyChowdhury et Subhransu Maji. « Bilinear CNN models for fine-grained visual recognition ». In : *Proceedings of the IEEE international conference on computer vision*. 2015, pages 1449-1457.

[156] Yingying Lin, Pek-Lan Khong, Zhiying Zou et Peng Cao. « Evaluation of pediatric hydronephrosis using deep learning quantification of fluid-to-kidney-area ratio by ultrasonography ». In : *Abdominal Radiology* 46 (2021), pages 5229-5239.

[157] Zhanye LIN, Zhengyi LI, Peng CAO, Yingying LIN, Fengting LIANG, Jiajun HE et Libing HUANG. « Deep learning for emergency ascites diagnosis using ultrasonography images ». In : *Journal of Applied Clinical Medical Physics* 23.7 (2022), e13695. DOI : `https://doi.org/10.1002/acm2.13695`. eprint : `https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/acm2.13695`. URL : `https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.13695`.

[158] Mengxue LIU, Wenjing LI, Fangzhen GE et Xiangjun GAO. « Weakly-supervised temporal action localization using multi-branch attention weighting ». In : *Multimedia Systems* 30.5 (2024), page 260.

[159] Xiang LIU, Jia Lin SONG, Shuo Hong WANG, Jing Wen ZHAO et Yan Qiu CHEN. « Learning to Diagnose Cirrhosis with Liver Capsule Guided Ultrasound Image Classification ». In : *Sensors* 17.1 (2017). ISSN : 1424-8220. DOI : `10.3390/s17010149`. URL : `https://www.mdpi.com/1424-8220/17/1/149`.

[160] Yiman LIU, Qiming HUANG, Xiaoxiang HAN, Tongtong LIANG, Zhifang ZHANG, Xiuli LU, Bin DONG, Jiajun YUAN, Yan WANG, Menghan HU et al. « Atrial septal defect detection in children based on ultrasound video using multiple instances learning ». In : *Journal of Imaging Informatics in Medicine* (2024), pages 1-11.

[161] Ze LIU, Han HU, Yutong LIN, Zhuliang YAO, Zhenda XIE, Yixuan WEI, Jia NING, Yue CAO, Zheng ZHANG, Li DONG et al. « Swin transformer v2 : Scaling up capacity and resolution ». In : *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pages 12009-12019.

[162] Ze LIU, Yutong LIN, Yue CAO, Han HU, Yixuan WEI, Zheng ZHANG, Stephen LIN et Baining GUO. « Swin transformer : Hierarchical vision transformer using shifted windows ». In : *Proceedings of the IEEE/CVF ICCV*. 2021, pages 9992-10002. DOI : `10.1109/ICCV48922.2021.00986`.

[163] Ze LIU, Jia NING, Yue CAO, Yixuan WEI, Zheng ZHANG, Stephen LIN et Han HU. « Video swin transformer ». In : *Proceedings of the IEEE/CVF CVPR*. 2022, pages 3192-3201. DOI : `10.1109/CVPR42600.2020.00028`.

[164] Jonathan LONG, Evan SHELHAMER et Trevor DARRELL. « Fully convolutional networks for semantic segmentation ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pages 3431-3440.

[165] S MARĈELJA. « Mathematical description of the responses of simple cortical cells ». In : *JOSA* 70.11 (1980), pages 1297-1300.

[166] José MAURÍCIO, Inês DOMINGUES et Jorge BERNARDINO. « Comparing vision transformers and convolutional neural networks for image classification : A literature review ». In : *Applied Sciences* 13.9 (2023), page 5521.

[167] Mona-Sabrine Mayouf et Florence Dupin de Saint-Cyr. « GH-CNN : A new CNN for coherent hierarchical classification ». In : *International Conference on Artificial Neural Networks*. Springer. 2022, pages 669-681.

[168] Deepak Mishra, Santanu Chaudhury, Mukul Sarkar et Arvinder Singh Soin. « Ultrasound Image Segmentation : A Deeply Supervised Network With Attention to Boundaries ». In : *IEEE Transactions on Biomedical Engineering* 66.6 (2019), pages 1637-1648. doi : 10.1109/TBME.2018.2877577.

[169] Anish Mittal, Rajiv Soundararajan et Alan C. Bovik. « Making a "Completely Blind" Image Quality Analyzer ». In : *IEEE Signal Processing Letters* 20.3 (2013), pages 209-212. doi : 10.1109/LSP.2012.2227726.

[170] Umar Farooq Mohammad et Mohamed Almekkawy. « Automated detection of liver steatosis in ultrasound images using convolutional neural networks ». In : *2021 IEEE International Ultrasonics Symposium (IUS)*. IEEE. 2021, pages 1-4.

[171] P. Mohana Shankar. « A general statistical model for ultrasonic backscattering from tissues ». In : *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 47.3 (2000), pages 727-736. doi : 10.1109/58.842062.

[172] Masoud Mokhtari, Neda Ahmadi, Teresa SM Tsang, Purang Abolmaesumi et Renjie Liao. « Gemtrans : A general, echocardiography-based, multi-level transformer framework for cardiovascular diagnosis ». In : *Machine Learning in Medical Imaging*. 2024, pages 1-10. doi : 10.1007/978-3-031-45676-3_1.

[173] Masoud Mokhtari, Teresa Tsang, Purang Abolmaesumi et Renjie Liao. « EchoGNN : Explainable Ejection Fraction Estimation with Graph Neural Networks ». In : *MICCAI 2022*. Tome 13434. 2022, pages 360-369. doi : 10.1007/978-3-031-16440-8_35.

[174] Rand Muhtaseb et Mohammad Yaqub. « EchoCoTr : Estimation of the left ventricular ejection fraction from spatiotemporal echocardiography ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pages 370-379.

[175] Phuc Nguyen, Ting Liu, Gautam Prasad et Bohyung Han. « Weakly supervised action localization by sparse temporal pooling network ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pages 6752-6761.

[176] Yuanlu Ni, Yang Cong, Chengqian Zhao, Jinhua Yu, Yin Wang, Guohui Zhou et Mengjun Shen. « Active learning based on multi-enhanced views for classification of multiple patterns in lung ultrasound images ». In : *Computerized Medical Imaging and Graphics* 118 (2024), page 102454.

[177] A. Nithya, Ahilan Appathurai, N. Venkatadri, D.R. Ramji et C. Anna Palagan. « Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images ». In : *Measurement* 149 (2020), page 106952. issn : 0263-2241. doi : https://doi.org/10.1016/j.measurement.2019.106952. url : https://www.sciencedirect.com/science/article/pii/S0263224119308188.

[178] Kyoko Ohno-Matsui, Ryo Kawasaki, Jost B Jonas, Chui Ming Gemmy Cheung, Seang-Mei Saw, Virginie J M Verhoeven, Caroline C W Klaver, Muka Moriyama, Kosei Shinohara, Yumiko Kawasaki, Mai Yamazaki, Stacy Meuer, Tatsuro Ishibashi, Miho Yasuda, Hidetoshi Yamashita, Akira Sugano, Jie Jin Wang, Paul Mitchell, Tien Yin Wong et META-analysis for Pathologic Myopia (META-PM) Study Group. « International photographic classification and grading system for myopic maculopathy ». In : *American journal of ophthalmology* 159.5 (2015), 877-83.e7. doi : 10.1016/j.ajo.2015.01.022.

[179] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz et al. « Attention u-net : Learning where to look for the pancreas ». In : *arXiv preprint arXiv :1804.03999* (2018).

[180] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin et Piotr Bojanowski. « DINOv2 : Learning Robust Visual Features without Supervision ». In : *Transactions on Machine Learning Research* (2024). issn : 2835-8856. url : https://openreview.net/forum?id=a68SUt6zFt.

[181] M. Outtas, L. Zhang, O. Deforges, W. Hammidouche, A. Serir et C. Cavaro-Menard. « A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images ». In : *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*. 2016, pages 308-313. doi : 10.1109/ISIVC.2016.7894006.

[182] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley et James Y Zou. « Video-based AI for beat-to-beat assessment of cardiac function ». In : *Nature* 580 (2020), pages 252-256. doi : 10.1038/s41586-020-2145-8.

[183] Nora Ouzir, Adrian Basarab, Olivier Lairez et Jean-Yves Tourneret. « Robust optical flow estimation in cardiac ultrasound images using a sparse

representation ». In : *IEEE transactions on medical imaging* 38.3 (2018), pages 741-752.

[184] Mehri Owjimehr, Habibollah Danyali et Mohammad Sadegh Helfroush. « An improved method for liver diseases detection by ultrasound image analysis ». In : *Journal of Medical Signals & Sensors* 5.1 (2015), pages 21-29.

[185] Martin Paralic, Kamil Zelenak, Patrik Kamencay et Robert Hudec. « Automatic Approach for Brain Aneurysm Detection Using Convolutional Neural Networks ». In : *Applied Sciences* 13.24 (2023), page 13313.

[186] Szymon Płotka, Michal K Grzeszczyk, Robert Brawura-Biskupski-Samaha, Paweł Gutaj, Michał Lipa, Tomasz Trzciński et Arkadiusz Sitek. « BabyNet : residual transformer module for birth weight prediction on fetal ultrasound video ». In : *MICCAI 2022*. Tome 13434. 2022, pages 350-359. doi : 10.1007/978-3-031-16440-8_34.

[187] Szymon Płotka, Adam Klasa, Aneta Lisowska et al. « Deep learning fetal ultrasound video model match human observers in biometric measurements ». In : *Phys. Med. Biol.* 67 (2022), page 045013. doi : 10.1088/1361-6560/ac4d85.

[188] Bin Pu, Ningbo Zhu, Kenli Li et Shengli Li. « Fetal cardiac cycle detection in multi-resource echocardiograms using hybrid classification framework ». In : *Future Generation Computer Systems* 115 (2021), pages 825-836. doi : /10.1016/j.future.2020.09.014.

[189] Marie-Therese Puth, Markus Neuhäuser et Graeme D. Ruxton. « Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits ». In : *Animal Behaviour* 102 (2015), pages 77-84. issn : 0003-3472. doi : https://doi.org/10.1016/j.anbehav.2015.01.010. url : https://www.sciencedirect.com/science/article/pii/S0003347215000196.

[190] U Raghavendra, U Rajendra Acharya, Hamido Fujita, Anjan Gudigar, Jen Hong Tan et Shreesha Chokkadi. « Application of Gabor wavelet and Locality Sensitive Discriminant Analysis for automated identification of breast cancer using digitized mammogram images ». In : *Applied Soft Computing* 46 (2016), pages 151-161.

[191] Aimon Rahman et Vishal M Patel. « UltraMAE : Multi-modal Masked Autoencoder for Ultrasound Pre-training ». In : *Medical Imaging with Deep Learning*. 2024.

[192] Tawsifur RAHMAN, Mahmoud Khatib AA AL-RUWEIDI, Md Shaheenur Islam SUMON, Reema Yousef KAMAL, Muhammad EH CHOWDHURY et Huseyin C YALCIN. « Deep Learning Technique for Congenital Heart Disease Detection using Stacking-based CNN-LSTM Models from Fetal Echocardiogram : A Pilot Study ». In : *IEEE Access* (2023).

[193] Khalid RASHEED, Faraz JUNEJO, Ayesha MALIK et Muhammad SAQIB. « Automated fetal head classification and segmentation using ultrasound video ». In : *IEEE Access* 9 (2021), pages 160249-160267. DOI : 10.1109/ACCESS.2021.3131518.

[194] D Santhosh REDDY, R BHARATH et P RAJALAKSHMI. « A Novel Computer-Aided Diagnosis Framework Using Deep Learning for Classification of Fatty Liver Disease in Ultrasound Imaging ». In : *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. 2018, pages 1-5. DOI : 10.1109/HealthCom.2018.8531118.

[195] Joseph REDMON et Ali FARHADI. « YOLO9000 : better, faster, stronger ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pages 7263-7271.

[196] Huan REN, Wenfei YANG, Tianzhu ZHANG et Yongdong ZHANG. « Proposal-based multiple instance learning for weakly-supervised temporal action localization ». In : *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pages 2394-2404.

[197] Shaoqing REN, Kaiming HE, Ross GIRSHICK et Jian SUN. « Faster R-CNN : towards real-time object detection with region proposal networks ». In : *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'15. Montreal, Canada : MIT Press, 2015, pages 91-99.

[198] Alfréd RÉNYI. « On measures of entropy and information ». In : *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1 : contributions to the theory of statistics*. Tome 4. University of California Press. 1961, pages 547-562.

[199] Hadrien REYNAUD, Athanasios VLONTZOS, Benjamin HOU, Arian BEQIRI, Paul LEESON et Bernhard KAINZ. « Ultrasound video transformers for cardiac ejection fraction estimation ». In : *MICCAI 2021*. Tome 12906. 2021, pages 495-505. DOI : 10.1007/978-3-030-87231-1_48.

[200] Se-Yeol RHYOU et Jae-Chern YOO. « Cascaded Deep Learning Neural Network for Automated Liver Steatosis Diagnosis Using Ultrasound Images ». In : *Sensors* 21.16 (2021). ISSN : 1424-8220. DOI : 10.3390/s21165304. URL : https://www.mdpi.com/1424-8220/21/16/5304.

[201] Ricardo Ribeiro, Rui Marinho et João Sanches. « Global and local detection of liver steatosis from ultrasound ». In : *Conference proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Tome 2012. IEEE. 2012, pages 6547-6550.

[202] Sanyal Rinella Mary E. et Arun J. « Management of NAFLD : a stage-based approach ». In : *Nature Reviews Gastroenterology & Hepatology* 13.4 (2016), pages 196-205. doi : `https://doi.org/10.1038/nrgastro.2016.3`.

[203] Elymar C Rivas, Franklin Moreno, Alimar Benitez, Villie Morocho, Pablo Vanegas et Ruben Medina. « Hepatic Steatosis detection using the co-occurrence matrix in tomography and ultrasound images ». In : *2015 20th Symposium on Signal Processing, Images and Computer Vision* (*STSIVA*). IEEE. 2015, pages 1-7.

[204] Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah et Mei Chen. « Pivotal : Prior-driven supervision for weakly-supervised temporal action localization ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pages 22992-23002.

[205] Olaf Ronneberger, Philipp Fischer et Thomas Brox. « U-Net : Convolutional Networks for Biomedical Image Segmentation ». In : *Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*). Springer, 2015, pages 234-241. doi : `10.1007/978-3-319-24574-4_28`. url : `https://arxiv.org/abs/1505.04597`.

[206] Madhusudan Roy. « Classification of ultrasonography images of human fatty and normal livers using GLCM textural features ». In : *Current Trends in Technology and Science* (2014).

[207] Subhankar Roy, Willi Menapace, Sebastiaan Oei, Ben Luijten, Enrico Fini, Cristiano Saltori, Iris Huijben, Nishith Chennakeshava, Federico Mento, Alessandro Sentelli, Emanuele Peschiera, Riccardo Trevisan, Giovanni Maschietto, Elena Torri, Riccardo Inchingolo, Andrea Smargiassi, Gino Soldati, Paolo Rota, Andrea Passerini, Ruud J. G. van Sloun, Elisa Ricci et Libertario Demi. « Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound ». In : *IEEE transactions on medical imaging* 39 (2020), pages 2676-2687. doi : `10.1109/TMI.2020.2994459`.

[208] Luca Saba, Nilanjan Dey, Amira S Ashour, Sourav Samanta, Siddhartha Sankar Nath, Sayan Chakraborty, João Sanches, Dinesh Kumar, RuiTato Marinho et Jasjit S Suri. « Automated stratification of liver disease in ultrasound : an online accurate feature classification paradigm ». In : *Computer methods and programs in biomedicine* 130 (2016), pages 118-134.

[209] Ashirbani SAHA et Qing Ming Jonathan WU. « Utilizing Image Scales Towards Totally Training Free Blind Image Quality Assessment ». In : *IEEE Transactions on Image Processing* 24.6 (2015), pages 1879-1892. DOI : 10.1109/TIP.2015.2411436.

[210] Sergio J. SANABRIA, Amir M. PIRMOAZEN, Jeremy DAHL, Aya KAMAYA et Ahmed EL KAFFAS. « Comparative Study of Raw Ultrasound Data Representations in Deep Learning to Classify Hepatic Steatosis ». In : *Ultrasound in Medicine and Biology* 48.10 (2022), pages 2060-2078. ISSN : 0301-5629. DOI : https://doi.org/10.1016/j.ultrasmedbio.2022.05.031. URL : https://www.sciencedirect.com/science/article/pii/S0301562922004288.

[211] Mark SANDLER, Andrew HOWARD, Menglong ZHU, Andrey ZHMOGINOV et Liang-Chieh CHEN. « MobileNetV2 : Inverted residuals and linear bottlenecks ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pages 4510-4520. DOI : 10.1109/CVPR.2018.00474.

[212] D. SANTHOSH REDDY, P. RAJALAKSHMI et M.A. MATEEN. « A deep learning based approach for classification of abdominal organs using ultrasound images ». In : *Biocybernetics and Biomedical Engineering* 41.2 (2021), pages 779-791. ISSN : 0208-5216. DOI : https://doi.org/10.1016/j.bbe.2021.05.004. URL : https://www.sciencedirect.com/science/article/pii/S0208521621000590.

[213] B. SCHMAUCH, P. HERENT, P. JEHANNO, O. DEHAENE, C. SAILLARD, C. AUBÉ, A. LUCIANI, N. LASSAU et S. JÉGOU. « Diagnosis of focal liver lesions from ultrasound using deep learning ». In : *Diagnostic and interventional imaging* 100.4 (2019), pages 227-233. DOI : https://doi.org/10.1016/j.diii.2019.02.009.

[214] Krista SCHMIEDT, Georgiana SIMION et Cătălin Daniel CĂLEANU. « Preliminary results on contrast enhanced ultrasound video stream diagnosis using deep neural architectures ». In : *2022 International Symposium on Electronics and Telecommunications (ISETC)*. IEEE. 2022, pages 1-4.

[215] Caroline A. SCHNEIDER, Wayne S. RASBAND et Kevin W. ELICEIRI. « NIH Image to ImageJ : 25 years of image analysis ». In : *Nature Methods* 9.7 (2012), pages 671-675. ISSN : 1548-7105. DOI : 10.1038/nmeth.2089. URL : https://doi.org/10.1038/nmeth.2089.

[216] Boris SEKACHEV, Nikita MANOVICH, Maxim ZHILTSOV, Andrey ZHAVORONKOV, Dmitry KALININ, Ben HOFF, TOSMANOV, Dmitry KRUCHININ, Artyom ZANKEVICH, Dmitriy SIDNEV, Maksim MARKELOV, JOHANNES222, Mathis CHENUET, A-ANDRE, TELENACHOS, Aleksandr MELNIKOV, Jijoong KIM, Liron ILOUZ, Nikita GLAZOV, PRIYA4607, Rush TEHRANI, Seungwon JEONG, Vladimir SKUBRIEV, Sebastian

YONEKURA, vugia TRUONG, ZLIANG7, LIZHMING et Tritin TRUONG. *opencv/cvat : v1.1.0*. URL : `https://doi.org/10.5281/zenodo.4009388`.

[217] Ramprasaath R. SELVARAJU, Michael COGSWELL, Abhishek DAS, Ramakrishna VEDANTAM, Devi PARIKH et Dhruv BATRA. « Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization ». In : *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pages 618-626. DOI : `10.1109/ICCV.2017.74`.

[218] S. SELVARANI et P. RAJENDRAN. « Detection of Renal Calculi in Ultrasound Image Using Meta-Heuristic Support Vector Machine ». In : *Journal of Medical Systems* 43.9 (2019), page 281. DOI : `10.1007/s10916-019-1407-1`.

[219] C. E. SHANNON. « A mathematical theory of communication ». In : *The Bell System Technical Journal* 27.3 (1948), pages 379-423. DOI : `10.1002/j.1538-7305.1948.tb01338.x`.

[220] Daniel E SHEA, Sourabh KULHARE, Rachel MILLIN, Zohreh LAVERRIERE, Courosh MEHANIAN, Charles B DELAHUNT, Dipayan BANIK, Xinliang ZHENG, Meihua ZHU, Ye JI et al. « Deep learning video classification of lung ultrasound features associated with pneumonia ». In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pages 3103-3112.

[221] Sophia SHI. « A novel hybrid deep learning architecture for predicting acute kidney injury using patient record data and ultrasound kidney images ». In : *Applied Artificial Intelligence* 35.15 (2021), pages 1329-1345. DOI : `10.1080/08839514.2021.1976908`. eprint : `https://doi.org/10.1080/08839514.2021.1976908`. URL : `https://doi.org/10.1080/08839514.2021.1976908`.

[222] Georgiana SIMION, Catalin CALEANU et Patricia Andreea BARBU. « Ultrasound liver steatosis diagnosis using deep convolutional neural networks ». In : *2021 IEEE 27th International Symposium for Design and Technology in Electronic Packaging (SIITME)*. IEEE. 2021, pages 326-329.

[223] Karen SIMONYAN et Andrew ZISSERMAN. « Two-stream convolutional networks for action recognition in videos ». In : *Advances in neural information processing systems* 27 (2014).

[224] Karen SIMONYAN et Andrew ZISSERMAN. « Very deep convolutional networks for large-scale image recognition ». In : *arXiv preprint arXiv :1409.1556* (2014).

[225] Mandeep SINGH, Sukhwinder SINGH et Savita GUPTA. « An information fusion based method for liver classification using texture analysis of ultrasound images ». In : *Information Fusion* 19 (2014), pages 91-96.

[226] Yuxin Song, Zhaoming Zhong, Baoliang Zhao, Peng Zhang, Qiong Wang, Ziwen Wang, Liang Yao, Faqin Lv et Ying Hu. « Medical ultrasound image quality assessment for autonomous robotic screening ». In : *IEEE Robotics and Automation Letters* 7.3 (2022), pages 6290-6296.

[227] Jaime Lynn Speiser. « A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data ». In : *Journal of Biomedical Informatics* 117 (2021), page 103763. issn : 1532-0464. doi : `https://doi.org/10.1016/j.jbi.2021.103763`. url : `https://www.sciencedirect.com/science/article/pii/S1532046421000927`.

[228] Vidya K Sudarshan, Muthu Rama Krishnan Mookiah, U Rajendra Acharya, Vinod Chandran, Filippo Molinari, Hamido Fujita et Kwan Hoong Ng. « Application of wavelet techniques for cancer diagnosis using ultrasound images : A Review ». In : *Computers in Biology and Medicine* 69 (2016), pages 97-111. issn : 0010-4825. doi : `https://doi.org/10.1016/j.compbiomed.2015.12.006`. url : `https://www.sciencedirect.com/science/article/pii/S0010482515003911`.

[229] S Sudharson et Priyanka Kokil. « An ensemble of deep neural networks for kidney ultrasound image classification ». In : *Computer Methods and Programs in Biomedicine* 197 (2020), page 105709. doi : `10.1016/j.cmpb.2020.105709`.

[230] Anlan Sun, Zhao Zhang, Meng Lei, Yuting Dai, Dong Wang et Liwei Wang. « Boosting Breast Ultrasound Video Classification by the Guidance of Keyframe Feature Centers ». In : *MICCAI 2023*. Tome 14224. 2023, pages 441-451. doi : `10.1007/978-3-031-43904-9_43`.

[231] Yun Sun, Yu Li, Weihang Zhang, Fengju Zhang, Hanruo Liu, Wang Ningli et Huiqi Li. « Automatic diagnosis of myopic maculopathy using continuous severity ranking labels ». In : *Cluster Computing* (2024). doi : `10.1007/s10586-024-04607-z`. url : `https://link.springer.com/10.1007/s10586-024-04607-z`.

[232] Piotr M. Szczypiński, Michał Strzelecki, Andrzej Materka et Artur Klepaczko. « MaZda—A software package for image texture analysis ». In : *Computer Methods and Programs in Biomedicine* 94.1 (2009), pages 66-76. issn : 0169-2607. doi : `https://doi.org/10.1016/j.cmpb.2008.08.005`. url : `https://www.sciencedirect.com/science/article/pii/S0169260708002083`.

[233] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke et Alexander A Alemi. « Inception-v4, inception-resnet and the impact of residual connections on learning ». In : *Proceedings of the AAAI*. Tome 31. 2017. doi : `10.1609/aaai.v31i1.11231`.

[234] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke et Andrew Rabinovich. « Going deeper with convolutions ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pages 1-9. DOI : 10.1109/CVPR.2015.7298594.

[235] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens et Zbigniew Wojna. « Rethinking the Inception Architecture for Computer Vision ». In : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pages 2818-2826. URL : https://api.semanticscholar.org/CorpusID:206593880.

[236] Mingxing Tan et Quoc Le. « Efficientnet : Rethinking model scaling for convolutional neural networks ». In : *International conference on machine learning*. PMLR. 2019, pages 6105-6114.

[237] Neha Tanwar et Khandakar Faridar Rahman. « Machine Learning in liver disease diagnosis : Current progress and future opportunities ». In : *IOP Conference Series : Materials Science and Engineering* 1022.1 (2021), page 012029. DOI : 10.1088/1757-899X/1022/1/012029. URL : https://dx.doi.org/10.1088/1757-899X/1022/1/012029.

[238] Elliot B Tapper et Anna S-F Lok. « Use of liver imaging and biopsy in clinical practice ». In : *New England Journal of Medicine* 377.8 (2017), pages 756-768. DOI : https://doi.org/10.1056/NEJMra1610570.

[239] Michael Taylor, John Guiver, Stephen Robertson et Tom Minka. « SoftRank : optimizing non-smooth rank metrics ». In : *Proceedings of the 2008 International Conference on Web Search and Data Mining*. WSDM '08. Palo Alto, California, USA : Association for Computing Machinery, 2008, pages 77-86. ISBN : 9781595939272. DOI : 10.1145/1341531.1341544. URL : https://doi.org/10.1145/1341531.1341544.

[240] The GIMP Development Team. *GIMP*. Version 2.10.12. 12 juin 2019. URL : https://www.gimp.org.

[241] Sarina Thomas, Qing Cao, Anna Novikova, Daria Kulikova et Guy Ben-Yosef. « EchoNarrator : Generating natural text explanations for ejection fraction predictions ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pages 634-644.

[242] Sarina Thomas, Andrew Gilbert et Guy Ben-Yosef. « Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pages 380-390.

[243] Thodsawit Tiyarattanachai, Terapap Apiparakoon, Sanparith Marukatat, Sasima Sukcharoen, Nopavut Geratikornsupuk, Nopporn Anukulkarnkusol, Parit Mekaroonkamol, Natthaporn Tanpowpong, Pamornmas Sarakul, Rungsun Rerknimitr et al. « Development and validation of artificial intelligence to detect and diagnose liver lesions from ultrasound images ». In : *PLoS One* 16.6 (2021), e0252882.

[244] Minh Nguyen Nhat To, Fahimeh Fooladgar, Paul Wilson, Mohamed Harmanani, Mahdi Gilany, Samira Sojoudi, Amoon Jamzad, Silvia Chang, Peter Black, Parvin Mousavi et al. « LensePro : Label noise-tolerant prototype-based network for improving cancer detection in prostate ultrasound with limited annotations ». In : *International Journal of Computer Assisted Radiology and Surgery* (2024), pages 1-8.

[245] Tong Tong, Jionghui Gu, Dong Xu, Ling Song, Qiyu Zhao, Fang Cheng, Zhiqiang Yuan, Shuyuan Tian, Xin Yang, Jie Tian, Kun Wang et Tian'an Jiang. « Deep learning radiomics based on contrast-enhanced ultrasound images for assisted diagnosis of pancreatic ductal adenocarcinoma and chronic pancreatitis ». In : *BMC medicine* 20 (2022), page 74. doi : 10.1186/s12916-022-02258-8.

[246] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani et Manohar Paluri. « Learning spatiotemporal features with 3d convolutional networks ». In : *Proceedings of the IEEE international conference on computer vision*. 2015, pages 4489-4497.

[247] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun et Manohar Paluri. « A closer look at spatiotemporal convolutions for action recognition ». In : *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pages 6450-6459.

[248] Majid Vafaeezadeh, Hamid Behnam et Parisa Gifani. « Ultrasound Image Analysis with Vision Transformers—Review ». In : *Diagnostics* 14.5 (2024). issn : 2075-4418. doi : 10.3390/diagnostics14050542. url : https://www.mdpi.com/2075-4418/14/5/542.

[249] Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, Ewout W Steyerberg, Topic Group `Evaluating diagnostic tests et prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. « Calibration : the Achilles heel of predictive analytics ». In : *BMC medicine* 17.1 (2019), page 230.

[250] P Varma, C Jayasekera, RN Gibson, DL Stella et AJ Nicoll. « The changing place of liver biopsy in clinical practice : an audit of an Australian tertiary

hospital ». In : *Internal Medicine Journal* 44.8 (2014), pages 805-808. DOI : https://doi.org/10.1111/imj.12503.

[251] A VASWANI. « Attention is all you need ». In : *Advances in Neural Information Processing Systems* (2017).

[252] Petar VELIČKOVIĆ, Guillem CUCURULL, Arantxa CASANOVA, Adriana ROMERO, Pietro LIO et Yoshua BENGIO. « Graph attention networks ». In : *arXiv preprint arXiv :1710.10903* (2017).

[253] Features VENKATANATH, PRANEETH, Maruthi Chandrasekhar BH., Sumohana S. CHANNAPPAYYA et Swarup S. MEDASANI. « Blind image quality evaluation using perception based features ». In : *2015 Twenty First National Conference on Communications (NCC)* (2015), pages 1-6. URL : https://api.semanticscholar.org/CorpusID:6917137.

[254] Peng WAN, Shukang ZHANG, Wei SHAO, Junyong ZHAO, Yinkai YANG, Wentao KONG, Haiyan XUE et Daoqiang ZHANG. « Correlation-Adaptive Multi-view CEUS Fusion for Liver Cancer Diagnosis ». In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pages 188-197.

[255] Binglu WANG, Yongqiang ZHAO, Le YANG, Teng LONG et Xuelong LI. « Temporal action localization in the deep learning era : A survey ». In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[256] Limin WANG, Bingkun HUANG, Zhiyu ZHAO, Zhan TONG, Yinan HE, Yi WANG, Yali WANG et Yu QIAO. « Videomae v2 : Scaling video masked autoencoders with dual masking ». In : *Proceedings of the IEEE/CVF CVPR*. 2023, pages 14549-14560. DOI : 10.1109/CVPR52729.2023.01398.

[257] Xin WANG, Yudong CHEN et Wenwu ZHU. « A survey on curriculum learning ». In : *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pages 4555-4576.

[258] Yuchen WANG, Zhongyu LI, Xiangxiang CUI, Liangliang ZHANG, Xiang LUO, Meng YANG et Sh CHANG. « Key-frame guided network for thyroid nodule recognition using ultrasound videos ». In : *MICCAI 2022*. Tome 13434. 2022, pages 238-247. DOI : 10.1007/978-3-031-16440-8_23.

[259] Ziwen WANG, Yuxin SONG, Baoliang ZHAO, Zhaoming ZHONG, Liang YAO, Faqin LV, Bing LI et Ying HU. « A soft-reference breast ultrasound image quality assessment method that considers the local lesion area ». In : *Bioengineering* 10.8 (2023), page 940.

[260] Jonatas WEHRMANN, Ricardo CERRI et Rodrigo BARROS. « Hierarchical multi-label classification networks ». In : *International conference on machine learning*. PMLR. 2018, pages 5075-5084.

[261] Ross WIGHTMAN. *PyTorch Image Models*. `https://github.com/rwightman/pytorch-image-models`. 2019. DOI : `10.5281/zenodo.4414861`.

[262] Ross WIGHTMAN, Hugo TOUVRON et Hervé JÉGOU. « Resnet strikes back : An improved training procedure in timm ». In : *arXiv preprint arXiv :2110.00476* (2021).

[263] Baoyuan WU, Zhilei LIU, Shangfei WANG, Bao-Gang HU et Qiang JI. « Multi-label learning with missing labels ». In : *2014 22nd International conference on pattern recognition*. IEEE. 2014, pages 1964-1968.

[264] Hong WU, Juan FU, Hongsheng YE, Yuming ZHONG, Xuebin ZHOU, Jianhua ZHOU et Yi WANG. « Multi-modality transrectal ultrasound vudei classification for identification of clinically significant prostate cancer ». In : *arXiv preprint arXiv :2402.08987* (2024).

[265] Jiawei WU, Fulong LIU, Weiqin SUN, Zhipeng LIU, Hui HOU, Rui JIANG, Haowei HU, Peng REN, Ran ZHANG et Xiao ZHANG. « Boundary-aware convolutional attention network for liver segmentation in ultrasound images ». In : *Scientific Reports* 14.1 (2024), page 21529.

[266] Jiaxiang WU, Pan ZENG, Peizhong LIU et Guorong LV. « Automatic classification method of liver ultrasound standard plane images using pre-trained convolutional neural network ». In : *Connection Science* 34.1 (2022), pages 975-989.

[267] Miao WU, Chuanbo YAN, Xiaorong WANG, Qian LIU, Zhihua LIU et Tao SONG. « Automatic classification of hepatic cystic echinococcosis using ultrasound images and deep learning ». In : *Journal of Ultrasound in Medicine* 41 (2022), pages 163-174. DOI : `https://doi.org/10.1002/jum.15691`.

[268] Stephen Gang WU, Forrest Sheng BAO, Eric You XU, Yu-Xuan WANG, Yi-Fan CHANG et Qiao-Liang XIANG. « A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network ». In : *2007 IEEE International Symposium on Signal Processing and Information Technology*. 2007, pages 11-16. DOI : `10.1109/ISSPIT.2007.4458016`.

[269] Ianto Lin XI, Jing WU, Jing GUAN, Paul J ZHANG, Steven C HORII, Michael C SOULEN, Zishu ZHANG et Harrison X BAI. « Deep learning for differentiation of benign and malignant solid liver lesions on ultrasonography ». In : *Abdominal Radiology* 46 (2021), pages 534-543. DOI : `https://doi.org/10.1007/s00261-020-02564-w`.

[270] Yiming XU, Bowen ZHENG, Xiaohong LIU, Tao WU, Jinxiu JU, Shijie WANG, Yufan LIAN, Hongjun ZHANG, Tong LIANG, Ye SANG, Rui JIANG, Guangyu WANG, Jie REN et Ting CHEN. « Improving artificial intelligence pipeline for liver malignancy diagnosis using ultrasound images and video frames ». In : *Briefings in Bioinformatics* 24 (2022), bbac569. DOI : `10.1093/bib/bbac569`.

[271] Li-Yun Xue, Zhuo-Yun Jiang, Tian-Tian Fu, Qing-Min Wang, Yu-Li Zhu, Meng Dai, Wen-Ping Wang, Jin-Hua Yu et Hong Ding. « Transfer learning radiomics based on multimodal ultrasound imaging for staging liver fibrosis ». In : *European radiology* 30 (2020), pages 2973-2983.

[272] Jingkang Yang, Kaiyang Zhou, Yixuan Li et Ziwei Liu. « Generalized out-of-distribution detection : A survey ». In : *International Journal of Computer Vision* (2024), pages 1-28.

[273] Shi Yin, Qinmu Peng, Hongming Li, Zhengqiang Zhang, Xinge You, Katherine Fischer, Susan L. Furth, Yong Fan et Gregory E. Tasian. « Multi-instance Deep Learning of Ultrasound Imaging Data for Pattern Classification of Congenital Abnormalities of the Kidney and Urinary Tract in Children ». In : *Urology* 142 (2020), pages 183-189. issn : 0090-4295. doi : `https://doi.org/10.1016/j.urology.2020.05.019`. url : `https://www.sciencedirect.com/science/article/pii/S0090429520305719`.

[274] Shi Yin, Qinmu Peng, Hongming Li, Zhengqiang Zhang, Xinge You, Hangfan Liu, Katherine Fischer, Susan L Furth, Gregory E Tasian et Yong Fan. « Multi-instance deep learning with graph convolutional neural networks for diagnosis of kidney diseases using ultrasound imaging ». In : *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging and Clinical Image-Based Procedures : First International Workshop, UNSURE 2019, and 8th International Workshop, CLIP 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 8*. Springer. 2019, pages 146-154.

[275] Zobair M Younossi, Aaron B Koenig, Dinan Abdelatif, Yousef Fazel, Linda Henry et Mark Wymer. « Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes ». In : *Hepatology* 64.1 (2016). doi : `https://doi.org/10.1002/hep.28431`.

[276] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar et Inderjit Dhillon. « Large-scale multi-label learning with missing labels ». In : *International conference on machine learning*. PMLR. 2014, pages 593-601.

[277] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga et George Toderici. « Beyond short snippets : Deep networks for video classification ». In : *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pages 4694-4702.

[278] H Zamanian, A Mostaar, P Azadeh et M Ahmadi. « Implementation of combinational deep learning algorithm for non-alcoholic fatty liver classification in ultrasound images ». In : *Journal of biomedical physics & engineering* 11.1 (2021), page 73.

[279] Yuanyi ZENG, Xiaoyu CHEN, Yi ZHANG, Lianfa BAI et Jing HAN. « Dense-U-Net : densely connected convolutional network for semantic segmentation with a small number of samples ». In : *Tenth international conference on graphics and image processing* (*ICGIP 2018*). Tome 11069. SPIE. 2019, pages 665-670.

[280] Can ZHANG, Meng CAO, Dongming YANG, Jie CHEN et Yuexian ZOU. « Cola : Weakly-supervised temporal action localization with snippet contrastive learning ». In : *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pages 16010-16019.

[281] Chen ZHANG, Xiangyao DENG et Sai Ho LING. « Next-Gen medical imaging : U-Net evolution and the rise of transformers ». In : *Sensors* 24.14 (2024), page 4668.

[282] Huili ZHANG, Lehang GUO, Juncheng LI, Jun WANG, Shihui YING et Jun SHI. « Multi-View disentanglement-based bidirectional generalized distillation for diagnosis of liver cancers with ultrasound images ». In : *Information Processing & Management* 61.6 (2024), page 103855.

[283] Huili ZHANG, Lehang GUO, Jun WANG, Shihui YING et Jun SHI. « Multi-view feature transformation based SVM+ for computer-aided diagnosis of liver cancers with ultrasound images ». In : *IEEE Journal of Biomedical and Health Informatics* 27.3 (2023), pages 1512-1523.

[284] Jiansong ZHANG, Yongjian CHEN, Pan ZENG, Yao LIU, Yong DIAO et Peizhong LIU. « Ultra-attention : automatic recognition of liver ultrasound standard sections based on visual attention perception structures ». In : *Ultrasound in Medicine & Biology* 49.4 (2023), pages 1007-1017.

[285] Lei ZHANG, Haijiang ZHU et Tengfei YANG. « Deep Neural Networks for fatty liver ultrasound images classification ». In : *2019 Chinese Control And Decision Conference* (*CCDC*). 2019, pages 4641-4646. DOI : 10.1109/CCDC.2019.8833364.

[286] Lingfeng ZHANG, Shishir K SHAH et Ioannis A KAKADIARIS. « Hierarchical multi-label classification using fully associative ensemble learning ». In : *Pattern Recognition* 70 (2017), pages 89-103.

[287] Min-Ling ZHANG et Zhi-Hua ZHOU. « A review on multi-label learning algorithms ». In : *IEEE transactions on knowledge and data engineering* 26.8 (2013), pages 1819-1837.

[288] Siyuan ZHANG, Yifan WANG, Jiayao JIANG, Jingxian DONG, Weiwei YI et Wenguang HOU. « CNN-Based Medical Ultrasound Image Quality Assessment ». In : *Complexity* 2021.1 (2021), page 9938367.

[289] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin et Jian Sun. « ShuffleNet : An Extremely Efficient Convolutional Neural Network for Mobile Devices ». In : *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pages 6848-6856. doi : 10.1109/CVPR.2018.00716.

[290] Xin Zhang, Rabab Abdelfattah, Yuqi Song et Xiaofeng Wang. « An effective approach for multi-label classification with missing labels ». In : *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*. IEEE. 2022, pages 1713-1720.

[291] Feipeng Zhao et Yuhong Guo. « Semi-supervised multi-label learning with incomplete labels ». In : *Twenty-fourth international joint conference on artificial intelligence*. 2015.

[292] Guojia Zhao, Dezhuag Kong, Xiangli Xu, Shunbo Hu, Ziyao Li et Jiawei Tian. « Deep learning-based classification of breast lesions using dynamic ultrasound video ». In : *European Journal of Radiology* 165 (2023), page 110885.

[293] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang et Jiaya Jia. « Pyramid Scene Parsing Network ». In : *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pages 6230-6239. doi : 10.1109/CVPR.2017.660.

[294] Qiang Zheng, Susan L Furth, Gregory E Tasian et Yong Fan. « Computer-aided diagnosis of congenital abnormalities of the kidney and urinary tract in children based on ultrasound imaging data by integrating texture image features and deep transfer learning image features ». In : *Journal of pediatric urology* 15.1 (2019), 75-e1. doi : https://doi.org/10.1016/j.jpurol.2018.10.020.

[295] Qiang Zheng, Steven Warner, Gregory Tasian et Yong Fan. « A Dynamic Graph Cuts Method with Integrated Multiple Feature Maps for Segmenting Kidneys in 2D Ultrasound Images ». In : *Academic Radiology* 25.9 (2018), pages 1136-1145. issn : 1076-6332. doi : https://doi.org/10.1016/j.acra.2018.01.004. url : https://www.sciencedirect.com/science/article/pii/S1076633218300163.

[296] Hongyu Zhou, Jianmin Ding, Yan Zhou, Yandong Wang, Lei Zhao, Cho-Chiang Shih, Jingping Xu, Jianan Wang, Ling Tong, Zhouye Chen et al. « Malignancy diagnosis of liver lesion in contrast enhanced ultrasound using an end-to-end method based on deep learning ». In : *BMC Medical Imaging* 24.1 (2024), page 68.

[297] Yue Zhou, Houjin Chen, Yanfeng Li, Qin Liu, Xuanang Xu, Shu Wang, Pew-Thian Yap et Dinggang Shen. « Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images ». In : *Medical Image Analysis* 70 (2021), page 101918.

[298] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh et Jianming Liang. « Unet++ : Redesigning skip connections to exploit multiscale features in image segmentation ». In : *IEEE transactions on medical imaging* 39.6 (2019), pages 1856-1867.

[299] Jun-Yan Zhu, Taesung Park, Phillip Isola et Alexei A. Efros. « Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks ». In : *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pages 2242-2251. doi : 10.1109/ICCV.2017.244.

[300] Minyan Zhu, Liyong Ma, Wenqi Yang, Lumin Tang, Hongli Li, Min Zheng et Shan Mou. « Elastography ultrasound with machine learning improves the diagnostic performance of traditional ultrasound in predicting kidney fibrosis ». In : *Journal of the Formosan Medical Association* 121.6 (2022), pages 1062-1072. issn : 0929-6646. doi : https://doi.org/10.1016/j.jfma.2021.08.011. url : https://www.sciencedirect.com/science/article/pii/S0929664621003879.