Deanship of Graduate Studies
Al-Quds University

A Systematic Comparative Evaluation of Classical Machine Learning Algorithms for Liver Lesion Classification in Ultrasound Imaging

Aesha Loay Ebrahim Enairat

M.Sc. Thesis

Jerusalem – Palestine

1447/ 2026

A Systematic Comparative Evaluation of Classical Machine Learning Algorithms for Liver Lesion Classification in Ultrasound Imaging

Prepared By:
Aesha Loay Ebrahim Enairat

B.Sc. of Medical Imaging / An-Najah National University / Palestine

Supervisor: Prof. Radwan Qasrawi

This Thesis was submitted in partial fulfillment of requirements for the degree of Master of Medical Imaging Technology, Faculty of Graduate studies, Al-Quds University.

1447/ 2026

Al-Quds University

Deanship of Graduate Studies

Thesis Approval

A Systematic Comparative Evaluation of Classical Machine Learning Algorithms for Liver Lesion Classification in Ultrasound Imaging

Prepared by: **Aesha Loay Ebrahim Enairat**

Registration No: **22310956**

Supervisor: **Prof. Radwan Qasrawi**

The master's thesis was submitted and accepted on: 14/01/2026

The names and signatures of the examining committee members are as follows:

Head of Committee: **Prof. Radwan Qasrawi**	Signature:	
Internal Examiner: **Dr. Hussein ALMasri**	Signature:	
External Examiner: **Dr. Iyad Tumar**	Signature:	

Jerusalem – Palestine
1447/ 2026

Dedication

I dedicate this work to my homeland, Palestine; to the capital of the heart, Jerusalem; to my mother city, Jenin; and to Gaza, its steadfast people and its martyrs. I also dedicate it to my parents, Loay and Nehal Nairat, the source of my ambition and the strength behind every step I have taken; to my sisters and brother, the source of my constant support in this journey; and to my grandmother, Aesha, and my aunt, Afaf may Allah grant them mercy and eternal peace. I extend my sincere dedication as well to the members of the Palestinian Students Scholarship Fund (PSSF), who funded my master's studies and make this achievement possible. To all of you, I humbly dedicate this work.

Declaration

I certify that this thesis submitted for the degree of Master is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or Institution.

Signed: 

Aesha Loay Ebrahim Enairat

Date: 14/01/2026

Acknowledgements

I express my profound gratitude to Allah for facilitating the completion of this work. I would like to convey my sincere appreciation to my outstanding supervisor, Professor Radwan Qasrawi, for his guidance, unwavering support, constructive feedback, and valuable mentorship throughout all stages of this thesis. I am also grateful to Mr. Suliman Thwib for his continuous assistance and support. Furthermore, I would like to extend my sincere appreciation to the Department of Medical Imaging Technology at Al-Quds University for providing an encouraging academic environment and for supporting my journey. I also express my deep gratitude to the examination committee members who supervised and assessed my thesis, Dr. Hussein ALMasri and Dr. Iyad Tumar, for their valuable time, insightful comments, and constructive evaluation. I extend my sincere gratitude to An-Najah National University Hospital (NNUH), particularly Prof. Sa'ed Zyoud, Dr. Jihad Hamaida, and Mr. Ahmed Abu Ali, as well as the Patient's Friends Society (PFS), for their collaboration, cooperation, and commitment throughout the data collection process. To all those who have facilitated my path and contributed to the completion of this thesis in any way. Your assistance and encouragement have been deeply appreciated.

Abstract

Liver ultrasound is very popular as it is accessible, safe and inexpensive, but the distinction between benign and malignant liver lesions is still difficult to make, owing to the noise, low contrast, and the overlap of the lesions. The thesis is an evaluation of the performance and generalization of classical machine learning techniques in liver lesion classification with ultrasound images in 3 binary tasks, where one of them is benign-normal, another one is malignant-normal, and the last one is benign-malignant.

We merged a localized clinical dataset with a publicly available dataset in Zenodo, which created an original set of 6,791 ultrasound images. By eliminating the duplicate and very similar images to avoid redundancy we were left with a curated set of unique images numbered 2,387. We compared the default preprocessing with contrast enhancement with Contrast Limited Adaptive Histogram Equalization (CLAHE) and tested various traditional classifiers, including ensemble model, support vector machines, linear model, instance-based model as well as probabilistic models. Stratified cross-validation, independent testing and a fully isolated holdout set were used in the evaluation of model performance.

The best and consistent performance of the ensemble-based classifiers was observed especially when it came to malignant- normal and benign- normal classification. Conversely, benign-malignant classification was the hardest to carry out because the lesion types had large visual overlaps. The use of CLAHE resulted in better sensitivity and lesion separability of a number of models, and task-specific advantages.

On the whole, the findings suggest that classical machine learning pipelines with the proper support of preprocessing and strict validation may be effective in terms of assisting the classification of ultrasound-based liver lesions and may serve as the basis of further advancement.

Keywords: Liver ultrasound; Liver lesion classification; Classical machine learning; Contrast Limited Adaptive Histogram Equalization (CLAHE); Ensemble learning; Support Vector Machines (SVM); Cross-validation; Holdout validation.

Table of contents

List of Figures

List of Tables

Abbreviations

AdaBoost	Adaptive Boosting
AI	Artificial Intelligence
AUC	Area Under Curve
CAD	Computer-Aided Diagnosis
CART	Classification and Regression Tree
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
FNH	Focal Nodular Hyperplasia
GBM	Gradient Boosting Machines
GLCM	Gray-Level Co-occurrence Matrix
HCC	Hepatocellular Carcinoma
KNN	K-Nearest Neighbors
LAB	Lightness A B (color space)
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
ML	Machine Learning
NPV	Negative Predictive Value
PPV	Positive Predictive Value
RBF	Radial Basis Function
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

1 Chapter One

Introduction

The clinical and methodological bases of the present study are described in this chapter. The chapter discusses the clinical significance of classification of liver lesions, the application of ultrasound imaging modalities and their limitations in the assessment of the liver, and an increasing interest in machine learning (ML) as a potential decision-support tool for medical imaging. Thereafter, it defines the research problem, objectives, questions, and hypotheses that oriented the study. With these sections, the justification for a structured comparative study on various ML algorithms for ultrasound-based classification of liver lesions is made; this will also provide a context for the results highlighted in the chapters to follow.

1.1 Thesis Outline

This thesis is divided into six chapters. Chapter One provides the background of the study, brings out the research problem and states the objectives, research questions, hypotheses and justification. Chapter Two gives a clinical and methodological background regarding the liver ultrasound imaging, image characteristics, and important concepts that relate to machine learning-based classification. Chapter Three presents the literature that has been published on the classical ML methods used in medical image analysis of ultrasound and liver lesion classification. Chapter Four explains the study methodology, such as the process of acquiring data, the preprocessing tasks, quality control, model training, and evaluation plan. Chapter Five refers to the experimental findings and compares the performance of the models in the three binary classification problems with and without Contrast Limited Adaptive Histogram Equalization (CLAHE) preprocessing. Lastly, Chapter Six will cover the findings in relation to the past research, clinical and methodological implications, limitations and summarize the thesis with recommendations on further research.

1.2 Overview

Liver diseases posed a major global health problem, and liver cancer is one of the leading causes of cancer deaths in the world (Kazi et al., 2024). Accurate characterization of liver lesions is, therefore, at the center of clinical decision-making, for prognosis and treatment options differ tremendously in benign versus malignant diseases (Barnard Giustini et al., 2023).
Medical imaging plays a crucial role, and ultrasound is still the most widely used first-line procedure for the examination of the liver due to its availability, safety, and low cost (Ronot et al., 2018).

Ultrasound has become the back bone of liver screening and surveillances as well as the initial assessment of lesions performed in a clinical practice environment with restricted access to high-level imaging techniques (Chou et al., 2015).

Liver lesions represent a wide spectrum of focal abnormalities within the hepatic parenchyma and are commonly classified as benign, malignant, or normal reference tissue in imaging-based examinations. Incidental findings of often benign lesions such as hemangiomas and focal nodular hyperplasia are conservatively managed, whereas malignant lesions, most notably hepatocellular carcinoma, exhibit the most aggressive biological behavior and dismal prognosis when diagnosed at advanced stages (Forner et al., 2018). Early detection of malignant liver lesions influences patient outcomes because curative options, especially surgical resection, ablation, or transplantation, are widely applicable to an early stage of the disease only (Llovet et al., 2021). Therefore, reliable differentiation of benign and malignant lesions still remains the main objective of diagnosis and is a considerable challenge in clinical imaging.

The global burden of liver cancer accentuates the need for a proper diagnostic approach. Hepatocellular carcinoma is the most common form of primary liver cancer and is a serious public health concern. In 2022, liver cancer was the sixth most common cancer globally, with about 865,000 new cases, and the third most frequent cause of cancer deaths, with nearly 758,000 deaths (Bray et al., 2024). Incidence and mortality are consistently high among men and in low- and middle-income areas. Regional disparities result from the differences in risk factors like chronic infections with hepatitis B and C, alcohol-related liver diseases, and metabolic disorders. In several areas in Asia, Africa, and Latin America, liver cancer continues to be one of the major causes of cancer deaths among men where advanced cross-sectional imaging is limited (Bray et al., 2024). In these contexts, dependence on ultrasound for liver surveillance and diagnosis is particularly high, making the need for improved ultrasound-based diagnostic accuracy even more relevant.

Ultrasound imaging finds itself strategically in the assessment of liver diseases because of its non-invasiveness, wide availability, and real-time imaging capability. It is used commonly for screening and surveillance in the high-risk groups where defined, such as cirrhotic patients, for whom regular ultrasound examinations are being encouraged to facilitate early tumor detection and the chances of improved survival (Qurashi et al., 2024). Grayscale ultrasound materially contributes to everyday clinical workflows by providing initial triage for patients, many of whom present with no focal abnormality on ultrasound and hence require no further diagnostic evaluation. Visual assessment rests on lesion morphology, echogenicity, and size. However, interpretation of ultrasound data is subjective by nature and highly operator dependent. Image quality is further affected by the operator's skill level, patient characteristics, and scanning condition (Kirk et al., 2025). All these factors affect the way in which lesions are characterized and hence make it difficult to differentiate the findings as either benign or malignant, especially in cases where the visual features start to overlap. Our findings in this study corroborate those restrictions whereby we noticed consistently lower performances in classification when it came to benign-malignant discrimination as opposed to differentiation between malignant versus normal.

With regard to these issues, ML has emerged as an important avenue for data-driven image analysis and decision support in medical imaging. Machine-learning algorithms learn discriminative patterns from labeled data and apply consistent decision rules across cases, giving rise to objective outputs that can support the human interpretive process (Dana et al., 2020). In ultrasound imaging, machine-learning methods are a potential solution for automating lesion detection, classification, and risk stratification. Texture-and statistical feature sets extracted from grayscale ultrasound images provided means of capturing subtle image features that may not be readily perceived through visual assessment alone (Maryam et al., 2022). These were keen attempts to reduce inter-observer variability and enhance diagnostic consistency.

The performance variations captured in the literature from the work reported by many results from differences in dataset composition, preprocessing strategies, model selection, evaluation protocols, and so on (Renard et al., 2020). Such limitations in terms of algorithm consideration or lack of validation by an independent study prevent proper assessment of robustness and generalizability (Simkó et al., 2024). This study addresses those shortcomings by applying a controlled comparative framework under which multiple ML algorithms are evaluated after having undergone similar preprocessing and validation conditions. Thus, the adoption of theoretical-methodological choices regarding performance upon clinically relevant classification tasks helps strengthen its findings within actual ultrasound workflows and guide better understanding of the strengths and limitations of ML as applied to liver lesion classification.

1.3 Problem Statement

Ultrasounds have gained popularity for the evaluation of the liver, whereas the reliable classification of liver lesions remains an enormous challenge. Amongst many visual interpretations, operator dependency, image quality variability, and major overlap in the appearance of benign and malignant lesions all play their relevant roles. These limitations may delay a diagnosis or lead to unnecessary follow-up procedures.

The world of ML is full of promise for automated image analysis, yet existing studies report conflicting results and often have ill-defined standardized evaluation protocols. There is an evident need for a systematic comparison of heterogeneous ML algorithms under a comparable set of preprocessing and validation conditions with clinically relevant classification tasks.

In this study, the mentioned need is fulfilled by evaluating various ML techniques comparatively with respect to their performance for benign versus normal, malignant versus normal, and benign versus malignant classification tasks while also evaluating the impact of image preprocessing techniques such as CLAHE on these classifier performances.

1.4 Research Objectives

The major aim of this research is to assess the feasibility of using ML approaches to classify liver lesions based on ultrasound images. This aim will be attained through the following specific objectives:

1. To compare and contrast the various ML algorithms with different paradigms of learning therein giving them the opportunity to classify liver ultrasound images into the clinically relevant categories.
2. To determine the effect of image preprocessing on the model performance with special consideration to CLAHE.
3. To evaluate the model performance across three clinically meaningful binary classification tasks, benign versus normal, malignant versus normal, and benign versus malignant.
4. Risk and generalization capability of developed models are evaluated through stratified cross-validation, independent test sets, and fully isolated holdout data sets.

1.5 Research Questions

This study is guided by the following research questions:

1. How do different ML algorithms perform in classifying liver lesions from ultrasound images across clinically relevant binary classification tasks?

2. How does contrast enhancement using CLAHE influence the diagnostic performance of ML models in liver ultrasound imaging?
3. How does classification performance differ between benign–normal, malignant–normal, and benign–malignant classification scenarios?
4. To what extent do the evaluated ML models generalize when applied to independent test and holdout datasets?

1.6 Research Hypotheses

The following hypotheses are formulated to guide the analysis and interpretation of the study results:

1. ML models can achieve reliable diagnostic performance in classifying liver ultrasound images into benign, malignant, and normal categories.
2. Ensemble-based ML methods will demonstrate stronger and more stable performance than linear or probabilistic models due to their ability to model complex image patterns.
3. CLAHE-based contrast enhancement will improve classification performance by enhancing discriminative image features in ultrasound images.
4. Model performance on independent test and holdout datasets will remain consistent with cross-validation results, indicating adequate generalization and limited overfitting.

1.7 Research Justification

Especially in areas where ultrasound remains the primary imaging technique, enhancing the classification of liver lesions is critical for improving early diagnosis and outcomes for the patients. ML provides one possible avenue through which increasingly objective interpretation can be made in a more standard way. However, this would best serve the imaging community if models were robust, generalizable, and clinically aligned.
Furthermore, systematic comparisons of several algorithms and preprocessing strategies supported with stringent validation provide evidence-based insights into which approaches are most suitable for ultrasound-based liver lesion classification. The findings can be further extended in developing dependable computer-aided diagnosis systems while also paving the way for the future work oriented toward enhancing clinical decision-making in liver imaging.

This chapter introduced the clinical importance of accurate liver lesion characterization and highlighted the limitations of ultrasound in distinguishing benign from malignant lesions. It also presented the research problem, objectives, research questions, hypotheses, and justification for conducting a comparative evaluation of classical ML models. Overall, Chapter One established the motivation and scope of the study and provided the foundation for the methodological and experimental work presented in the following chapters.

2 Chapter Two

Background

The chapter establishes the clinical and methodological foundations to comprehend ultrasound-based liver lesion classification. It discusses the characteristics of focal liver lesions visualized in B-mode ultrasound and discusses the inherent diagnostic challenges that stem from overlapping echogenic and textural features. The chapter highlights some fundamental properties of images that may influence automated analysis: speckle noise, texture heterogeneity, and redundancy introduced by frame-based ultrasound data. Moreover, the key preprocessing issues will be discussed, especially considering contrast enhancement using CLAHE in terms of maximizing visibility and stability of features. Finally, this chapter covers basic principles of classical ML, ensemble strategies, and standard evaluation metrics often used in medical ultrasound studies.

2.1 Liver Lesions in Ultrasound Imaging

The areas of abnormal tissue present in the liver are liver lesions, most of which are discovered incidentally during abdominal imaging. For the initial examination of the liver, ultrasound, and in particular B-mode ultrasound, is the preferred imaging modality. In this mode, lesions appear as areas with differing echogenicity from that of surrounding liver parenchyma, which can be hypoechoic, hyperechoic, or isoechoic, according to the specific type and composition of the lesion (James et al., 2023).

Benign hepatic lesions usually include cysts, hemangiomas, focal nodular hyperplasia (FNH), and hepatic adenomas. Hemangiomas are the most common among these and usually appear as homogeneously hyperechoic nodules with posterior acoustic enhancement. FNH and other adenomas may demonstrate variable echogenicities, sometimes mimicking malignant features

(Flannery, 2020). Despite being benign, they can usually be distinguished from malignant counterparts because of intrinsic overlap in sonographic features.
The following are considered malignant liver lesions: hepatocellular carcinoma (HCC) and secondary tumors (metastases). HCC develops primarily in the background of chronic liver disease (cirrhosis) and initially presents as small hypoechoic nodules whose appearance progresses to heterogeneous or mottled as the lesion enlarges. Liver metastases can appear as hypoechoic or target-like lesions in accordance with their primary source (James et al., 2023). The ultrasound appearance of such malignant lesions can vary greatly, sometimes mimicking benign features to lay down diagnostic confusion.

In a B-mode ultrasound, the typical liver appears as a homogeneous medium-gray structure with a fine echotexture. Such a presentation is an important baseline visual reference for identifying any lesion. Disruption of the normal appearance either by mass effect, textural difference, or echogenic contrast is often the first indication of liver pathology (der Medizinischen Fakultät et al., 2018).

Though ultrasound diagnosis of liver lesions can be difficult due to high resemblance, benign and malignant lesions are often echogenic or show similar shapes. For example, some metastases may strongly resemble hemangiomas in being hyperechoic, while early-stage HCC may appear almost isoechoic, blending with normal liver tissue (Flannery, 2020). A gap exists here with the need for extra imaging or computation to assist the clinician in improving the diagnostic accuracy.

2.2 B-Mode Ultrasound Imaging

B-mode is the most popular imaging modality for medical diagnostics. It creates a 2-dimensional gray image by converting returning echo amplitude into brightness on the corresponding pixels. For each echo from tissue interfaces, a corresponding pixel is formed; brighter pixels indicate stronger echoes (Flannery, 2020).

The grayscale intensities of B-mode ultrasound are directly related to the reflection co-efficient at tissue boundaries. Therefore, the larger the acoustic impedance difference between adjacent tissues, the stronger the echo returned, resulting in the high levels of grayscale brightness. These echoes are encoded and mapped to pixels in the 2D array and rendered into the ultrasound image (James et al., 2023).

A B-mode image is such that every pixel that is displayed indicates the amplitude of reflection from the ultrasound waves. Thus, for instance, when a hyper-echoic area, such as fat or calcifications with denser structures, emerges in an image, it appears white, while hypoechoic areas with features such as fluid-filled cysts appear dark. Such mapping enables interpretation

by the clinician of anatomical and pathological structures without having to resort to invasive procedures, which, it is recorded, is according to (Al-Karawi, 2019).
Due to those, B-mode ultrasound imaging is most commonly recognized in liver imaging because of its real-time operations, low cost, non-ionizing nature, and availability. It can detect hepatic focal lesions, monitor changes in cirrhotic conditions, and assists in guiding operations for biopsy. The frontline tool of hepatocellular carcinoma surveillance in patients with chronic liver disease (Roman & Detti, 2019).

Notwithstanding its merits, B-mode ultrasound has a fair share of disadvantages. One of them is its poor contrast resolution regarding soft tissues, thus making it impossible to distinguish between benign and malignant liver lesions when their echogenicity patterns overlap. Also, the quality of the images tends to go downward when they are operator-dependent, patient habitus limited, and artifacts such as speckle noise and shadowing are present (Al-Karawi, 2019; James et al., 2023). The above-mentioned limitations are most critical when understanding computer-aided diagnosis or ML tasks based on a B-mode image.

2.3 Speckle Noise

An example of granulation interference is speckle noise, which is found prominently in ultrasound images. Differently from random noise, speckle is an example of multiplicative noise that originates from the coherence assumed by ultrasound waves, which reflect off many small, unresolved scatterers located within the tissue. So, this noise pattern is not a result of malfunctions in the equipment but is inbuilt into the very process of ultrasound imaging.

The fluctuations and mottling of ultrasound images emanate from the constructive and destructive interference of ultrasonic waves reflected from objects lesser than ultrasound wavelength. The object has different phase delays for the reflected sound waves, and when those waves will recombine to form at the detector, a fluctuating intensity pattern is formed. Therefore, a gravelly effect that is not corresponding to any real anatomical structures (Al-Karawi, 2019).

In contrast to additive random noise, which is invariably assumed to be Gaussian, speckle is signal-dependent and multiplicative. Speckle can vary spatially with tissue type and imaging parameters. Whereas random noise uniformly degrades an image, speckle has structured patterns, further complicating their filtering and denoising (Al-Karawi, 2019).

Image clarity is significantly reduced by speckle noise, especially when the subtle boundaries between tissues or lesions need to be detected. This further alters the textural properties of the tissues, which are important for feature-based analysis in ML. Speckle, thereby leading to

diagnostic confusion, may conceal or copy fine textures that would otherwise be instrumental in differentiating between benign and malignant lesions (Qasem, 2018).

When it comes to the classification of images, traditional or deep learning models primarily rely on texture and intensity. Speckle creates some artificial variation in the image which is easily misrepresented as features included in the model, thus causing either overfitting of the model or less robustness. Besides that, speckle reduces the diversity of datasets and patient populations; hence, making model validation complicated (Al-Karawi, 2019).

2.4 Texture in Liver Ultrasound Images

Texture in medical imaging implies the spatial variation pixel intensity patterns possess in tissues. It captures visual cues like smoothness, roughness, granularity, and regularity that give a hint on the underlying structure and composition of biological tissues. A texture analysis in ultrasound is important for distinguishing tissue types, detecting abnormalities, and contributing to computer-aided diagnosis systems (Roman & Detti, 2019).

Hepatic lesions usually show heterogeneous texture with different echo intensity and pattern in them. This is due to the fact that malignant tumors tend to be more disorganized and coarser in texture caused by necrosis, vascular infiltration, and rapid cell proliferation: hepatocellular carcinoma or liver metastases. The benign lesions usually have a more homogeneous or smoother-textured characteristic lesion within and are good indicators of internal heterogeneity (Roman & Detti, 2019).

On the other hand, malignant liver lesions often resemble benign ones with clinically relevant visual and textural differences in just a few cases in B-mode ultrasound. For example, focal nodular hyperplasia may appear heterogeneous because of fibrous components or vascular structures, while well-differentiated hepatocellular carcinoma may appear relatively homogeneous. This overlap limits human interpretation as well as some traditional rule-based diagnostic systems (Zhang, 2020).

Currently, texture characterization methods include statistically derived and measurable quantitative parameters such as entropy, contrast, correlation, and homogeneity that widely can be extracted for incorporation in machine-learning models for differentiating types of lesions. Thus, it will allow automated systems to eventuate a rating of minor variations that the human eye can't see, which will increase performance standards of classification and reproducibility. This makes texture analysis vital for radiologist support, along with the potential for developing CAD tools (computer-aided diagnosis) (Zhang, 2020).

2.5 Frame-Based Ultrasound Datasets

Unlike in the past, when ultrasound systems offered capture of single static images, modern systems are capable of capturing real-time video sequences. This temporal information is recorded during clinical procedures to visualize the movement of organs or vascular flow or in guidance of interventional procedures. During research, these video flows are frequently stored in DICOM or MP4 formats and later used for training ML models (Jha, 2021).

Component frames are extracted from videos at prescribed intervals (for example, every 1 to 5 frames) to develop frame-dependent datasets. This process enables one to augment the dataset in size and either apply image classification or image segmentation techniques at each frame. Such an assumption stands, however, that every frame is an independent data point; this might not hold true, particularly due to the high visual redundancy that may exist (Waibel, 2019).

Consecutives of ultrasound frames are highly similar, especially when there are minimal movements or probe holds still. There is a very high frame rate, that is 30+ fps, and hence they are quite adjacent in content, resulting in the temporal redundancy. Such similarities in the frames will lead to several problems in ML, especially in that their training and test sets will hence include frames from the same video, creating a risk of leakage (Jha, 2021).

The redundancy occurs in a situation where a dataset contains numerous frames that are almost identical. While this may insist on an artificial increase in dataset size, they might confuse the performance metrics by essentially making the classification task easier than it is in reality. In the clinical environment, the generalization means that you should really be learning from diverse anatomical and pathological patterns and not memorizing subtle variations across nearly identical frames (Waibel, 2019).

2.6 Data Leakage

This is how to define data leakage: it is when, in a sense, you unknowingly include information in the training set that is also in the test set. This results in an optimistic performance measure for the model in validation but poorly generalizes to unseen data. Leakage may occur when a part of the training process unintentionally gets access to data that may not be available at inference time (Jha, 2021).

In medical imaging, particularly in ultrasound video datasets, researchers often extract frames to evaluate model performance. Data leakage typically occurs when frames from the same patient or video appear in both the training and testing sets. Because consecutive frames tend to be visually similar, the model may learn redundant patterns, which can artificially inflate performance metrics and compromise the validity of the evaluation (Jha, 2021). This issue

becomes more critical when labels (e.g., lesion type) remain unchanged across highly similar frames.

Data leakage can affect key evaluation metrics such as accuracy, sensitivity, and AUC. Consequently, a model may appear highly accurate during validation yet fail to generalize to truly independent test data (Jha, 2021). This can lead to overconfidence in diagnostic performance and may undermine clinical trust in AI-assisted tools. Moreover, leakage reduces reproducibility and limits fair comparison across studies.

Leakage may also cause models to "cheat" by memorizing low-level textures, speckle noise patterns, or imaging artifacts instead of learning generalizable pathological features. In ultrasound imaging, where speckle noise and redundant frames are common, leakage can result in misleadingly high performance that does not translate to real-world deployment (Jha, 2021).

2.7 Contrast Enhancement

Contrast enhancement is a key preprocessing strategy in medical imaging that improves the visibility of anatomical structures by modifying pixel intensity values. In B-mode ultrasound, inherent low contrast often limits the visualization of soft tissue boundaries, particularly in liver imaging, where lesion margins may be subtle and can resemble the surrounding parenchyma. Therefore, contrast enhancement aims to improve both human interpretation and automated lesion detection (Bebis et al., 2023).

Contrast enhancement techniques are generally classified as global or local methods. Global approaches, such as standard histogram equalization, apply uniform adjustments across the entire image and may be suboptimal for heterogeneous ultrasound scans. In contrast, local methods, such as CLAHE, enhance contrast within small regions while preserving edge details and reducing noise amplification. This makes local enhancement more suitable for liver ultrasound images, where brightness and texture can vary substantially. By improving lesion boundary visibility and enhancing diagnostically relevant texture features, contrast enhancement can support more reliable feature extraction and potentially improve classification performance in both clinical and computational workflows (Bebis et al., 2023).

2.8 Contrast Limited Adaptive Histogram Equalization (CLAHE)

Contrast Limited Adaptive Histogram Equalization (CLAHE), which is an advanced image enhancement technique, would totally improve localized contrast in an image while restricting the enhancement of noise. Unlike the normal histogram equalization operation, where uniform contrast enhancement is done all over the image, CLAHE basically works with small tiles or sub-regions and performs localized histogram equalization. The method suits medical images

like ultrasound, which usually possess inferior contrast features and have high levels of noise, such as speckle (P. Singh et al., 2020a).

Generally, the standard histogram equalization techniques tend to over-enhance noise in homogeneous areas, making them doubly disadvantageous in the case of low-contrast images such as medical images. To overcome these limitations, CLAHE was developed to limit contrast enhancement using a predetermined clipping threshold, known as the "clip limit." This avoids over-saturation in areas where only minor intensity variation occurs, thus being well suited to preserving the relevant anatomical structures while suppressing the irrelevant variations (Zheng et al., 2019). CLAHE divides the image into non-overlapping tiles, then applies histogram equalization to each tile. However, prior to equalization, the histogram of each tile is clipped at a user-defined threshold, and its clipped portions are then equally redistributed back across the histogram. This contrast limiting step controls the amount by which noise is amplified in regions with uniform textures. The last image is obtained by interpolating between tile boundaries to avoid introducing any artificial edge (R. Priyah & Kamalakkannan, 2025). Liver ultrasound images are susceptible to a low dynamic range accompanied by a large amount of speckle noise, which may hinder the visualization of lesions and their boundaries. CLAHE enhances subtle intensity differences and displays features such as lesion margins, vessels, and texture more clearly, with little introduction of artifacts. It is computationally efficient and can be readily incorporated into pre-processing pipelines for classical or deep-learning-based models.(R. Priyah & Kamalakkannan, 2025).

2.9 Classical Machine Learning

Classical machine learning (ML) refers to a group of methodologic approaches from learning on the basis of structured and pre-defined feature space for predictive or classification purposes. In contrast to automated hierarchical learning on raw data performed by various deep learning methods, classical approaches rely heavily on features that are manually engineered and often rely upon domain knowledge, such as intensity, texture, or shape. Popular algorithms in this subset of ML would include logistic regression, k-nearest neighbors (KNN), decision trees, support vector machines (SVM), and the random forests (Al-Karawi, 2019).

While deep learning models, such as convolutional neural networks (CNNs), are, in fact, powerful vehicles for automated feature extraction, but they need to deal with vast labeled datasets and high resource-consumption computational power. By comparison, classical ML methods can be better interpreted, require less computational power, and are indeed candidates for solving tasks when training data are not abundant, which is often the case in medical imaging, especially ultrasound analysis. For instance, texture-based features are extracted manually from grayscale ultrasound images using gray-level co-occurrence matrices (GLCM)

and then used as inputs to classifiers like SVMs or random forests for classifying lesions (Al-Karawi, 2019).

In traditional ML, the extraction of features is basically done as a pre-processing step with expert input as to what the most relevant and informative features are. These often include statistical descriptors like mean and variance, texture measures such as entropy and contrast, along with shape and edge characteristics (Al-Karawi, 2019). Hence, the classification task usually relies on how well and discriminative the selected features are during this stage.

The assessment is limited to B-mode liver ultrasound images, which frequently exhibit poor noise properties, high image noise, low contrast, and limited availability of annotated training samples. Under those conditions, classical machine learning is particularly fruitful. It reliably facilitates work with small datasets, requires significantly less computational time for training and evaluation than deep learning, and allows interpretability, therefore, shedding light on the importance of features and reasons for decision-making. For the mentioned reasons, classical machine learning thus provides a strong and pragmatic approach to automated liver lesion classification in constrained clinical settings (Al-Karawi, 2019).

2.10 Classification Algorithms

Medical imaging is mainly dependent on these classification algorithms which serve as major tools in detecting and labeling patterns in data for automated disease detection and lesion classification in ultrasound images. Algorithms can be referred to in general classifications according to their mathematical base and learning strategy.

- Linear Classifiers: Logistic Regression

Among the linear classifiers, logistic regression is very popular for estimating the probability of class labels in recognition based on a linear combination of the entire feature set. They are easily interpretable, ensure quick training, and show great effectiveness in binary-classification problems where the benign and malignant lesions are distinguishable when a set of features can be linearly separable (Zhang, 2020).

- Distance-Based Classifiers: k-Nearest Neighbors (KNN)

KNN, a non-parametric classifier, assigns a label based on a majority vote from the nearest data points in feature space. It is intuitive, with no training phase, though performance can be sensitive to choice of distance metric and value of k. KNN is often used in ultrasound classification, where texture or intensity features are encapsulated in a multi-dimensional space (Imran et al., 2021).

- Probabilistic Classifiers: Naïve Bayes

Its classification is performed according to naive Bayes theorem, which is a probabilistic method derived from Bayes' Theorem, and has the essential assumption of independent features. The performance is quite good in many medical applications, particularly when computational efficiency is required. It will play a vital role in ultrasound tasks where the feature distributions are reasonably modeled by a probability density (Imran et al., 2021).

- Tree-Based Classifiers: Decision Trees

Decision trees create hierarchical structures for classification based on recursive partitioning of the feature space using decision rules. Although they are relatively easy to visualize and interpret, decision trees are prone to overfitting. In medical imaging, they assist in modelling non-linear decision boundaries based on combinations of features such as intensity and texture (Zhang, 2020).

- Ensemble Classifiers

The method of ensemble learning works by gathering the predictions made by several underlying learners and merging them into a prediction it believes would be more accurate and more robust to noise-for example:

1. Random Forest grows by aggregating predictions from many decision trees that are trained on bootstrapped data.
2. Extra Trees includes even more randomness in the mix.
3. AdaBoost and Gradient Boosting do the sequential correction of the learner errors.

In the presence of noise in the data that might arise from ultrasound, the methods are able to control variance while enhancing generalization of results (A. P. Singh et al., 2025).

- Kernel-Based Classifiers: Support Vector Machines (SVM)

SVMs are powerful classifiers that create optimal separating hyperplanes in high-dimensional space. Data that are not linearly separable are handled by kernel functions (e.g., RBF (Radial Basis Function), polynomial). These SVMs are often credited with achieving best performance in the ultrasound-based classification of lesions because of their resistance to overfitting in high-dimensional feature spaces (Zhang, 2020).

2.11 Ensemble Learning

Ensemble ML seeks to achieve higher accuracy and robustness through the use of multiple base classifiers. Using a single learning algorithm to base itself on creates reliance on that algorithm, whereas ensemble methods take the output predictions of many different models' learning algorithms as an input. The ensemble learning method has three sub-states: bootstrapping, boosting, and stacking. Bagging methods, like Random Forest, construct several decision trees over various bootstrap samples and combine their outputs, predominately through majority voting, so as to mitigate overfitting and variance. On the other hand, boosting algorithms, for example, AdaBoost and Gradient Boosting Machines (GBM), operate as a sequence of trained classifiers, where incorrect instances receive heavier weight so that reliability can be enhanced.

Stacking or stacked generalization is defined as a method of training several independent diverse classifiers and combining their outputs through a meta-learner that has learned the optimal way to compose predictions.(M Kumar, 2022).

These approaches have recently increased their usage in analysis and procedures involving medical images due to their robustness and flexibility in noise, class imbalance, and heterogeneity in medical imaging data. The results have shown that in ultrasound imaging where the quality of the input images is degraded by speckle noise and the anatomy is heterogeneous, the ensemble models demonstrate better performance in lesion classification tasks. Previous studies have shown that combining Random Forest and SVM classifier results by stacking or attaching them to deep learning backbone could increase diagnosis accuracy significantly in liver, breast, and pancreatic imaging (Zhang et al., 2024). Not only that, but ensemble methods are also best suited for datasets of smaller or moderate sizes, which form common scenarios in clinical research and promising to eradicate overfitting while retaining high generalizability (M Kumar, 2022).

2.12 Model Evaluation and Validation

Model evaluation and validation remain pertinent steps wherein medical image analysis would require assurance of reliability, robustness, and generalizability of predictive systems. Evaluation deals with quantifying the performance of a model with respect to the classification or prediction on a set of known outcomes, while validation measures performance over data it has never encountered before. Acceptable evaluation strategies in medical imaging, especially where diagnostic applications are concerned such as classifying liver lesions in ultrasound, are mandatory for confidence in clinical practice (González-Luna et al., 2019).

Developing a confusion matrix is one of the essential evaluation tools, one that provides us with detailed information regarding true positive, true negative, false positive, and false negative classifications. From this matrix, important parameters such as sensitivity (recall), specificity, accuracy, and F1-score can be calculated.(Abdullah et al., 2024) These functions indicate how well a model can separate diseased forms from the non-diseased ones. High sensitivity is important for the detection of malignant liver lesions so as to avoid patients slipping through the cracks; while, on the other hand, specificity stops benign cases from being misclassified.

An alternative standard metric is the ROC curve, with the true positive rate versus the false positive rate at different threshold settings on the axes. This AUC assesses performance in class discriminating capabilities, with 1.0 indicating excellent capabilities. ROCs and AUCs become more important in unbalanced datasets than relying on accuracy (González-Luna et al., 2019).

There are different ways to validate the model. One is holdout validation, in which the dataset is split into separate training and testing sets so that model evaluation is done on data not seen during training. This method is straightforward, but it may introduce bias depending on how the data is split. Another possibility is cross-validation-in this case k-fold cross-validation provides a more robust estimation of generalizability. It splits the data into k subsets and trains and validates the model k times, each time holding one subset as the validation set, thus reducing variance in performance metrics and enabling a more stable estimation of the model's true capability, especially on limited-size datasets (González-Luna et al., 2019).

Both the evaluation and validation techniques work in concert to ensure that machine-learning models in medical imaging are accurate not only in terms of training data but are also reliable for clinical deployment, where new and unseen cases have to be classified with a high level of confidence.

This chapter provided the essential clinical and technical background for ultrasound-based liver lesion classification. It discussed liver lesions in B-mode ultrasound, key image characteristics such as speckle noise and texture patterns, and challenges related to frame-based datasets including redundancy and data leakage. The chapter also introduced contrast enhancement techniques, with emphasis on CLAHE, and reviewed fundamental concepts of classical ML and model evaluation. These concepts support understanding of the processing pipeline and classification framework used in this thesis.

3 Chapter Three

Literature Review

This chapter evaluates the accomplishments obtained so far in the application of conventional machine learning techniques for the classification of liver lesions in ultrasound images. The application of computer techniques in the analysis of liver ultrasound has shown some promise in assisting radiologists by improving the consistency and objectivity of the diagnosis, especially in problematic cases involving low contrast and speckle noise. The performance of ML is typically highly dependent on the design of the entire processing pipeline that itself includes the preprocessing strategies, feature extraction methods, and the selection of classifiers. For this reason, we shall describe the development and evaluation of various ML workflows that have been utilized within the area of liver ultrasound imaging, with special emphasis on comparative analysis of classical algorithms under varying experimental conditions. By analyzing the various methodologies mentioned along with the results reported, this chapter aims to lay a solid foundation for appreciating current practices and eventually identifying the gaps that motivate the comparative framework of this study.

3.1 Machine Learning Applications in Liver Ultrasound Imaging

A definition generally offered to artificial intelligence (AI) in medical imaging is using computational methods to combine most aspects of human intelligence in interpreting and classifying images. Within liver ultrasound imaging, much research effort went into non-deep-learning ML techniques for the characterization of focal liver lesions, including but not limited to benign–malignant discrimination and multi-class lesion classification, using grayscale B-mode ultrasound images and handcrafted or radiomics-derived features (Vetter et al., 2023).

It motivates the application of ML to ultrasound of the liver in terms of reliability and consistency in diagnostic decision making that derives from such subtle and overlapping grayscale image patterns with respect to the weight of diagnostic entities. B-mode ultrasound is still considered first-line imaging modality in liver evaluation because of its accessibility and safety; however, it poses some inherent challenges in diagnosis (Mohammed et al., 2023). These include limited sensitivity for small lesions, especially those in cirrhotic livers, coarsely textured backgrounds, and speckle noise that may obscure lesion boundaries and hamper visual discriminability.

The entire literature shows a standard methodology composed of a standardized processing pipeline. The beginning of these pipelines is typically some identification or selection of a region of interest within the ultrasound image, followed by enhancement or transformation steps of the image, feature extraction, and classification with classical ML algorithms (Moghassemi et al., 2018). The outlined pipeline demonstrates an attempt to quantify highly complex visual information for computerized analysis.

In studies analyzing focal liver lesions, explicit lesion or region-of-interest delineation is commonly factored into the study design. Most times, the delineation of lesion boundaries is manual and done by experienced clinical personnel, so that feature extraction is defined within diagnostically relevant regions. Yet, these conditions are relaxed in instances where deliberate choices want to reduce the amount of methodological input by operators, who do then process in larger image regions or even entire cropped frames (Acharya et al., 2018). Such a difference clearly emphasizes other methodological varieties across studies in relation to a balance between automation and clinical supervision.

The specific postcard ultrasound unit image degradations motivate the preprocessing strategies used, most importantly, speckle noise, which appears as random grainy texture and degrades edge definition and contrast. To this end, histogram modification techniques for contrast enhancement usually are applied in conjunction with diffusion-based or adaptive filtering methods to stabilize intensity distributions and maintain texture consistency prior to feature extraction, thus improving the robustness against analysis downstream (Moghassemi et al., 2018).

Extracting features is at the heart of classical ML methods for liver ultrasound classification. Most studies focus on handcrafted texture descriptors based on the above rationale: that texture patterns encode vital information about the heterogeneity of the underlying tissue. Often, locally extracted features from gray-level co-occurrence matrices or local binary patterns- either alone or together with transform-domain representations such as wavelet-based features- are the most frequently reported, with the aim of enhancing the power to discriminate between lesion types (Moghassemi et al., 2018).

The classifiers employed after feature extraction include a variety of well-known algorithms such as linear models, margin-based classifiers, tree-based ensembles, probabilistic methods, and distance-based algorithms. SVM, random forests, logistic regression, and k-nearest neighbors consistently represent the most frequently employed classifiers across research studies on focal liver lesions classification and parenchymal liver disease evaluation (Vetter et al., 2023).

The variation of reported classification performance depends on the specific diagnostic task, feature design, and validation strategy. For instance, while binary benign-malignant classification, multi-class lesion differentiation, and broader parenchymal disease classification are using similar feature sets and classifiers for performance evaluations, their specificity among those evaluations is diverse (Shen et al., 2020; Vetter et al., 2023). This difference highlights the varying level of performance that ML achieves for different tasks in liver ultrasound applications.

Most utilizes evaluation methodologies for validation include cross-validation and hold-out tests, whereas independent external validation cohorts are quite limited. Systematic reviews have shown that the lack of external validation and heterogeneity of lesion types within individual datasets represent major barriers for clinical translation and generalizability (Vetter et al., 2023).

As a whole, many extensive studies present encouraging evidence for conventional ML approaches applied to B-mode liver ultrasound imaging, but vast differences in processing pipelines, the composition of datasets, and evaluation protocols reduce the feasibility of direct comparison across studies (Moghassemi et al., 2018). These points necessitate well-controlled comparative studies where different conventional ML algorithms can be evaluated under harmonized preprocessing and validation frameworks for a stronger basis for clinical adoption.

3.2 Comparative Studies of Traditional Machine Learning Algorithms in Ultrasound Medical Image Classification

In medical imaging applications, comparative studies play a fundamental role in critically examining and validating the usefulness of different computational methods. Due to the diversity of clinical applications and imaging modalities, it is not possible to find any ML algorithm that can be considered the best, with every application showing consistent superiority (Harikumar & Karthikamani, 2021). The performance of a given classifier is greatly sensitive to dataset characteristics, types of extracted features, and the clinical question investigated (Yap et al., 2009). Therefore, it is vital for a set of algorithms to be studied under the same experimental conditions, creating a balanced setting for comparisons (Wahdan et al., 2014) These studies emphasize task-specific algorithm suitability, as opposed to absolute superiority.

Most often, comparative studies in medical image classification are organized in a structured framework design experiment, aiming at fair and reproducible evaluations (Martínez-Más et al., 2019). Therefore, it begins from acquiring datasets and then proceeds with standardized preprocessing, feature extraction, and application of multiple classifiers (Yap et al., 2009).

Dataset characteristics vary widely across studies, from size to source to class balance characteristics (Martínez-Más et al., 2019). Stock pipelines are markedly set in order to remove artefact interference from the imaging and to standardize imaging characteristics. In ultrasound imaging, these types of pipelines mostly include speckle noise reduction, intensity normalization, and contrast enhancement (Suganya et al., 2022).

Feature representation is one of the crucial factors determining the comparative effectiveness of different traditional ML algorithms (N. Singh & Jindal, 2012). As these classical classifiers do not learn features directly with raw image data, the handcrafted features will efficiently evoke and contain the significant discriminative characteristics in the tissue samples. Various studies have shown that usually, the type and quality of features extracted often create a greater difference in classification performance than the classifier itself (Nieniewski & Chmielewski, 2020).

Some of the commonly used descriptors of texture include GLCM, Gabor filters, Local Binary Patterns (LBP), morphological, and first-order statistical features (Alivar et al., 2014; Wan et al., 2021). The performance of systems typically improves when multiple categories of features are combined (Alivar et al., 2014); but high-dimensional feature spaces can also lead to redundancy and overfitting. Therefore, feature selection and dimensionality reduction techniques are widely used to enhance robustness and generalization (Sathish Kumar et al., 2020).

Comparative ultrasound investigations typically consist of a core set of classical classifiers which include SVM variants, random forests, KNN, logistic regression, and Linear Discriminant Analysis (LDA). This base set is often expanded upon with boosting algorithms such as AdaBoost, XGBoost, LightGBM, and shallow neural networks such as multilayer perceptron or artificial neural networks depending on the scope of the study (Wan et al., 2021).

In no single study did any algorithm outperform all the other algorithms across the datasets and tasks (Harikumar & Karthikamani, 2021). SVM and ensemble methods like Random Forest seems to show pretty good performance, provided they are used with well-designed features (Mishra et al., 2021). KNN shows highly variable results depending on feature representation and data characteristics, whereas simple linear classifiers might perform competitively in well-structured feature space (Koseoglu et al., 2023).

Destructor M-studies find multiple metrics suitable enough for the classification performance rating (Harikumar & Karthikamani, 2021). The accuracy is put forth most of the time, but sometimes it might present a false view with imbalanced data (Jangra et al., 2022). Sensitivity and specificity are paramount due to their clinical relevance, especially in diagnostic contexts (Wan et al., 2021). The Area Under the Receiver Operating Characteristic (ROC) Curve is considered an important threshold-independent measure of performance (Martínez-Más et al., 2019). Precision, recall, and F1-score metrics are favored for reporting, especially in studies addressing the class imbalance (Wan et al., 2021).

Validation strategies differ in usage across research studies, but the most commonly used approaches are k-fold cross-validation and hold-out testing (Fleury & Marcomini, 2019). While internal validation generates useful estimates of performance, not having cohorts for independent external validation remains a major limitation. Models that work well in the internal validation may not generalize to data gained from other institutions or imaging systems (García Ocaña, 2019).

A very interesting feature of comparative ML literature is the variability and inconsistency of the findings reported. Different studies sometimes identify different algorithms as top performers even for similar clinical tasks. Such variability arises from differences in dataset composition, feature extraction methods, preprocessing pipelines, validation strategies, and hyperparameter tuning procedures. Therefore, reported performance rankings tend to be dataset-specific and hard to generalize (Putra et al., 2025).

Such methodological shortcomings are often encountered by many studies undertaken for comparative purposes and these include having inadequate sample size, one-centre data, and limited lesion variations along with class imbalance. Additional variability resulting from reliance on manual delineation of regions of interest makes it difficult to scale. Inconsistent preprocessing and feature extraction protocols limit reproducibility and comparison across studies (Chen et al., 2024).

Moreover, when comparative studies contribute immensely to understanding algorithm performance, the absence of standardized benchmarking frameworks and unified experimental pipelines designates limited interpretability and further clinical translation of the findings. Without doubt, there is still a need for systematic and controlled comparative evaluations which would adhere to shared protocols of preprocessing, feature extraction, validation, and metric-reporting.

Although previous comparative studies have evaluated classical ML algorithms in ultrasound medical image classification, direct comparison across studies remains limited due to differences in dataset characteristics, preprocessing pipelines, feature representations, and evaluation protocols. In contrast, the present thesis provides a controlled and task-oriented

comparative framework by evaluating multiple classical ML classifiers under identical preprocessing and validation conditions. Specifically, this study examines three clinically meaningful binary classification tasks (benign vs. normal, malignant vs. normal, and benign vs. malignant), assesses the impact of CLAHE contrast enhancement, and applies a rigorous evaluation strategy that includes stratified cross-validation, an independent test set, and a fully isolated holdout set. In addition, the study minimizes redundancy and potential data leakage by removing duplicate and near-duplicate frames prior to dataset splitting, thereby supporting a more reliable estimate of model generalization.

This chapter reviewed prior research on the use of classical ML methods in liver ultrasound imaging and broader ultrasound medical image classification. It highlighted common processing pipelines, feature extraction strategies, and classifier choices reported in the literature, as well as limitations such as dataset variability and inconsistent validation protocols. The chapter also emphasized the importance of controlled comparative evaluation. This review established the research gap addressed by the present thesis and motivated the comparative methodology adopted in the next chapter.

4 Chapter Four

Methodology

This study in computational ML was aimed at assessing the performance of classical ML techniques for classifying liver ultrasound images into normal, benign, and malignant. With this purpose, we constructed a heterogeneous liver ultrasound image dataset sourced from clinical image repositories as well as from a publicly available dataset. The methodological pipeline started with some level of preprocessing to ensure consistency and appropriateness for analysis, such as preprocessing with contrast enhancement and without contrast enhancement in order to assess their impact on classification performance. With that, a set of diverse classical ML algorithms will be trained and evaluated for this classification task. By systematically comparing the performances of the models using several evaluation metrics, this study aims at forming a robust consensus concerning the performance of ML-based approaches for ultrasound imaging liver lesion classification.

4.1 Dataset and Image Acquisition

This study was conducted on liver ultrasound images obtained from two sources: a locally collected clinical dataset and a publicly available dataset. The local dataset was derived from routine ultrasound examinations, during which video recordings were captured as part of the clinical assessment. One ultrasound video was acquired per patient, and each video was processed individually to extract frames and construct a slice-based image dataset. The local dataset included 1,552 frames extracted from 13 videos of benign liver lesions, 3,399 frames extracted from 14 videos of malignant liver lesions, and 1,105 frames extracted from 14 videos showcasing normal liver tissue.

Since one video corresponds to one unique patient, the number of local videos reflects the number of patients included in each diagnostic category (13 benign, 14 malignant, and 14 normal cases). Although the analysis was performed at the frame level, consecutive frames extracted from the same video are often highly correlated and may share similar visual patterns. To reduce redundancy and minimize the risk of data leakage, duplicate and highly similar frames were removed during the quality-control stage, and all frames originating from the same video were kept within the same dataset split during training and evaluation.

To increase dataset diversity and reduce single-center bias, a publicly available external dataset from Zenodo (DOI: 10.5281/zenodo.7272660) was incorporated. This dataset contributed 200 benign images, 435 malignant images, and 100 normal liver images. All images were reviewed to ensure consistency in imaging modality, anatomical region, and diagnostic labeling. After merging both sources, the initial combined dataset consisted of 6,791 ultrasound images across three diagnostic categories: benign, malignant, and normal liver tissue.

4.2 Image Preprocessing

All ultrasound images were resized to a standard spatial resolution of 128 × 128 pixels to ensure uniform input dimensions across classifiers. As an extra assurance for correct channel representation and to avoid conflicting with machine-learning pipelines, the images were converted to Red Green Blue (RGB).

Pre-processing of images by two strategies was evaluated. In the baseline preprocessing pipeline, the images were resized and directly flattened into one-dimensional feature vectors of 49,152 features (128 × 128 × 3). CLAHE was incorporated in a contrast-enhanced preprocessing pipeline to improve local contrast and visibility of subtle texture patterns related to the lesions before feature extraction.

CLAHE processing first required conversion of the image from RGB to LAB color space, wherein adaptive histogram equalization was performed only L (luminance) channel, applying a clip limit of 2.0 and a tile grid of 8 × 8. The enhanced L channel was then merged with the original A and B channels, and the image was converted back to RGB for flattening. CLAHE has been operated in OpenCV using the createCLAHE function.

4.3 Data Quality Control

The quality control pipeline was designed to prevent the possible leakage of data and to confirm data integrity; it thus imparted a level of detail sufficient to erase duplicates and near duplicates from the data. Exact duplicates were first determined through hash values calculated on the flattened array representation of each image. Next, perceptual hashing (pHash) with a hash size of 8 was implemented to detect perceptually similar images; thus, images showing a pHash

difference of 5 or less were flagged as similar. In this case, extra diligence was exercised to identify similar images across different diagnostic categories, as such cross-class similarities could be an indication of labeling errors or dataset contamination that might forcefully inflate classification performance. From each group of duplicate or similar image classes, only one representative image was included in the dataset.

The final curated dataset available for analysis, following quality control, consisted of 2,387 unique ultrasound pictures: 1,205 normal liver slices, 736 malignant slices, and 446 benign slices. This heavy reduction from the original 6,791 images to 2,387 highlights the critical need to rigorously identify and remove duplicates in datasets coming from video frames, where consecutive frames often appear quite similar. All duplicate detection and removal processes were done prior to any type of splitting of the dataset, thus allowing complete independence between training, testing, and holdout sets.

4.4 Dataset Partitioning

After data quality control, the cleaned dataset was partitioned using stratified sampling to maintain proportional class distribution across all subsets. The data were divided into a training set comprising 70% of the images, while the remaining 30% was equally split between a test set and a holdout set, each containing 15% of the total data. Stratification ensured that each diagnostic category was represented in the same proportion across training, test, and holdout partitions, preventing class imbalance issues during model evaluation. The holdout set was treated as a completely independent validation set, isolated from all model development, hyperparameter tuning, and model selection decisions to provide an unbiased estimate of final model performance.

4.5 Classification Models

Twelve classification algorithms, which represent various ML paradigms, were put to the test in the present study. Ensemble methods used were Random Forests, using 100 decision trees with bootstrap aggregation; Extra Trees, using 100 extremely randomized trees; Gradient Boosting, with 100 sequential boosting iterations; the last method being AdaBoost consisting of 100 adaptive boosting iterations. SVM were trialed with both RBF kernel and linear kernel options, with probability estimates on for ROC analysis. Linear models included Logistic regression with L2 regularization and a limit of 1000 iterations and Ridge classifier using L2-regularized least squares. Instance-based learning was by K-Nearest Neighbors with k=3 and k=5 neighbors. Gaussian Naive Bayes, assuming Gaussian feature distributions, and one Classification and Regression Tree (CART) Decision Tree were implemented. The models were implemented via scikit-learn, with the default hyperparameters used except for the ones stated above, and the random states were set to 42 for results' reproducibility.

For KNN, we evaluated two commonly used odd neighborhood sizes (k=3 and k=5). These values reduce the possibility of tie voting and allow comparison between a more local decision rule (k=3) and a slightly smoother, more stable decision boundary (k=5). This choice provides a simple sensitivity analysis of KNN performance with respect to neighborhood size, which is particularly relevant for ultrasound data characterized by speckle noise and overlapping texture patterns.

4.6 Feature Scaling and Model Training

For ameliorating model convergence with varying magnitude of features, StandardScaler normalization was performed on all of the features in such a way that each feature has a zero mean and unit variance. It is essential to mention that in every experimental iteration, scaling parameters were fitted only on training data before they were applied to validation, test, and holdout sets while preventing any data leakage. Fivefold stratified cross-validation was done on the training set to assess model stability and estimate generalization performance. In each fold, the training data were divided into four folds for model training and one fold for validation, where feature scaling was fitted separately on each training partition. The models were trained on the scaled training folds and evaluated on the corresponding validation fold, with performance metrics accumulated across all five folds for computing their means and standard deviations. After the cross-validation, each model was retrained on the complete training set, followed by sequential assessment on the test set for model selection and performance measure and on the holdout set for unbiased performance estimation. Such three-tier evaluation procedures yielded robust estimates of the generalization ability of models, while at the same time keeping any possible overfitting at bay.

4.7 Model Validation Strategy

Model validation was performed using a three-stage evaluation framework. First, stratified 5-fold cross-validation was applied on the training set to estimate performance stability. Second, model performance was evaluated on an independent test set that was not used during training. Finally, a fully isolated holdout set was used to report unbiased generalization performance on unseen data. Ground-truth labels were based on clinical diagnosis supported by radiology reports and confirmed by expert physicians. To prevent data leakage, feature scaling parameters were fitted only on the training data and then applied to validation, test, and holdout sets.

4.8 Performance Metrics

In a comprehensive way, the evaluation of classification capabilities was performed on several metrics suited for binary classification problems. Accuracy measures the overall proportion of

correct predictions, while sensitivity (recall) indicates the true positive rate and the ability of the model to identify positive cases correctly. Specificity, on the other hand, indicates how well the model identifies negative cases by measuring the true negative rate. Positive predictive value (PPV), (precision) states how many of the positive predictions were correctly predicted, while negative predictive value (NBV) states how many of the negative predictions were correctly predicted. The F1 score, the harmonic mean of precision and recall, allowed a balanced measure that takes into both false positives and false negatives. For probability-estimating models, the AUC-ROC was calculated, measuring the discriminative ability at all classification thresholds.

Thus, all these metrics were in turn calculated for cross-validation folds (mean ± standard deviation) and separately across test and hold-out sets, thereby giving a whole spectrum of the model's performance and generalization.

4.9 Binary Classification Framework

The dataset was classified into three classes: benign, malignant, and normal liver ultrasound images; thus, each of the three different binary classifications was performed to test the diagnostic capability in detail. This included classifying benign lesions from normal liver tissue in the first task, differentiating malignant lesions from normal tissue in the next, and drawing a line between benign and malignant lesions in the third task. Each binary classification problem subsetted the dataset so that only images from the two pertinent diagnostic categories were included, and the class labels were remapped to binary values of 0 and 1. All preprocessing steps, quality control procedures, data partitioning, scaling features, and model training protocols were performed independently on each of the three binary classification tasks. It let the experimental design remain intact and let the targeted assessment of the classifier performance on clinically relevant diagnostic distinctions.

4.10 Statistical Analysis and Visualization

On their part, systematic aggregation and comparison happened among all performance metrics with respect to all classification models in the preprocessing conditions and binary task. ROC curves were created in every model to show possibly trade-offs between sensitivity and specificity over various classification levels, with AUCs computed for comparison purposes.
Several confusion matrices were constructed for exhibiting true positives, true negatives, false positives, and false negatives across classifiers. These included performance metrics across models in holdout set using bar charts, model performance patterns through all binary classification tasks using heatmap, and direct comparison of CLAHE versus censused preprocessing impact on classification. All analysis on computation was carried out by using Python in conjunction with NumPy for operations using numbers, pandas for data manipulation, scikit-learn for ML implementations, OpenCV for image processing, and field specific

Matplotlib and Seaborn for visualization. The complete analysis pipeline, from loading the data to finally visualizing the analysis, was designed to be completely reproducible with fixed random seeds.

This chapter described the study methodology, including dataset acquisition from local clinical ultrasound videos and a public dataset, image preprocessing with and without CLAHE, and the quality control process used to remove duplicate and highly similar frames. It also presented the dataset partitioning strategy and detailed the classical ML classifiers evaluated in this work. Finally, the chapter explained the training workflow, feature scaling approach, and evaluation framework using cross-validation, independent testing, and holdout validation across three binary classification tasks.

5 Chapter Five

Results

This chapter presents the experimental results of the proposed ML framework for liver ultrasound image classification. Results are organized according to study objectives based on the evaluation of model performance across three binary diagnostic tasks: benign versus normal, malignant versus normal, and benign versus malignant classification. Performance evaluation will therefore be done by way of stratified cross-validation, independent testing, and holdout evaluation, with and without CLAHE preprocessing. The reported metrics will capture the discrimination ability and the robustness and generalization behavior across classifiers and preprocessing strategies.

5.1 Cross-Validation Performance for Benign versus Normal Liver Classification

This analysis aimed to investigate the ability of various ML classifiers to reliably distinguish benign liver lesions from normal liver tissue using ultrasound-derived features. Such a clinically relevant task requires reliable differentiation of benign lesions from healthy parenchyma. The five-fold stratified cross-validation was carried out to evaluate classification stability and robustness with respect to standard preprocessing and CLAHE-based contrast enhancement.

Table 5.1 presents quantitative results in the form of five-fold stratified cross-validation comparison results for benign-normal liver classification under both standard preprocessing and CLAHE-based contrast enhancement. All classifiers exhibited reproducible and consistent performance across the cross-validation; these ensemble-based models scored maximums in discrimination with least variability among different classifiers.

5.1.1 Performance without CLAHE

In the absence of contrast enhancement, the Extra Trees classifier showed the highest cross-validated accuracy (0.9137 ± 0.0359), highest cross-validated F1 score (0.9099 ± 0.0371), and highest AUC (0.9769 ± 0.0166) indicating strong separation between normal and benign liver tissue with small inter-fold variance. Random Forest followed closely with an accuracy of 0.8927 ± 0.0374 and AUC of 0.9742 ± 0.0141, with balanced sensitivity (0.8722 ± 0.0385) and specificity (0.9115 ± 0.0507).

Competitively accurate; an RBF kernel SVM scored 0.9068 ± 0.0416 in accuracy and 0.9334 ± 0.0467 in specificity, thus demonstrating the merit of nonlinear decision boundaries in ultrasound texture discrimination. On the contrary, the linear SVM performed worse in both accuracy (0.8787 ± 0.0337) and AUC (0.9505 ± 0.0225), indicating relatively diminished expressiveness.

It was evident from the instance-based classifiers that they had a real imbalance between sensitivity and specificity. KNN (k = 5), for instance, had a very high sensitivity (0.9657 ± 0.0248) and NPV (0.9627 ± 0.0278), coupled with a low specificity (0.8315 ± 0.0854) and higher variance. Gaussian Naive Bayes, on the other hand, was close to perfect in terms of sensitivity (0.9852 ± 0.0196) and NPV (0.9806 ± 0.0244) but had a specificity of low values (0.6679 ± 0.0790) and the lowest AUC (0.8266 ± 0.0409), which indicate a big class bias and minimal robustness.

The algorithms classified under linear classifiers (Logistic Regression and Ridge) reached a moderate performance: logistic regression being stable at an accuracy of 0.8858 ± 0.0298 and ridge regression being stable at an accuracy of 0.8672 ± 0.0332. The single Decision Tree reached an accuracy of 0.8834 ± 0.0355 and AUC of 0.8827 ± 0.0345, with more consistent underperformance than ensemble tree-based counterparts.

5.1.2 Effect of CLAHE on Cross-Validation Performance

Large improvements performed on the Random Forest raised its precision to 0.9228 ± 0.0151, and its AUC to 0.9830 ± 0.0051, while increasing precision up to 0.9469 ± 0.0401, with the result that Gradient Boosting also improved contrast, increasing accuracy to 0.9134 ± 0.0182, with an AUC of 0.9823 ± 0.0022.

Under the CLAHE conditions, SVM achieved moderate outcomes. RBF SVM recorded an accuracy of 0.9040 ± 0.0138 and an AUC of 0.9692 ± 0.0122, while the Linear SVM showed improved accuracy reaching 0.8908 ± 0.0224 with an AUC of 0.9692 ± 0.0114. The specificity of KNN classifiers made a remarkable improvement under the CLAHE preprocessing condition, with KNN (k = 3) gaining an accuracy value of 0.9021 ± 0.0172 and an F1 score of 0.9129 ± 0.0127.

Yet though Naive Bayes did show a minor advancement in accuracy (0.8569 ± 0.0190) and F1 score (0.8829 ± 0.0143), specificity stood low (0.7014 ± 0.0432) to make evident that some restrictions were still present even though contrast enhancement was done. Linear models and the Ridge Classifier showed very little consistency in results, whereas the Decision Tree improved its sensitivity (0.9276 ± 0.0571) but still couldn't match those that employ the ensemble approaches in performance comparison.

Ensemble-based classifiers, particularly Extra Trees and Random Forest, achieved consistently higher cross-validated accuracy and AUC across both preprocessing strategies, with CLAHE further reducing variability while enhancing performance. Contrast enhancement improved mostly the sensitivity and general discriminative stability, hence throwing light on its role in textural-feature-learning facilitation pertaining to benign-normal liver differentiation in ultrasonographic images.

Table 5.1a: Cross-Validation Performance of Benign-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.8927 ± 0.0374	0.8722 ± 0.0385	0.9115 ± 0.0507	0.9005 ± 0.0553	0.8881 ± 0.0332	0.8853 ± 0.0392	0.9742 ± 0.0141
Logistic Regression	0.8858 ± 0.0298	0.8916 ± 0.0461	0.8805 ± 0.0437	0.8717 ± 0.0417	0.9023 ± 0.0369	0.8806 ± 0.0324	0.9623 ± 0.0169
SVM (RBF)	0.9068 ± 0.0416	0.8773 ± 0.0505	0.9334 ± 0.0467	0.9237 ± 0.0526	0.8948 ± 0.0422	0.8992 ± 0.0451	0.9577 ± 0.0191
SVM (Linear)	0.8787 ± 0.0337	0.8865 ± 0.0344	0.8716 ± 0.0516	0.8632 ± 0.0495	0.8959 ± 0.0266	0.8739 ± 0.0345	0.9505 ± 0.0225
Gradient Boosting	0.8997 ± 0.0338	0.8867 ± 0.0396	0.9113 ± 0.0489	0.9020 ± 0.0478	0.9003 ± 0.0325	0.8934 ± 0.0346	0.9718 ± 0.0116
AdaBoost	0.8857 ± 0.0310	0.8866 ± 0.0406	0.8846 ± 0.0621	0.8784 ± 0.0579	0.8984 ± 0.0292	0.8806 ± 0.0302	0.9627 ± 0.0109
Extra Trees	0.9137 ± 0.0359	0.9166 ± 0.0450	0.9114 ± 0.0581	0.9061 ± 0.0563	0.9254 ± 0.0390	0.9099 ± 0.0371	0.9769 ± 0.0166
KNN (k=5)	0.8950 ± 0.0533	0.9657 ± 0.0248	0.8315 ± 0.0854	0.8426 ± 0.0736	0.9627 ± 0.0278	0.8988 ± 0.0492	0.9531 ± 0.0294
KNN (k=3)	0.8763 ± 0.0380	0.9310 ± 0.0102	0.8271 ± 0.0670	0.8321 ± 0.0539	0.9297 ± 0.0127	0.8779 ± 0.0332	0.9352 ± 0.0327
Naive Bayes	0.8180 ± 0.0431	0.9852 ± 0.0196	0.6679 ± 0.0790	0.7300 ± 0.0461	0.9806 ± 0.0244	0.8377 ± 0.0323	0.8266 ± 0.0409

Table 5.1b: Cross-Validation Performance of Benign-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
NO CLAHE							
Decision Tree	0.8834 ± 0.0355	0.8718 ± 0.0365	0.8935 ± 0.0605	0.8845 ± 0.0587	0.8866 ± 0.0284	0.8767 ± 0.0348	0.8827 ± 0.0345
Ridge	0.8672 ± 0.0332	0.8720 ± 0.0395	0.8629 ± 0.0379	0.8515 ± 0.0400	0.8830 ± 0.0342	0.8613 ± 0.0353	0.9149 ± 0.0250
CLAHE							
Random Forest	0.9228 ± 0.0151	0.9138 ± 0.0567	0.9339 ± 0.0525	0.9469 ± 0.0401	0.9058 ± 0.0553	0.9277 ± 0.0164	0.9830 ± 0.0051
Logistic Regression	0.8983 ± 0.0190	0.9069 ± 0.0507	0.8884 ± 0.0569	0.9102 ± 0.0392	0.8931 ± 0.0523	0.9066 ± 0.0182	0.9739 ± 0.0063
SVM (RBF)	0.9040 ± 0.0138	0.9034 ± 0.0641	0.9050 ± 0.0639	0.9244 ± 0.0440	0.8943 ± 0.0623	0.9107 ± 0.0160	0.9692 ± 0.0122
SVM (Linear)	0.8908 ± 0.0224	0.9103 ± 0.0428	0.8676 ± 0.0503	0.8940 ± 0.0338	0.8928 ± 0.0465	0.9009 ± 0.0203	0.9692 ± 0.0114
Gradient Boosting	0.9134 ± 0.0182	0.9172 ± 0.0560	0.9090 ± 0.0493	0.9266 ± 0.0356	0.9073 ± 0.0598	0.9199 ± 0.0184	0.9823 ± 0.0022
AdaBoost	0.8945 ± 0.0264	0.8862 ± 0.0641	0.9047 ± 0.0383	0.9197 ± 0.0270	0.8748 ± 0.0649	0.9008 ± 0.0279	0.9763 ± 0.0072
Extra Trees	0.9266 ± 0.0070	0.9345 ± 0.0399	0.9173 ± 0.0448	0.9337 ± 0.0324	0.9245 ± 0.0430	0.9327 ± 0.0074	0.9841 ± 0.0064
KNN (k=5)	0.8983 ± 0.0108	0.9345 ± 0.0428	0.8551 ± 0.0459	0.8876 ± 0.0277	0.9201 ± 0.0480	0.9092 ± 0.0110	0.9690 ± 0.0030
KNN (k=3)	0.9021 ± 0.0172	0.9345 ± 0.0201	0.8634 ± 0.0598	0.8941 ± 0.0404	0.9179 ± 0.0198	0.9129 ± 0.0127	0.9573 ± 0.0128
Naive Bayes	0.8569 ± 0.0190	0.9862 ± 0.0201	0.7014 ± 0.0432	0.7997 ± 0.0228	0.9784 ± 0.0302	0.8829 ± 0.0143	0.8438 ± 0.0203
Decision Tree	0.8965 ± 0.0320	0.9276 ± 0.0571	0.8592 ± 0.0571	0.8902 ± 0.0411	0.9135 ± 0.0632	0.9069 ± 0.0299	0.8934 ± 0.0316
Ridge	0.8927 ± 0.0161	0.9138 ± 0.0308	0.8676 ± 0.0537	0.8947 ± 0.0365	0.8954 ± 0.0306	0.9030 ± 0.0132	0.9685 ± 0.0131

5.2 Cross-Validation Performance for Malignant versus Normal Liver Classification

This analysis aims at measuring the discriminative performances of the classifiers assessed for distinguishing malignancy in liver lesions and normal liver tissue. Given the pronounced pathological differences between malignant lesions and healthy liver parenchyma, it is expected that this task shall exhibit greater separability versus benign-normal classification. The cross-validation results are analyzed to quantify model accuracy, stability, and sensitivity to contrast enhancement.

To systematically assess these hypotheses, Table 5.2 presents the five-fold stratified cross-validation performance of all evaluated ML models for the malignant–normal liver classification task, under both standard preprocessing and CLAHE-based contrast enhancement. This task yielded higher performance levels than benign–normal classification, with most models

achieving accuracies above 0.95 and AUC values approaching unity, indicating strong separability between malignant lesions and normal liver tissue.

5.2.1 Performance without CLAHE

Overall, neither the ensemble classifiers nor the linear ones were of any poor or unstable performance under standard preprocessing. However, the Extra Trees classifier was able to achieve the highest performance, with cross-validated accuracy at 0.9813 ± 0.0133, F1 score at 0.9682 ± 0.0225, and an AUC at 0.9988 ± 0.0009 - indicating close discrimination which does not vary among folds nearly at all. Logistic regression also did well, with an accuracy of 0.9770 ± 0.0070 and an F1 score of 0.9606 ± 0.0123, in addition to a very high AUC of 0.9930 ± 0.0103, indicating that the malignant-normal separation is largely linearly separable in the feature space extracted.

Both Random Forest and Gradient Boosting showed consistency with accuracies 0.9669 ± 0.0098 and 0.9741 ± 0.0141, respectively. Their AUC values surpassed 0.996. SVM performed relatively well, with the linear SVM achieving an accuracy of 0.9741 ± 0.0117 and an AUC of 0.9918 ± 0.0114, marginally outdoing the RBF SVM concerning sensitivity (0.9704 ± 0.0286 vs. 0.9212 ± 0.0285).

Since they are instance-based classifiers, they are sensitive but exhibit a lot of variance. KNN (k = 5) achieved a sensitivity of 0.9656 ± 0.0330 and an accuracy of 0.9669 ± 0.0168, while with a slight decrease in performance, KNN (k = 3) yielded an accuracy of 0.9568 ± 0.0137. Gaussian Naive Bayes was observed to differ largely in accuracy (0.8763 ± 0.0286) and AUC (0.8882 ± 0.0292) owing to specificity being lower at 0.8598 ± 0.0346; thus, it will show limited robustness for this classification task.

The single Tree could perform mediocrely with monotony of accuracy equal to 0.9439 ± 0.0154 and somewhat AUC equal to 0.9300 ± 0.0155, with lesser amalgamated tree methods. The latter outcome came up from the performance of Ridge Classifier as per the accuracy of 0.9612 ± 0.0148 along with an AUC of 0.9871 ± 0.0174, showing that the model already performed for establishing stable and slightly lower discrimination power than with use of an ensemble approach.

5.2.2 Effect of CLAHE on Cross-Validation Performance

CLAHE application enabled more performance improvements for more classifiers, particularly ensemble-based ones. The maximum overall performance was under CLAHE preprocessing for Extra Trees, where accuracy increased to 0.9848 ± 0.0064, the F1 score rose to 0.9791 ± 0.0091,

and AUC was around 0.9993 ± 0.0003, coupled with reduced variance, indicating improving stability and feature consistency.

Although it has been adopted from CLAHE, Gradient Boosting received a merit establishing this with an accuracy of 0.9760 ± 0.0084, an AUC of 0.9982 ± 0.0016, and an increased precision of acceptance at 0.9758 ± 0.0176. It influenced Random Forest in the same way, improving sensitivity but maintaining a high value of AUC at 0.9974 ± 0.0020; the increased sensitivity was set at 0.9621 ± 0.0201.

KNN classifiers exhibited marked sensitivity gains under CLAHE, with KNN (k = 5) achieving a sensitivity of 0.9897 ± 0.0138 and KNN (k = 3) reaching 0.9931 ± 0.0138, accompanied by high NBV exceeding 0.99. There was slightly reduced specificity associated with these improvements when compared to ensemble-based approaches.

The SVM exhibited only marginal improvements under CLAHE's procedures with both linear and RBF kernels attaining accuracies hovering near 0.963 and AUC values beyond 0.994, indicating little added benefit from contrast enhancement. Logistic Regression showed just a slight decrement in accuracy (0.9621 ± 0.0098) in comparison with the non-CLAHE condition, implying successful linear separability was already well captured without contrast enhancement.

Naive Bayes remained the worst performer under CLAHE, with accuracy reducing to 0.8369 ± 0.0188 and AUC to 0.8523 ± 0.0119, while still retaining relatively high sensitivity (0.9103 ± 0.0229). The Decision Tree and Ridge Classifier have shown a slight gain in performance, though significantly underperformed when compared to ensemble methods.

Performing malignant–normal liver classification using either of these preprocess methods led to very high and consistent cross-validation performance. Of all the ensemble-based classifiers, Extra Trees produced the most accurate, stable, and discriminative results. CLAHE preprocessing increased the sensitivity and reduced the inter-fold variability for a number of models; however, it had less dramatic effects than in the benign–normal situation, indicating stronger inherent texture and intensity differences between malignant lesions and normal liver tissue.

Table 5.2a: Cross-Validation Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.9669 ± 0.0098	0.9507 ± 0.0154	0.9736 ± 0.0152	0.9382 ± 0.0353	0.9796 ± 0.0061	0.9439 ± 0.0160	0.9969 ± 0.0023
Logistic Regression	0.9770 ± 0.0070	0.9654 ± 0.0295	0.9817 ± 0.0119	0.9573 ± 0.0272	0.9860 ± 0.0120	0.9606 ± 0.0123	0.9930 ± 0.0103

Table 5.2b: Cross-Validation Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
SVM (RBF)	0.9683 ± 0.0086	0.9212 ± 0.0285	0.9878 ± 0.0100	0.9698 ± 0.0237	0.9683 ± 0.0110	0.9444 ± 0.0151	0.9945 ± 0.0023
SVM (Linear)	0.9741 ± 0.0117	0.9704 ± 0.0286	0.9757 ± 0.0121	0.9431 ± 0.0279	0.9878 ± 0.0117	0.9561 ± 0.0200	0.9918 ± 0.0114
Gradient Boosting	0.9741 ± 0.0141	0.9557 ± 0.0182	0.9818 ± 0.0225	0.9584 ± 0.0494	0.9818 ± 0.0073	0.9561 ± 0.0227	0.9976 ± 0.0019
AdaBoost	0.9727 ± 0.0184	0.9705 ± 0.0286	0.9737 ± 0.0208	0.9397 ± 0.0465	0.9877 ± 0.0118	0.9542 ± 0.0305	0.9970 ± 0.0026
Extra Trees	0.9813 ± 0.0133	0.9706 ± 0.0284	0.9858 ± 0.0138	0.9667 ± 0.0323	0.9879 ± 0.0116	0.9682 ± 0.0225	0.9988 ± 0.0009
KNN (k=5)	0.9669 ± 0.0168	0.9656 ± 0.0330	0.9675 ± 0.0134	0.9248 ± 0.0305	0.9856 ± 0.0138	0.9445 ± 0.0283	0.9936 ± 0.0008
KNN (k=3)	0.9568 ± 0.0137	0.9509 ± 0.0217	0.9594 ± 0.0143	0.9067 ± 0.0312	0.9793 ± 0.0092	0.9280 ± 0.0224	0.9916 ± 0.0017
Naive Bayes	0.8763 ± 0.0286	0.9165 ± 0.0450	0.8598 ± 0.0346	0.7325 ± 0.0568	0.9617 ± 0.0202	0.8129 ± 0.0417	0.8882 ± 0.0292
Decision Tree	0.9439 ± 0.0154	0.8966 ± 0.0285	0.9635 ± 0.0218	0.9126 ± 0.0500	0.9578 ± 0.0109	0.9034 ± 0.0257	0.9300 ± 0.0155
Ridge	0.9612 ± 0.0148	0.9411 ± 0.0246	0.9695 ± 0.0170	0.9285 ± 0.0377	0.9755 ± 0.0105	0.9343 ± 0.0243	0.9871 ± 0.0174
CLAHE							
Random Forest	0.9722 ± 0.0130	0.9621 ± 0.0201	0.9780 ± 0.0172	0.9629 ± 0.0287	0.9782 ± 0.0112	0.9622 ± 0.0173	0.9974 ± 0.0020
Logistic Regression	0.9621 ± 0.0098	0.9690 ± 0.0201	0.9580 ± 0.0184	0.9316 ± 0.0279	0.9819 ± 0.0111	0.9494 ± 0.0126	0.9957 ± 0.0029
SVM (RBF)	0.9633 ± 0.0046	0.9310 ± 0.0327	0.9820 ± 0.0133	0.9690 ± 0.0217	0.9615 ± 0.0173	0.9489 ± 0.0078	0.9971 ± 0.0016
SVM (Linear)	0.9633 ± 0.0093	0.9690 ± 0.0316	0.9600 ± 0.0179	0.9349 ± 0.0276	0.9822 ± 0.0175	0.9509 ± 0.0125	0.9949 ± 0.0044
Gradient Boosting	0.9760 ± 0.0084	0.9586 ± 0.0176	0.9860 ± 0.0102	0.9758 ± 0.0176	0.9764 ± 0.0096	0.9670 ± 0.0115	0.9982 ± 0.0016
AdaBoost	0.9658 ± 0.0190	0.9586 ± 0.0234	0.9700 ± 0.0179	0.9491 ± 0.0300	0.9759 ± 0.0137	0.9538 ± 0.0256	0.9960 ± 0.0025
Extra Trees	0.9848 ± 0.0064	0.9724 ± 0.0207	0.9920 ± 0.0075	0.9863 ± 0.0127	0.9844 ± 0.0112	0.9791 ± 0.0091	0.9993 ± 0.0003
KNN (k=5)	0.9760 ± 0.0084	0.9897 ± 0.0138	0.9680 ± 0.0147	0.9479 ± 0.0231	0.9940 ± 0.0079	0.9681 ± 0.0109	0.9967 ± 0.0026
KNN (k=3)	0.9772 ± 0.0110	0.9931 ± 0.0138	0.9680 ± 0.0133	0.9477 ± 0.0210	0.9959 ± 0.0082	0.9698 ± 0.0146	0.9941 ± 0.0046
Naive Bayes	0.8369 ± 0.0188	0.9103 ± 0.0229	0.7943 ± 0.0403	0.7219 ± 0.0368	0.9393 ± 0.0124	0.8042 ± 0.0157	0.8523 ± 0.0119
Decision Tree	0.9608 ± 0.0177	0.9552 ± 0.0176	0.9640 ± 0.0250	0.9406 ± 0.0398	0.9738 ± 0.0101	0.9474 ± 0.0228	0.9596 ± 0.0159
Ridge	0.9419 ± 0.0145	0.9483 ± 0.0289	0.9381 ± 0.0073	0.8985 ± 0.0133	0.9693 ± 0.0165	0.9227 ± 0.0202	0.9836 ± 0.0145

5.3 Cross-Validation Performance for Benign versus Malignant Liver Classification

Benign-malignant liver classification represents the toughest diagnostic task among others in view of great visual and textural overlap between lesion types in ultrasound images. This further complicates classification, resulting in lesser accuracies and marked sensitivity-specificity trade-offs across various models. The cross-validation analysis was carried out in that context to investigate the model's discrimination behavior and any possible effects of CLAHE-based contrast enhancement, as summarized in Table 3, which summarizes the five-fold stratified cross-validation performance of the evaluated ML models for the benign-malignant liver classification task under standard preprocessing and CLAHE-based contrast enhancement. By far, the task was proved to be more difficult than benign-normal and malignant-normal classification, with lower overall accuracies and pronounced sensitivity-specificity trade-offs across models, thus attesting to great visual and textural similarity between benign and malignant liver lesions in ultrasound images.

5.3.1 Performance without CLAHE

Under the usual preprocessing, ensemble classifiers again showed the greatest performance, although less pronounced than the other classification tasks. Indeed, the Extra Trees classifier exhibited the best value for cross-validated accuracy (0.8719 ± 0.0319), F1 score (0.9084 ± 0.0230), and AUC (0.9456 ± 0.0268), which suggests the most reliable discrimination between benign and malignant lesions. Random Forest came second by getting an accuracy of 0.8538 ± 0.0405 and an F1 score of 0.8960 ± 0.0294 but with much lower specificity (0.7078 ± 0.0727) which indicated increased confusion in the classes.

SVM were seen to be very sensitive with decreased specificity. Among the models, the RBF SVM, with sensitivity of 0.9412 ± 0.0390, recorded the highest sensitivity, while specificity dropped to 0.6591 ± 0.0597, culminating in an accuracy of 0.8524 ± 0.0373 with AUC 0.9126 ± 0.0376. The linear SVM showed a similar trend with a sensitivity of 0.9026 ± 0.0434 but a specificity of 0.7251 ± 0.1065.

A fair balancing act process was rather moderately thorough for both Gradient Boosting and AdaBoost, which yielded accuracies of 0.8426 ± 0.0324 and 0.8496 ± 0.0359. Their respective AUC was 0.92. KNN classifiers exhibited good sensitivity (0.9291 ± 0.0472 for k = 5), but specificity was relatively low (0.6772 ± 0.0592), thus displaying a bias toward predicting more malignant cases.

With an average accuracy of 0.7994±0.0293 and an area under a curve of 0.7179±0.0427, Gaussian Naive Bayes was the most poorly performing model overall; low specificity (0.4912±0.0697) was largely responsible for this, while high sensitivity (0.9411±0.0241) was

noted. The individual Decision Tree and Ridge Classifier also lagged in performance behind ensemble methods, scoring an accuracy below 0.82 and AUC below 0.86.

5.3.2 Effect of CLAHE on Cross-Validation Performance

But applying CLAHE has small and consistent performance improvements among classifiers, usually on ensemble models and instance-based ones. The best results under CLAHE were caused due to Extra Trees: the accuracy increased to 0.8761±0.0247, F1 score to 0.9105±0.0160, and AUC to 0.9463±0.0122, with its decreased inter-fold variability.

Random Forest gained accuracy as high as 0.8680 ± 0.0134 and an F1 score of 0.9050 ± 0.0076, both demonstrating stability improvements under CLAHE. Gradient Boosting benefited from CLAHE, acquiring an accuracy of 0.8586 ± 0.0255 and an AUC of 0.9322 ± 0.0155.

With CLAHE techniques, KNN classifiers showed a remarkable improvement in both accuracy and F1 score. KNN (k = 5) produced an accuracy of 0.8720 ± 0.0111 and an F1 score of 0.9081 ± 0.0070, while KNN (k = 3) appeared to score comparably with an accuracy of 0.8720 ± 0.0250 and an F1 score of 0.9074 ± 0.0160. Specificity improvement from the non-CLAHE condition was the primary driver of these improvements.

SVM are least sensitive to changes under CLAHE, while RBF SVM maintained a high sensitivity (0.9422 ± 0.0229) coupled with low specificity (0.6724 ± 0.0504), giving just a small edge to accuracy. Logistic Regression and Ridge classifiers showed slight stability improvements yet were again outperformed by ensemble-based models.

After contrast enhancement, among all classes of classifiers, Naive Bayes has remained the weakest performer recording an accuracy of mere 0.8005 ± 0.0100 and an AUC of 0.7330 ± 0.0082, indeed indicating its inherent weaknesses in modeling overlapping benign-malignant feature distributions.

Benign-malign liver classification is the most difficult diagnostic classification task in the three scenarios. Keep seeing lower and more varying performance from benign-normal and malignant-normal classification. Ensemble-based classifiers, specifically Extra Trees and Random Forest, displayed the most promising and evenly distributed performance across both preprocessing strategies. Moderate increases for accuracy, F1 score, and stability were shown with CLAHE pre-processing while ensemble models and KNN benefited far more compared to other classifiers. None of these alterations fundamentally addressed the inherent sensitivity-specificity imbalance in this task, though. These results highlight the considerable similarity of the ultrasound appearance characteristics of the benign and malignant liver lesions and further

very strongly suggest that for this particular classification scenario, even more discriminative feature representations or more advanced modeling approaches are needed.

Table 5.3a: Cross-Validation Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.8538 ± 0.0405	0.9209 ± 0.0400	0.7078 ± 0.0727	0.8732 ± 0.0290	0.8087 ± 0.0840	0.8960 ± 0.0294	0.9403 ± 0.0274
Logistic Regression	0.8482 ± 0.0647	0.9026 ± 0.0550	0.7296 ± 0.0996	0.8792 ± 0.0440	0.7781 ± 0.1106	0.8905 ± 0.0476	0.9106 ± 0.0428
SVM (RBF)	0.8524 ± 0.0373	0.9412 ± 0.0390	0.6591 ± 0.0597	0.8575 ± 0.0240	0.8437 ± 0.0915	0.8971 ± 0.0269	0.9126 ± 0.0376
SVM (Linear)	0.8468 ± 0.0513	0.9026 ± 0.0434	0.7251 ± 0.1065	0.8787 ± 0.0424	0.7757 ± 0.0831	0.8899 ± 0.0370	0.9040 ± 0.0443
Gradient Boosting	0.8426 ± 0.0324	0.9108 ± 0.0474	0.6947 ± 0.0494	0.8668 ± 0.0186	0.7913 ± 0.0866	0.8876 ± 0.0254	0.9280 ± 0.0349
AdaBoost	0.8496 ± 0.0359	0.9006 ± 0.0398	0.7386 ± 0.0660	0.8829 ± 0.0264	0.7776 ± 0.0698	0.8912 ± 0.0269	0.9150 ± 0.0401
Extra Trees	0.8719 ± 0.0319	0.9269 ± 0.0314	0.7520 ± 0.0642	0.8911 ± 0.0251	0.8282 ± 0.0632	0.9084 ± 0.0230	0.9456 ± 0.0268
KNN (k=5)	0.8496 ± 0.0203	0.9291 ± 0.0472	0.6772 ± 0.0592	0.8632 ± 0.0189	0.8293 ± 0.0896	0.8939 ± 0.0170	0.9031 ± 0.0322
KNN (k=3)	0.8510 ± 0.0270	0.9209 ± 0.0406	0.6991 ± 0.0330	0.8696 ± 0.0127	0.8104 ± 0.0810	0.8940 ± 0.0210	0.8889 ± 0.0395
Naive Bayes	0.7994 ± 0.0293	0.9411 ± 0.0241	0.4912 ± 0.0697	0.8015 ± 0.0238	0.7955 ± 0.0763	0.8655 ± 0.0192	0.7179 ± 0.0427
Decision Tree	0.8120 ± 0.0411	0.8457 ± 0.0463	0.7386 ± 0.0446	0.8755 ± 0.0215	0.6916 ± 0.0718	0.8600 ± 0.0327	0.7921 ± 0.0402
Ridge	0.8189 ± 0.0531	0.8782 ± 0.0481	0.6896 ± 0.0820	0.8606 ± 0.0356	0.7258 ± 0.0960	0.8690 ± 0.0391	0.8583 ± 0.0355
CLAHE							
Random Forest	0.8680 ± 0.0134	0.9301 ± 0.0167	0.7390 ± 0.0718	0.8823 ± 0.0275	0.8381 ± 0.0247	0.9050 ± 0.0076	0.9384 ± 0.0108
Logistic Regression	0.8490 ± 0.0272	0.8862 ± 0.0205	0.7714 ± 0.0552	0.8902 ± 0.0239	0.7652 ± 0.0400	0.8881 ± 0.0197	0.9194 ± 0.0172
SVM (RBF)	0.8545 ± 0.0216	0.9422 ± 0.0229	0.6724 ± 0.0504	0.8571 ± 0.0191	0.8509 ± 0.0566	0.8974 ± 0.0150	0.9229 ± 0.0094
SVM (Linear)	0.8504 ± 0.0313	0.8923 ± 0.0128	0.7631 ± 0.0712	0.8876 ± 0.0305	0.7712 ± 0.0393	0.8898 ± 0.0215	0.9106 ± 0.0260
Gradient Boosting	0.8586 ± 0.0255	0.9321 ± 0.0240	0.7057 ± 0.0823	0.8695 ± 0.0307	0.8355 ± 0.0441	0.8992 ± 0.0168	0.9322 ± 0.0155
AdaBoost	0.8396 ± 0.0181	0.9062 ± 0.0205	0.7014 ± 0.0375	0.8634 ± 0.0149	0.7837 ± 0.0382	0.8841 ± 0.0133	0.9180 ± 0.0177
Extra Trees	0.8761 ± 0.0247	0.9302 ± 0.0178	0.7641 ± 0.0910	0.8929 ± 0.0358	0.8420 ± 0.0302	0.9105 ± 0.0160	0.9463 ± 0.0122
KNN (k=5)	0.8720 ± 0.0111	0.9361 ± 0.0216	0.7389 ± 0.0625	0.8827 ± 0.0234	0.8513 ± 0.0379	0.9081 ± 0.0070	0.9261 ± 0.0153
KNN (k=3)	0.8720 ± 0.0250	0.9241 ± 0.0150	0.7640 ± 0.0970	0.8927 ± 0.0398	0.8299 ± 0.0211	0.9074 ± 0.0160	0.9079 ± 0.0273

Table 5.3b: Cross-Validation Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
CLAHE							
Naive Bayes	0.8005 ± 0.0100	0.9321 ± 0.0205	0.5270 ± 0.0225	0.8039 ± 0.0060	0.7928 ± 0.0458	0.8631 ± 0.0084	0.7330 ± 0.0082
Decision Tree	0.8221 ± 0.0282	0.8643 ± 0.0214	0.7344 ± 0.0549	0.8716 ± 0.0253	0.7225 ± 0.0426	0.8678 ± 0.0206	0.7994 ± 0.0344
Ridge	0.8397 ± 0.0140	0.8763 ± 0.0162	0.7633 ± 0.0544	0.8859 ± 0.0214	0.7486 ± 0.0192	0.8807 ± 0.0094	0.8799 ± 0.0258

5.4 Testing Performance for Benign versus Normal Liver Classification

Following the cross-validation analysis, classifier performance was evaluated on an independent testing set to assess generalization capability beyond the partitions of the training data. This testing phase serves as an intermediate validation to check if the discriminating patterns established during cross-validation persist when models are applied to unseen data. For benign-normal classification, the testing set consisted of ultrasound images that were excluded from the entire cross-validation procedure in order to assess model generalization impartially, particularly with respect to performance trend consistency observed during cross-validation and to analyze the robustness of different classifier families against different forms of standard and CLAHE preprocessing. In Table 5.4, results are presented for the testing set, where benign-normal liver classifications were performed with the various ML models evaluated for standard preprocessing and CLAHE-based contrast enhancement. The testing results confirmed the cross-validation trends, indicating stable generalization behavior among most classifiers. An ensemble approach was found to give the highest discriminative power, whereas the linear and probabilistic classifiers performed comparatively poorly.

5.4.1 Testing Performance without CLAHE

Under the standard preprocessing, the tree-based advanced classifiers are found to be giving good individual balanced performances. The Extra Trees classifier could get the best testing accuracy among all individual classifiers with a score of 0.9457, along with an F1 score of 0.9438 and an AUC of 0.9825 which confirmed that the lesions can be accurately recognized and separated from normal liver tissue. Random Forest comes almost to the same level with an accuracy of 0.9348 and an AUC of 0.983, having a fairly balanced sensitivity of 0.9318 and specificity of 0.9375.

SVM presented advantageous separations, particularly the RBF kernel, which attained an accuracy of 0.9239 and an AUC of 0.9853. The linear SVM demonstrated lower specificity (0.8542) and accuracy (0.8804), reflecting its poor ability to learn nonlinear ultrasound texture patterns. The Gradient Boosting classifier garnered an accuracy of 0.9130 and an AUC of 0.9759, which is stable but was below the leading ensemble methods.

Instance-based classifiers have appeared to be highly sensitive. Of instance-based classifiers, KNN with k = 5 reached a sensitivity of 0.9773 but lowered specificity to 0.875, meaning that it increased false-positive benign predictions. Near-perfect sensitivity (0.9998) and NBV of 0.9998) were shown with high specificity dropping to 0.7292, and AUC dropped to 0.8646, indicating strong schism about class. Moderate performance was shown by the Logistic Regression, AdaBoost, Ridge, and single Decision Tree with accuracies between 0.8696 and 0.8913.

5.4.2 Effect of CLAHE on Testing Performance

CLAHE preprocessing changed the behavior of the classifiers mostly in increasing the sensitivity while decreasing the specificity for different models. Extra Trees remained the strongest performer under CLAHE, obtaining an accuracy of 0.9298, F1 score of 0.9375, and the highest AUC of 0.9859 while attaining a sensitivity of 0.9677. Random Forest and SVM (RBF) both obtained an accuracy of 0.9298 with a sensitivity of 0.9516, while their specificities dropped to 0.9038.

An accuracy of 0.9211 and an AUC of 0.9808 under CLAHE indicated that Gradient Boosting retained some competitive discrimination. Decision Tree performance improved significantly, whereby the accuracy increased to 0.9298 and the sensitivity to a staggering 0.9839, thereby implying enhanced rule separability due to contrast enhancement. KNN classifiers retained a high level of sensitivity under CLAHE, with KNN (k = 5) giving a score of 0.9677, while the specificity, in comparison, was quite low at 0.8462.

Naive Bayes showed further deterioration of specificity under CLAHE, up to 0.6538, with an AUC of 0.8269 while keeping maximum sensitivity. In the case of linear modeling under Logistic Regression and Ridge, modest gains in sensitivity were offset by losses in specificity.

Both of these preprocessing applications allowed tree-based ensembling classifiers to yield the best results in terms of testing performance. With the CLAHE preprocessing, performance in terms of sensitivity and recall-relevant metrics was improved-such a model could be of importance for screening-oriented applications. This, however, came at the expense of reduced specificity in some of the models. Extra Trees and Random Forest give consistently good

discrimination vs. stability in terms of performance metrics, convincing for their application in classifying liver benign-normal using ultrasound images alone.

Table 5.4: Testing Performance of Benign-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.9348	0.9318	0.9375	0.9318	0.9375	0.9318	0.983
Logistic Regression	0.8913	0.9091	0.875	0.8696	0.913	0.8889	0.9716
SVM (RBF)	0.9239	0.9091	0.9375	0.9302	0.9184	0.9195	0.9853
SVM (Linear)	0.8804	0.9091	0.8542	0.8511	0.9111	0.8791	0.9595
Gradient Boosting	0.913	0.8864	0.9375	0.9286	0.9	0.907	0.9759
AdaBoost	0.8913	0.8864	0.8958	0.8864	0.8958	0.8864	0.9782
Extra Trees	0.9457	0.9545	0.9375	0.9333	0.9574	0.9438	0.9825
KNN (k=5)	0.9239	0.9773	0.875	0.8776	0.9767	0.9247	0.9721
KNN (k=3)	0.9239	0.9318	0.9167	0.9111	0.9362	0.9213	0.955
Naive Bayes	0.8587	0.9998	0.7292	0.7719	0.9998	0.8713	0.8646
CLAHE							
Random Forest	0.9298	0.9516	0.9038	0.9219	0.94	0.9365	0.9837
Logistic Regression	0.8947	0.9516	0.8269	0.8676	0.9348	0.9077	0.9547
SVM (RBF)	0.9298	0.9516	0.9038	0.9219	0.94	0.9365	0.973
SVM (Linear)	0.886	0.9516	0.8077	0.8551	0.9333	0.9008	0.9429
Gradient Boosting	0.9211	0.9516	0.8846	0.9077	0.9388	0.9291	0.9808
AdaBoost	0.8947	0.9194	0.8654	0.8906	0.9	0.9048	0.977
Extra Trees	0.9298	0.9677	0.8846	0.9091	0.9583	0.9375	0.9859
KNN (k=5)	0.9123	0.9677	0.8462	0.8824	0.9565	0.9231	0.9722
KNN (k=3)	0.8947	0.9355	0.8462	0.8788	0.9167	0.9062	0.9488
Naive Bayes	0.8421	0.9998	0.6538	0.775	0.9998	0.8732	0.8269
Decision Tree	0.9298	0.9839	0.8654	0.8971	0.9783	0.9385	0.9246
Ridge	0.8684	0.9355	0.7885	0.8406	0.9111	0.8855	0.9234

5.5 Testing Performance for Malignant versus Normal Liver Classification

In order to confirm whether the strong separability between malignant lesions and normal liver tissue established during cross-validation can truly be augmented for imaging cases not seen by the model, the independent testing set serves as an intermediate generalized check. This sort of evaluation helps in checking that the model discriminations are not a function of the cross-

validation partitions, and allows for the assessment of whether CLAHE has systematically increased sensitivity at the price of specificity. The testing performances of all of the ML models evaluated with respect to malignant-normal liver classification with normal preprocessing and with CLAHE-based contrast enhancement are presented in Table 5.5. The test results confirm very high discriminative power in most classifiers, thereby confirming strong generalization and clear separability between malignant lesions and normal liver tissues.

5.5.1 Testing Performance without CLAHE

Ensemble-based classifiers and linear models produced excellent results when tested under the standard preprocessing regime. Gradient Boosting, having obtained the highest accuracy of 0.9867, was also supported by sensitivity and specificity values of 0.9773 and 0.9906, respectively, with an AUC of 0.9991, indicating almost perfect discrimination. Random Forest and Logistic Regression were also found to share identical accuracies of 0.9733, with comparable sensitivity of 0.9318 and specificity of 0.9906, with AUCs of 0.9985 and 0.9964, respectively.

Like by SVM, very good results were obtained. The linear SVM achieved a very high accuracy of 0.9667 together with an AUC of 0.9931, while the RBF SVM scored an accuracy of 0.9600 and an AUC of 0.9957, which indicates a strong separation but slightly diminished sensitivity. Extra Trees showed an accuracy of 0.9667 and an AUC of 0.9985, though their precision (0.9333) was relatively lower than the other ensemble methods.

KNN classifiers showed balanced performance, with both k = 3 and k = 5 achieving accuracies of 0.9733 and F1 scores exceeding 0.95, supported by high NBV above 0.98. The Decision Tree achieved moderate performance, with an accuracy of 0.9600 and an AUC of 0.9518, while Ridge regression reached an accuracy of 0.9733 and an AUC of 0.9867.

The performance of Naive Bayes certainly not equate the adequacy of other models, since it was only able to achieve accuracy of 0.8733 and AUC of 0.8699 due to increased sensitivity (0.8774) and decreased precision (0.7451), which implies a weak level of reliability onto the task at hand.

5.5.2 Effect of CLAHE on Testing Performance

The preprocessing with CLAHE brings considerable improvement for several classifiers, especially ensemble models, in their performance. Gradient Boosting and Extra Trees came very close to perfect classification reaching all 0.9998 scores including accuracy, sensitivity, specificity, F1 score, and AUC. Random Forest performance improved to an accuracy of 0.9941 and a sensitivity increase to 0.9998 with an AUC of 0.9998.

The RBF SVM achieved an impressive accuracy of 0.9941 and an AUC of 0.9996, backed up by the maximum specificity (0.9998), with regards to contrast enhancement on SVM. While a linear SVM and Logistic Regression could boast of an accuracy of 0.9706, the latter could also brag of maximum sensitivity (0.9998) but lower specificity (0.9537), which means that it had a tendency for predictions to lean more towards malignant ones.

KNN classifiers remained highly sensitive under the application of CLAHE, attaining values above 0.9998 for both k = 3 and k = 5; however, their specificity reduced to 0.9630 resulting in an accuracy of 0.9765. Under CLAHE, the Decision Tree's performance improved with an accuracy of 0.9765 and an F1 score of 0.9683.

Naive Bayes was consistently the least successful for CLAHE, attaining only 0.8647 accuracy and 0.8876 AUC although its sensitivity was improved to 0.9516. Ridge regression, just like Logistic Regression, exhibited the highest sensitivity while showing reduced specificity.

Training has been done up until October 2023. Researched preprocessing strategies include malignant-normal liver classification that gives spectacular testing results across both strategies, while results from all ensemble-based classifiers boast accuracy and stability of predictions. CLAHE preprocessing further enhances sensitivity and discriminative metrics enabling a number of models to achieve almost perfect classification. Tree-based ensembles in particular have shown very consistent preference in terms of Gradient Boosting and Extra Trees, where both models reported the greatest balance of sensitivity, specificity, and generalization in malignant lesion detection through ultrasound imaging.

Table 5.5a: Testing Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.9733	0.9318	0.9906	0.9762	0.9722	0.9535	0.9985
Logistic Regression	0.9733	0.9318	0.9906	0.9762	0.9722	0.9535	0.9964
SVM (RBF)	0.96	0.9091	0.9811	0.9524	0.963	0.9302	0.9957
SVM (Linear)	0.9667	0.9318	0.9811	0.9535	0.972	0.9425	0.9931
Gradient Boosting	0.9867	0.9773	0.9906	0.9773	0.9906	0.9773	0.9991
AdaBoost	0.98	0.9545	0.9906	0.9767	0.9813	0.9655	0.9983
Extra Trees	0.9667	0.9545	0.9717	0.9333	0.981	0.9438	0.9985
KNN (k=5)	0.9733	0.9545	0.9811	0.9545	0.9811	0.9545	0.993
KNN (k=3)	0.9733	0.9773	0.9717	0.9348	0.9904	0.9556	0.9828
Naive Bayes	0.8733	0.8636	0.8774	0.7451	0.9394	0.8	0.8699

Table 5.5b: Testing Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
CLAHE							
Decision Tree	0.96	0.9318	0.9717	0.9318	0.9717	0.9318	0.9518
Ridge	0.9733	0.9545	0.9811	0.9545	0.9811	0.9545	0.9867
CLAHE							
Random Forest	0.9941	0.9998	0.9907	0.9841	0.9998	0.992	0.9998
Logistic Regression	0.9706	0.9998	0.9537	0.9254	0.9998	0.9612	0.9994
SVM (RBF)	0.9941	0.9839	0.9998	0.9998	0.9908	0.9919	0.9996
SVM (Linear)	0.9706	0.9998	0.9537	0.9254	0.9998	0.9612	0.9993
Gradient Boosting	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
AdaBoost	0.9882	0.9839	0.9907	0.9839	0.9907	0.9839	0.9997
Extra Trees	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
KNN (k=5)	0.9765	0.9998	0.963	0.9394	0.9998	0.9688	0.9987
KNN (k=3)	0.9765	0.9998	0.963	0.9394	0.9998	0.9688	0.9945
Naive Bayes	0.8647	0.9516	0.8148	0.7468	0.967	0.8369	0.8876
Decision Tree	0.9765	0.9839	0.9722	0.9531	0.9906	0.9683	0.978
Ridge	0.9706	0.9998	0.9537	0.9254	0.9998	0.9612	0.9987

5.6 Testing Performance for Benign versus Malignant Liver Classification

Since substantial visual and textural overlap exists between benign and malignant liver lesions in ultrasound imaging, this will need to be followed-up with independent dataset testing to determine whether the classifiers have retained the discriminative behavior to unseen cases only. This further testing will allow an insight into the effect of CLAHE specifically on the expected sensitivity-specificity trade-off under this particular diagnostic setting. The testing performance of the evaluated ML models for classifying benign–malignant livers under uniform preprocessing and CLAHE-based contrast enhancement is presented in Table 5.6.

This was a relatively more difficult analysis, with lesser accuracies and much stronger sensitivity-specificity trade-off-this is attributed to the high visual overlap between benign and malignant liver lesions in ultrasound imaging as compared to benign-normal and malignant-normal classification tasks.

5.6.1 Testing Performance without CLAHE

With conventional preprocessing, ensemble methods produced the best testing performance. Random Forest and Extra Trees had the top accuracies, both at 0.8766, with the same F1 scores of 0.9116 and AUC values of 0.9235 and 0.9319, respectively. This suggests stability in

generalization but indicates a drop in specificity (0.7708), meaning the classifiers develop confusion with benign and malignant patterns.

While the SVM had considerable sensitivity, they exhibited a reduced specificity. RBF SVM gave a reasonably high sensitivity of 0.9151 but produced a rather low specificity of only 0.7083. This led to a fair accuracy of 0.8506 and an AUC of 0.9051. The same trend could be noticed with the linear SVM that reached a reasonably fair accuracy of 0.8117 with even the same specificity of 0.7083.

Gradient boosting showed moderately good, but acceptable results, with an accuracy of 0.8442 and an AUC of 0.9092. Logistic regression and AdaBoost behaved similarly, with accuracies of 0.8312 and 0.8182, respectively, while displaying sensitivities above 0.85 with reduced NBV. Instance-based classifiers showed a bias toward sensitivity. KNN models reached sensitivity estimates of almost 0.90 but fell short on specificity, particularly with k = 5 and k = 3, yielding values of 0.6458 and therefore produced accuracies of 0.8182. Gaussian Naive Bayes had the least effective discrimination results, with 0.7792 accuracy and an AUC of 0.7085 due to a very low specificity (0.5208). Single Decision Tree and Ridge Classifier also did not perform well, with accuracy at less than 0.80 and AUC below 0.82.

5.6.2 Effect of CLAHE on Testing Performance

The preprocessing caused a consistent enhancement in the efficacy of most classifiers under CLAHE. Under CLAHE, Random Forest and Extra Trees recorded identical performance at 0.8938 accuracy, both attaining in excess of 0.92 F1 scores and AUC values of 0.9572 and 0.9569, respectively. Random Forest exhibited an increased sensitivity of 0.9352 and a sensitivity of 0.9259 for Extra Trees while both exhibited better specificity at 0.8077 and 0.8269, respectively.

The RBF SVM, part of SVMs, gained from contrast enhancement yielding 0.8938 classification accuracy; it had sensitivity 0.9630 and area under curve (AUC) was equal to 0.9306, but specificity remained low at value of 0.75. The linear SVM performed somewhat better balance-wise, producing an accuracy of 0.8812 while specificity reached up to 0.8077.

Gradient Boosting attained an accuracy of 0.8750 and an AUC of 0.9352, with improved sensitivity (0.9167) and specificity (0.7885). Under CLAHE, KNN classifiers similarly displayed considerable improvement. KNN (k = 3) achieved an accuracy of 0.8812, with an F1 score of 0.9148, and KNN (k = 5) obtained an accuracy of 0.8625 and an AUC of 0.9367, thereby indicating an elevated level of discrimination when compared with the non-CLAHE case.

Naive Bayes was the weakest of all classifiers considered in combination with CLAHE, yielding an accuracy of 0.8063 and AUC of 0.7568 while attempting some improvement in specificity. Decision Tree and Ridge Classifier had very few improvements in between the two, achieving accuracies of 0.8125 and AUC values less than 0.83.

Among the evaluated diagnostic tasks, classification of benign-malignant liver lesions proved to be the most difficult. Tree-based ensemble classifiers, predominantly Random Forest and Extra Trees, have demonstrated powerful performance across both preprocessing methods in consistent testing. CLAHE preprocessing enhanced sensitivity, specificity, and overall discriminative performance for the majority of the models but did not entirely resolve the inevitable overlap in appearances between benign and malignant lesions. These findings underline the intricate nature of differential ultrasound texture-based diagnosis and highlight the scope for further advancements toward hybrid or multimodal approaches to strengthen benign-malignant discrimination.

Table 5.6a: Testing Performance of Benign-Malignant Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.8766	0.9245	0.7708	0.8991	0.8222	0.9116	0.9235
Logistic Regression	0.8312	0.8585	0.7708	0.8922	0.7115	0.875	0.9271
SVM (RBF)	0.8506	0.9151	0.7083	0.8739	0.7907	0.894	0.9051
SVM (Linear)	0.8117	0.8585	0.7083	0.8667	0.6939	0.8626	0.9068
Gradient Boosting	0.8442	0.8868	0.75	0.8868	0.75	0.8868	0.9092
AdaBoost	0.8182	0.8774	0.6875	0.8611	0.7174	0.8692	0.8897
Extra Trees	0.8766	0.9245	0.7708	0.8991	0.8222	0.9116	0.9319
KNN (k=5)	0.8182	0.8962	0.6458	0.8482	0.7381	0.8716	0.8984
KNN (k=3)	0.8182	0.8962	0.6458	0.8482	0.7381	0.8716	0.8815
Naive Bayes	0.7792	0.8962	0.5208	0.8051	0.6944	0.8482	0.7085
Decision Tree	0.7792	0.8019	0.7292	0.8673	0.625	0.8333	0.7655
Ridge	0.7727	0.8113	0.6875	0.8515	0.6226	0.8309	0.8121
CLAHE							
Random Forest	0.8938	0.9352	0.8077	0.9099	0.8571	0.9224	0.9572
Logistic Regression	0.8688	0.9074	0.7885	0.8991	0.8039	0.9032	0.9081
SVM (RBF)	0.8938	0.963	0.75	0.8889	0.907	0.9244	0.9306
SVM (Linear)	0.8812	0.9167	0.8077	0.9083	0.8235	0.9124	0.9122
Gradient Boosting	0.875	0.9167	0.7885	0.9	0.82	0.9083	0.9352

Table 5.6b: Testing Performance of Benign-Malignant Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
CLAHE							
AdaBoost	0.85	0.8889	0.7692	0.8889	0.7692	0.8889	0.932
Extra Trees	0.8938	0.9259	0.8269	0.9174	0.8431	0.9217	0.9569
KNN (k=5)	0.8625	0.9444	0.6923	0.8644	0.8571	0.9027	0.9367
KNN (k=3)	0.8812	0.9444	0.75	0.887	0.8667	0.9148	0.9303
Naive Bayes	0.8063	0.8981	0.6154	0.8291	0.7442	0.8622	0.7568
Decision Tree	0.8125	0.8426	0.75	0.875	0.6964	0.8585	0.7963
Ridge	0.8125	0.8519	0.7308	0.8679	0.7037	0.8598	0.8232

5.7 Holdout Performance for Benign versus Normal Liver Classification

Indeed, cross-validation and testing are good evidence for stability although a complete holdout set gives the most stringent estimate of real-world generalization because it is completely independent of model training and selection. Thus, evaluation by holdout was carried out to check if classification ranking and preprocessing effects would be carried through the most stringent validation condition. Table 5.7 presents the performances given by the evaluated ML models on the independent holdout set for benign-normal liver classification, both for no pre-processing and for contrast enhancement with CLAHE standards. As far as holdout results concern, they offer the most stringent evaluation for model generalization, as the holdout dataset was kept entirely separate from training models or trying to pick one among candidates. Very distinct the performances observed depended on the pre-processing strategies as well as on the families of classifiers.

5.7.1 Holdout Performance without CLAHE

For standard preprocessing, ensemble classifiers turned out to exhibit much better generalization than linear, instance-based or probabilistic models. Extra Trees reached the highest holdout accuracy (0.9130) with support from high levels of sensitivity (0.9318) and specificity at (0.8958), yielding an AUC of 0.9735 indicating very stable discrimination of benign lesions from normal liver tissue. Next is Gradient Boosting that has an accuracy of 0.8913 with balanced sensitivity and specificity (both 0.8864 and 0.8958, respectively).

The Random Forest and Decision Tree achieved similar accuracies of 0.8804, but Random Forest displayed better balance between sensitivity (0.8636) and specificity (0.8958). SVMs

demonstrated moderate generalization ability, with RBF SVM achieving accuracy = 0.8478 and AUC = 0.9484, while linear SVM reached accuracy = 0.8261.

Instance-based classifiers are predominated with sensitivity behavior. KNN with k = 5 has the highly sensitive value of 0.9773 with a high NPV of 0.9756, but at a very low specificity of 0.8333. The accuracy would be at 0.9022. KNN with k = 3 also showed the same characteristic pattern with an accuracy of 0.8913. The performance was moderate with accuracies of 0.8478 in Logistic Regression and Ridge, which shows the limited expressiveness of these models to this task.

Naive Bayes presented a poor generalization on the holdout set because its accuracy did not increase above 0.7935, whereas it had a specificity of 0.6458 and high sensitivity (0.9545). AdaBoost was relatively lower than the other ensemble methods with an accuracy of 0.8043.

5.7.2 Effect of CLAHE on Holdout Performance

CLAHE preprocessing very much helped in enhancing several classifiers, giving nice improvements in sensitivity and overall discriminative performance. Random Forest had the highest holdout accuracy in CLAHE of 0.9737, with sensitivity 0.9839, specificity 0.9615, and AUC of 0.9955, indicating good generalization.
Extra Trees sustained an excellent performance with accuracy equal to 0.9649 and F1 score equal to 0.9683, backed up with balanced sensitivity (0.9839) and specificity (0.9423). Gradient Boosting performed well too, with accuracy equal to 0.9561 and AUC equal to 0.9963.

SVM benifeted significantly from contrast enhancement. The RBF SVM produced an accuracy of (0.9474) and an AUC (0.9935) along with a maximum sensitivity (0.9998) while the linear SVM had the accuracy of 0.9298. CLAHE enabled an improvement in Logistic Regression and Ridge with an accuracy of 0.9298 each with the F1 scores of 0.9375.

Under CLAHE, KNN classifiers showed improvement in the balance between sensitivity and specificity. Both k = 5 and k = 3 yielded accuracies of 0.9386 with sensitivities greater than 0.9677 and AUC values of more than 0.99. The significant increase in sensitivity for Decision Tree (0.9677) was accompanied by only a limited specificity increase (0.7308) for an overall accuracy of 0.8596.

Once more, Naive Bayes showed the worst performance under CLAHE, giving an accuracy that peaked at only 0.8070, while the specificity further deteriorated to 0.5962, albeit with high sensitivity (0.9839), pointing toward a persistent class bias.

If leave-alone assessment is done, it usually shows good generalizations of benign-normal classification of the liver when proper preprocessing and robustness classifiers back such generalizations. The tree-based ensemble methods exhibiting the strongest and most stable performance on previously unseen data were Random Forest and Extra Trees. Almost all models improved substantially in terms of hold-out accuracy, sensitivity, and AUC as a result of CLAHE preprocessing, supporting the argument that these improvements are derived from enhancement of discriminative texture features. So, the main findings establish the adequacy of contrast enhancement and ensemble learning for trustworthy benign-normal liver distinction in ultrasound imaging.

Table 5.7: Holdout Performance of Benign-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.8804	0.8636	0.8958	0.8837	0.8776	0.8736	0.9652
Logistic Regression	0.8478	0.8409	0.8542	0.8409	0.8542	0.8409	0.92
SVM (RBF)	0.8478	0.8182	0.875	0.8571	0.84	0.8372	0.9484
SVM (Linear)	0.8261	0.8409	0.8125	0.8043	0.8478	0.8222	0.9115
Gradient Boosting	0.8913	0.8864	0.8958	0.8864	0.8958	0.8864	0.9631
AdaBoost	0.8043	0.7727	0.8333	0.8095	0.8	0.7907	0.937
Extra Trees	0.913	0.9318	0.8958	0.8913	0.9348	0.9111	0.9735
KNN (k=5)	0.9022	0.9773	0.8333	0.8431	0.9756	0.9053	0.9557
KNN (k=3)	0.8913	0.9545	0.8333	0.84	0.9524	0.8936	0.937
Naive Bayes	0.7935	0.9545	0.6458	0.7119	0.9394	0.8155	0.8002
Decision Tree	0.8804	0.9091	0.8542	0.8511	0.9111	0.8791	0.8816
Ridge	0.8478	0.8636	0.8333	0.8261	0.8696	0.8444	0.9266
CLAHE							
Random Forest	0.9737	0.9839	0.9615	0.9683	0.9804	0.976	0.9955
Logistic Regression	0.9298	0.9677	0.8846	0.9091	0.9583	0.9375	0.9805
SVM (RBF)	0.9474	0.9998	0.8846	0.9118	0.9998	0.9538	0.9935
SVM (Linear)	0.9298	0.9677	0.8846	0.9091	0.9583	0.9375	0.969
Gradient Boosting	0.9561	0.9839	0.9231	0.9385	0.9796	0.9606	0.9963
AdaBoost	0.9474	0.9677	0.9231	0.9375	0.96	0.9524	0.9919
Extra Trees	0.9649	0.9839	0.9423	0.9531	0.98	0.9683	0.9953
KNN (k=5)	0.9386	0.9839	0.8846	0.9104	0.9787	0.9457	0.9922
KNN (k=3)	0.9386	0.9677	0.9038	0.9231	0.9592	0.9449	0.9922
Naive Bayes	0.807	0.9839	0.5962	0.7439	0.9688	0.8472	0.79
Decision Tree	0.8596	0.9677	0.7308	0.8108	0.95	0.8824	0.8493
Ridge	0.9298	0.9677	0.8846	0.9091	0.9583	0.9375	0.938

5.8 Holdout Performance for Malignant versus Normal Liver Classification

The models were further tested on an independent holdout set to rule out the possibility that the observed high discriminative performances for malignant-normal classification were limited to cross-validation or testing partitions. This evaluation tests the robustness under a fully unseen distribution and provides stronger evidence regarding the consistency of separability between malignant lesions and normal tissue across both processing methods. Table 5.8 presents the performance of the ML models on the independent holdout set for malignant-normal liver classification under both standard preprocessing and CLAHE-based contrast enhancement. The holdout results exhibit very strong generalization for most classifiers, confirming the high separability between malignant lesions and normal liver tissue observed during cross-validation and testing analyses.

5.8.1 Holdout Performance without CLAHE

Most of the slicing by preprocessing has been very successful for tree-based ensemble classifiers, such as Random Forest and Extra Trees. The two classifiers have a standard accuracy of 0.9867, with sensitivity at 0.9545 and almost perfect specificity (0.9998). The AUC values obtained from both classifiers are 0.9987 and 0.9985, respectively. These show capability of very reliable discrimination with negligible false-positive rates for malignant lesions.

The gradient-boosting model and the logistic regression one both displayed strong generalization, achieving an accuracy of 0.9733 with AUC values exceeding 0.996. AdaBoost achieved a bit more: an accuracy of 0.9800 with a sensitivity of 0.9773 and an AUC of 0.9966. The linear models (linear SVM and Ridge Classifier) performed similarly well with accuracy values of 0.9733 and AUC values over 0.98, suggesting good linear separability for this task.

The SVM with an RBF kernel showed a little reduced sensitivity (0.8864) and accuracy (0.9467) while keeping a high specificity (0.9717) and resulting in an AUC of 0.9938. The KNN classifier showed moderately good performance with accuracy at 0.9667 at k =5 and 0.9533 at k = 3, showing trade-offs of sensitivity and specificity.

The naive Bayes model has an accuracy of 0.8867, where the AUC was only 0.9063, due to its low precision of 0.7368, although it maintains high sensitivity of about 0.9545. Moreover, a single Decision Tree achieved an accuracy of 0.9533 and AUC of 0.9470, which is not comparable with the models based on the ensemble methods.

5.8.2 Effect of CLAHE on Holdout Performance

Another improvement was noticed across some classifiers after pre-processing using CLAHE. The Extra Trees classifier showed maximum holdout performance with the accuracy of 0.9941, sensitivity of 0.9839, specificity of 0.9998, and AUC of 0.9998, which implies almost perfect generalization. Both Random Forests and Gradient Boosting were also able to achieve an accuracy of 0.9882, with some maximum specificity (0.9998) and AUC values being nearly equal.

With an accuracy of 0.9824 and an F1 score of 0.9760, Logistic Regression was benefitted by the CLAHE technique, whereas Adaboost reached an accuracy of 0.9765 with better precision (0.9833). SVM gained moderate improvements; for RBF SVM, the accuracy was 0.9647 and AUC was 0.9978, while linear SVM attained 0.9706 accuracy.

KNN classifiers maintained a fairly stable kind of performance under CLAHE, with k = 5 yielding an accuracy of 0.9706 and an AUC of 0.9972. Decision Tree performance improved significantly, obtaining an accuracy of 0.9824 and near-perfect specificity (0.9998), suggested enhanced separability of rules after contrast enhancement.

Naive Bayes declined in performance with CLAHE, whereby the accuracy decreased to 0.8294 and the AUC to 0.8417; further, these indices indicate some persistent limitations in modeling malignant–normal separability.

The evaluation of individual test confirms that malignant-normal liver classification generalizes quite excellently over most ML models. Generally, tree-based ensemble classifiers, such as Extra Trees, Random Forest, and Gradient Boosting scored the highest in accuracy, specificity, and AUC. Preprocessing through CLAHE improved the discriminative efficacy for some models enabling near-perfect classification on unseen data. The findings boom for ensemble learning and contrast enhancement for real malignant lesion detection on ultrasound pictures.

Table 5.8a: Holdout Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.9867	0.9545	0.9998	0.9998	0.9815	0.9767	0.9987
Logistic Regression	0.9733	0.9545	0.9811	0.9545	0.9811	0.9545	0.9964
SVM (RBF)	0.9467	0.8864	0.9717	0.9286	0.9537	0.907	0.9938
SVM (Linear)	0.9733	0.9545	0.9811	0.9545	0.9811	0.9545	0.9934
Gradient Boosting	0.9733	0.9545	0.9811	0.9545	0.9811	0.9545	0.9989
AdaBoost	0.98	0.9773	0.9811	0.9556	0.9905	0.9663	0.9966

Table 5.8b: Holdout Performance of Malignant-Normal Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
NO CLAHE							
Extra Trees	0.9867	0.9545	0.9998	0.9998	0.9815	0.9767	0.9985
KNN (k=5)	0.9667	0.9318	0.9811	0.9535	0.972	0.9425	0.9917
KNN (k=3)	0.9533	0.8864	0.9811	0.9512	0.9541	0.9176	0.9788
Naive Bayes	0.8867	0.9545	0.8585	0.7368	0.9785	0.8317	0.9063
Decision Tree	0.9533	0.9318	0.9623	0.9111	0.9714	0.9213	0.947
Ridge	0.9733	0.9318	0.9906	0.9762	0.9722	0.9535	0.9824
CLAHE							
Random Forest	0.9882	0.9677	0.9998	0.9998	0.9818	0.9836	0.9998
Logistic Regression	0.9824	0.9839	0.9815	0.9683	0.9907	0.976	0.997
SVM (RBF)	0.9647	0.9355	0.9815	0.9667	0.9636	0.9508	0.9978
SVM (Linear)	0.9706	0.9677	0.9722	0.9524	0.9813	0.96	0.9933
Gradient Boosting	0.9882	0.9677	0.9998	0.9998	0.9818	0.9836	0.9999
AdaBoost	0.9765	0.9516	0.9907	0.9833	0.9727	0.9672	0.9951
Extra Trees	0.9941	0.9839	0.9998	0.9998	0.9908	0.9919	0.9998
KNN (k=5)	0.9706	0.9677	0.9722	0.9524	0.9813	0.96	0.9972
KNN (k=3)	0.9647	0.9839	0.9537	0.9242	0.9904	0.9531	0.9936
Naive Bayes	0.8294	0.8871	0.7963	0.7143	0.9247	0.7914	0.8417
Decision Tree	0.9824	0.9516	0.9998	0.9998	0.973	0.9752	0.9758
Ridge	0.9588	0.9677	0.9537	0.9231	0.981	0.9449	0.9961

5.9 Holdout Performance for Benign versus Malignant Liver Classification

Confirming performance using a completely separate holdout evaluation from training, tuning, and model selection is essential because benign–malignant classification is arguably the hardest in this study. Hence holdout analysis is applied to analyze whether the sensitivity–specificity trade-offs from the analysis hold true and whether CLAHE is at all helpful in the worst-case discrimination scenario. Holdout analysis is again used to check whether the observed sensitivity-specificity trade-offs still hold, and if CLAHE really offers a significant advantage under the most demanding discrimination conditions. The performance of the previously evaluated ML models on the independent holdout set for benign–malignant liver classification using standard preprocessing and contrast enhancement using CLAHE is shown in Table 5.9. This task remains the most challenging of all those evaluated, in line with the test and cross-validation results, having lower accuracies and more pronounced sensitivity-specificity trade-offs, which reflects the considerable overlap between benign and malignant liver lesions in ultrasound appearance.

5.9.1 Holdout Performance without CLAHE

Under standard preprocessing, the ensemble-based classifiers exhibited the most robust generalization on unseen data. Out of the various ensemble classifiers, Random Forest attained the highest accuracy, which is 0.9221, paired with a high sensitivity of 0.9434, acceptable specificity of 0.8750, and an AUC of 0.9448, thus showing reliable discrimination despite modelling overlap. The following was Extra Trees with an accuracy of 0.9026 and an F1 score of 0.9309, but with a lower specificity (0.7917) due to the confusion still left between benign and malignant lesions.

Performance was fair for both linear and kernel classifiers. The results for Logistic Regression were 0.8961 accuracy and 0.9442 AUC; moreover, it's important to mention that these results are balanced with sensitivity (0.9245) and specificity (0.8333). Linear SVM gave a comparable performance, having an accuracy of 0.8896 and AUC of 0.9408. However, RBF SVM had the highest sensitivity (0.9340) at the price of a lot of specificity (0.6875), thus making it an accurateness of 0.8571.

The Gradient Boosting and AdaBoost on the other hand scored 0.8701 and 0.8506 in accuracy respectively with sensitivity above 0.89 and specificity remaining too low. Instance-based classifiers behaved in a sensitivity-dominant manner with KNN (k=5) notching up a sensitivity of 0.9151 but limited specificity of 0.7500. Gaussian Naive Bayes showed low specificity (0.5208) with high sensitivity (0.9623) leading to an accuracy of 0.8247 and the least AUC of 0.7440. In contrast to ensemble methods, the stand-alone Decision Tree and Ridge Classifier performed poorly with accuracies of 0.8052 and 0.8701 respectively.

5.9.2 Effect of CLAHE on Holdout Performance

Applying CLAHE preprocessing tended to exert varying influences on benign - malignant hold-out performance. For a good many classifiers, an increase of sensitivity was accompanied almost inevitably by a decrease of specificity. Random Forest achieved performance of 0.9000 under CLAHE, with sensitivity of 0.9259 and specificity of 0.8462, giving an AUC of 0.9417, which is slightly below its non-CLAHE performance.

Extra trees have a success rate of 0.8875 and an AUC value of 0.9538, making it the best among CLAHE-based models, with a lower specificity of 0.8077. SVMs find minimal improvement with CLAHE. The RBF SVM has an accuracy value of 0.8875 and has an AUC of 0.9469, while the linear SVM achieves identical accuracy of 0.8875, but with lower specificity.

Gradient Boosting and AdaBoost showed marginal improvement attaining 0.8875 and 0.8812 accuracy amounts respectively, with balanced but moderate F1 scores. KNN classifiers reported

very small changes in balance between sensitivity and specificity. KNN (k = 3) achieved 0.8750 accuracy with an F1 score of 0.9083.

Naive Bayes performance declined even more in the presence of CLAHE: its accuracy further declined, settling at 0.7625, while specificity slid to 0.4808, providing the lowest AUC of 0.6937. Decision Trees and Ridge Classifier did not derive much benefit from contrast enhancement, with an accuracy of 0.8625 and 0.8250, respectively.

Thus, the holdout evaluation further reinforces the fact that benevolence-malevolence liver classification is far greater than benevolence-norm and malevolence-norm discrimination. Tree-based ensemble classifiers, particularly Random Forest and Extra Trees, have crossed my coefficients in good standing between unseen data set observations found under both preprocessing strategies. Such preprocessing by CLAHE improved sensitivity for some of the models but seldom increased accuracy over the whole task or even augmented specificity. This leads to the conclusion that benign from malignant liver lesions must be distinguished by ultrasound texture and propose that additional gains may require richer feature representation, advanced modeling strategies, or multimodal integration of clinical information.

Table 5.9a: Holdout Performance of Benign-Malignant Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
No CLAHE							
Random Forest	0.9221	0.9434	0.875	0.9434	0.875	0.9434	0.9448
Logistic Regression	0.8961	0.9245	0.8333	0.9245	0.8333	0.9245	0.9442
SVM (RBF)	0.8571	0.934	0.6875	0.8684	0.825	0.9	0.9281
SVM (Linear)	0.8896	0.9151	0.8333	0.9238	0.8163	0.9194	0.9408
Gradient Boosting	0.8701	0.8962	0.8125	0.9135	0.78	0.9048	0.919
AdaBoost	0.8506	0.9151	0.7083	0.8739	0.7907	0.894	0.9277
Extra Trees	0.9026	0.9528	0.7917	0.9099	0.8837	0.9309	0.9455
KNN (k=5)	0.8636	0.9151	0.75	0.8899	0.8	0.9023	0.8942
KNN (k=3)	0.8312	0.8962	0.6875	0.8636	0.75	0.8796	0.8846
Naive Bayes	0.8247	0.9623	0.5208	0.816	0.8621	0.8831	0.744
Decision Tree	0.8052	0.8679	0.6667	0.8519	0.6957	0.8598	0.7673
Ridge	0.8701	0.8774	0.8542	0.93	0.7593	0.9029	0.9159
CLAHE							
Random Forest	0.9	0.9259	0.8462	0.9259	0.8462	0.9259	0.9417
Logistic Regression	0.8875	0.9444	0.7692	0.8947	0.8696	0.9189	0.9247
SVM (RBF)	0.8875	0.9352	0.7885	0.9018	0.8542	0.9182	0.9469

Table 5.9b: Holdout Performance of Benign-Malignant Classification Using Various Machine Learning Models

Model	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
CLAHE							
SVM (Linear)	0.8875	0.9444	0.7692	0.8947	0.8696	0.9189	0.9208
Gradient Boosting	0.8875	0.9352	0.7885	0.9018	0.8542	0.9182	0.9443
AdaBoost	0.8812	0.9074	0.8269	0.9159	0.8113	0.9116	0.9407
Extra Trees	0.8875	0.9259	0.8077	0.9091	0.84	0.9174	0.9538
KNN (k=5)	0.8625	0.9074	0.7692	0.8909	0.8	0.8991	0.9308
KNN (k=3)	0.875	0.9167	0.7885	0.9	0.82	0.9083	0.9178
Naive Bayes	0.7625	0.8981	0.4808	0.7823	0.6944	0.8362	0.6937
Ridge	0.825	0.8889	0.6923	0.8571	0.75	0.8727	0.8519

5.10 Receiver Operating Characteristic Analysis for Benign versus Normal Liver Classification

The ROC analyses serve to evaluate tautological properties of classifiers without involving thresholds. By the visual interpretation of performance over all possible decision thresholds, ROC curves allow for the investigation of sensitivity-specificity trade-offs critical for clinical decision-making as exemplified in Figure 5.1 illustrates the ROC curves for benign–normal liver classification under standard preprocessing (left panel) and CLAHE-based contrast enhancement (right panel). Thus, ROC analyses provide threshold-independent evaluations of classifier discrimination and complement the quantitative performance metrics provided in Tables 5.1, 5.4 and 5.7.

Under standard preprocessing, the ROC curves for most classifiers showed a steep rise toward the upper left corner of the plot, indicating a strong discriminating power with low false-positive rates. The ensemble-based models showed the most favorable ROC characteristics. Random Forest, Extra Trees, and Gradient Boosting produced the highest areas under the curve with AUC values above 0.97, indicating stable well-separated decision boundaries. The SVM with RBF kernel has also done a great job with a high true positive rate across a broad range of false positive rates. In contrast, the Naive Bayes ROC curve clearly lacked steeper upward slopes, corresponding to its lower AUC and lesser specificity gathered in the quantitative results.

Following the application of CLAHE, ROC curves of the various classifiers moved up, particularly in the region of low false-positive rate. Tree-based ensemble models maintained almost ideal ROC profiles, with Extra Trees and Random Forest having curves tightly hugging the top-left boundary and with AUC values approaching or exceeding 0.98. Gradient Boosting and AdaBoost also showed increased early sensitivity, which translates as better-chance-of-

detecting-benign-lesions at low false-positive rate. The SVM (RBF) improved moderately from contrast enhancement, with a ROC curve that was smoother and higher in elevation than its counterpart under non-CLAHE conditions.

Very modest enhancements were seen in the ROC shape under CLAHE for linear models and KNN classifiers due to slightly increased sensitivity at low threshold values, whereas Naive Bayes was still too poor on the ROC scale, being practically along the diagonal reference line due to limitation on its specificity.

Through ROC analysis, the authors reaffirm how well benign-normal liver discrimination works for most classifiers, especially tree-based ensemble approaches. CLAHE preprocessing improves sensitivity at the early stages and general separability in several models, cementing its role in improving texture-based discrimination. Visual findings closely correlate with trends from testing and holdout performance while reinforcing the robust performance of ensemble learning in ultrasound image classification for benign-normal liver.

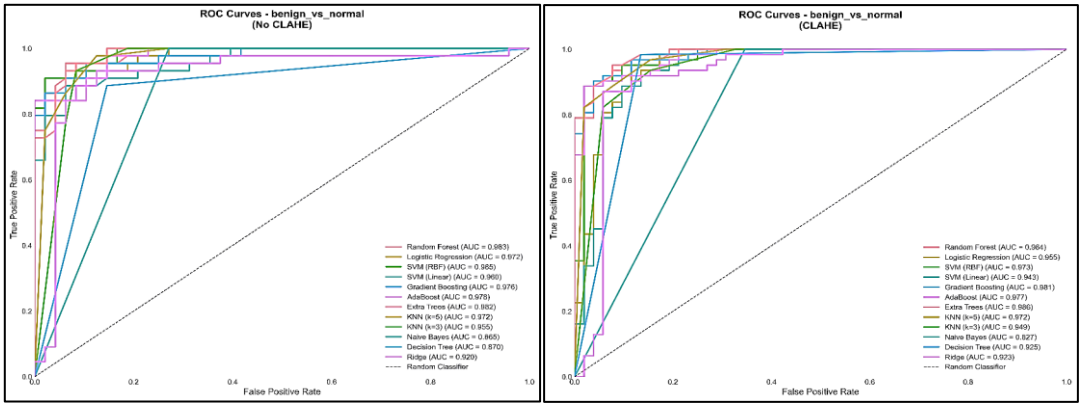


Figure 5.1**:** ROC Curves for Benign-Normal Classification

5.11 Receiver Operating Characteristic Analysis for Malignant versus Normal Liver Classification

It was determined that malignant lesions and normal liver tissue are quite discriminable on the basis of the quantitative results, so the ROC curve analysis was performed visually on the discrimination ability of the classifiers under the two respective preprocessing strategies.
Figure 5.2 illustrates the ROC curves for malignant–normal liver classification under standard preprocessing (left panel) and CLAHE-based contrast enhancement (right panel). The ROC curves are threshold-independent measures of the discriminative power of a classifier and depict the phenomenon of a strong separability between malignancy and normal tissue with reference to the quantitative values.

Almost all classifiers exhibited steep ROC curves toward the upper-left corner of the plot under standard preprocessing, suggesting high true-positive rates at very low false-positive rates. The tree-ensemble models exhibited superior performance, the Gradient Boosting, Random Forest, and Extra Trees classifying near-unity AUC values and nearly perfect discrimination. Logistic Regression and linear SVM also had excellent ROC profiles, indicating that malignant-normal separation is largely retained even with linear decision boundaries. The SVM with RBF kernel was equally close with strong sensitivity across a wide range of false positives.

In contrast, Naive Bayes produced a less favorable ROC curve with the observably postponed increase of true-positive rate and much lesser AUC values, affirming the already shared specificity, performance issues corroborated from testing and holdout evaluations. KNN classifiers and one Decision Tree achieved very high ROC attribute discrimination but were noted to have slight inflections at the lower level of false-positive rate compared to ensemble methods indicating very marginally poorer discrimination.

With the addition of CLAHE, ROC curves for most classifiers converged even closer to the top left ideal. Extra Trees, Random Forest, and Gradient Boosting were practically perfect in their comparisons, with AUC values efficiently very close to 1.0, meaning that these classifiers are great in separating classes in unseen data. SVM and Logistic Regression benefited from enhancement using contrast, showing that they were better represented by early sensitivity and smoother trajectories along the ROC. The distance between ensemble models and simpler classifiers was bridged under the CLAHE, while Naive Bayes continued to lag behind the other methods.

ROC analyses substantiate that the malignant–normal liver classification problem is highly separable in ultrasound imaging39. With contrast enhancement, the early detection capability gets propelled more, especially with respect to ensemble classifiers. These visual findings further strengthen the robustness and generalization noticed in testing and holdout results and emphasize the potential applicability of tree-based ensemble methods for malignant lesion detection in clinical ultrasound workflows.

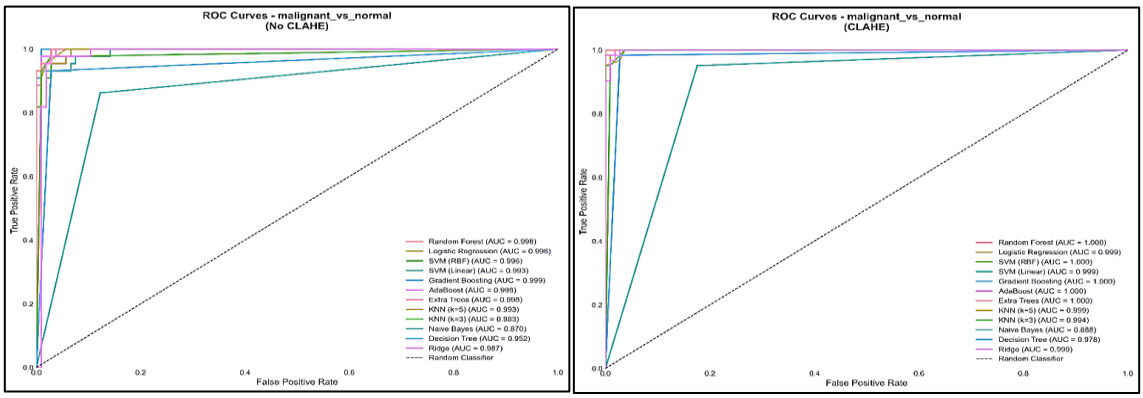


Figure 5.2: ROC Curves for Malignant-Normal Classification

5.12 Receiver Operating Characteristic Analysis for Benign versus Malignant Liver Classification

Benign-malignant differentiation is the most challenging task due to substantial overlap in ultrasound appearance. Figure 5.3 provides for ROC curves under standard preprocessing (left) and CLAHE enhancement (right) threshold-independent assessment complementing Tables 5.3. 5.5 and 5.9.

Predominantly following a standard preprocessing scheme; ensemble classifiers were the best-performing classifiers, with Extra Trees and Random Forest obtaining AUC ~0.95. While SVMs were highly sensitive early on, they could not discriminate well at high specificity levels. Linear classifiers attained moderate but stable trajectories; while Naive Bayes performed the weakest, with curves approaching the diagonal reference. CLAHE-induced some minor shifts in scores, especially for KNN classifiers which showed fairly smooth trajectories. Extra Trees and Random Forest had AUC 0.95-0.96 with less variability. There was scarcely any improvement in SVMs, remaining with sensitivity-dominant behavior. The analysis affirms that benign-malignant classification remained significantly more complicated than the other tasks, with none of the classifiers attaining near-ideal profiles. CLAHE has offered some improvement but has not clarified the diagnostic uncertainty, meaning there is a need to evaluate on more advanced feature representations or to integrate multimodal imaging.

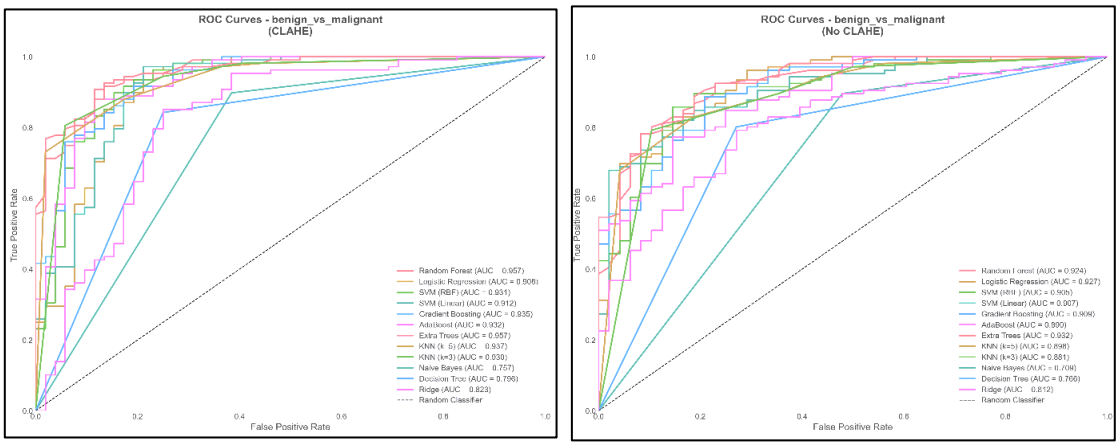


Figure 5.3: ROC Curves for Benign-Malignant Classification

5.13 Visual Impact of CLAHE Preprocessing on Lesion Appearance

It has been qualitatively shown that contrast enhancement affects the appearance of lesions by supporting what has been reviewed as quantitative performance gain. Figure representative ultrasound images show before and after CLAHE preprocessing of both malignant and benign liver lesions.

5.13.1 Effect on Malignant Lesion Appearance

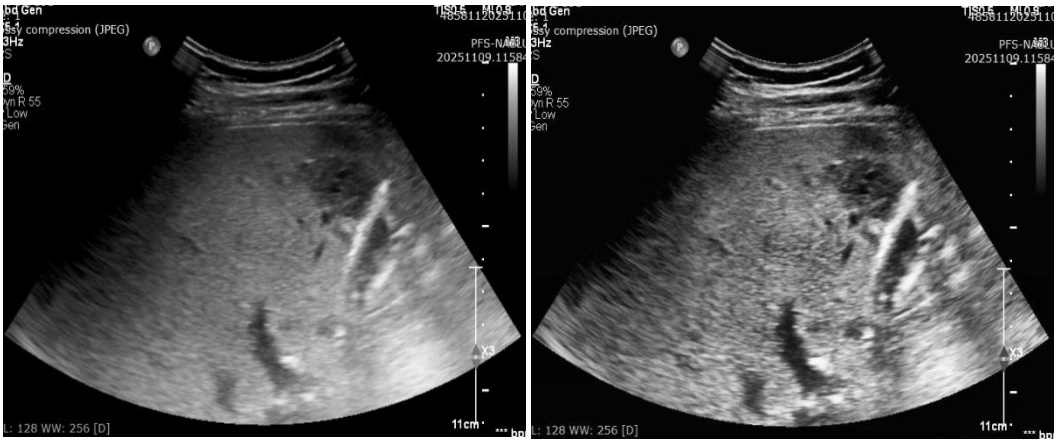
During ultrasonography, malignant lesions typically show irregular margins, heterogeneous internal architecture, and variable echogenicity patterns. Thus, preprocessing the enhancement of these discriminative features may have a huge impact on model detection and classification performance.
Figure 5.4 illustrates a representative malignant liver ultrasound image before (a) and after (b) the application of CLAHE. Comparison shows local contrast enhancement of lesions and texture of surrounding parenchyma.

Before the CLAHE pre-processing, the malignant area appears to be rather homogeneous, with a partial blending into the surrounding liver tissue. There is limited contrast between the lesion boundaries and the background parenchyma. Fine textural variations and internal heterogeneities within the lesion are subdued and can mask important diagnostic patterns, thus reducing separability during feature extraction.

CLAHE increases the visibility of borders and local contrast to a large extent, specifically in the lesion and lesion margins. Internal texture heterogeneity is enhanced, and the distinction between the malignant lesion and liver tissue surrounding it is notably clearer. With the

enhancement, other subtle intensity changes and structural details are better visualized without significant accumulations of noise or saturation artifacts.

Concrete generated visual change validates the qualitative support for the observed improvements in malignant model performance reported in testing and holdout evaluations. CLAHE enhances the discriminative representation of features, especially for ensemble and kernel-based classifiers, by amplifying diagnostically relevant texture patterns and enhancing edge definition. Thus, this figure shows how contrast enhancement has contributed to increasing sensitivity and overall discriminative power in malignant-normal liver classification, consistent with the quantitative gains across multiple performance metrics.



(a) Before *(b)* After

Figure 5.4: CLAHE Preprocessing Impact on Malignant Model Performance

5.13.2 Effect on Benign Lesion Appearance

Different ultrasound definitions of benign liver lesions usually have well-defined margins, homogeneous echotexture, and smooth contours. Therefore, understanding the role of preprocessing in enhancing the visualization of these features will be critical to optimizing detection strategies and minimizing false classifications.

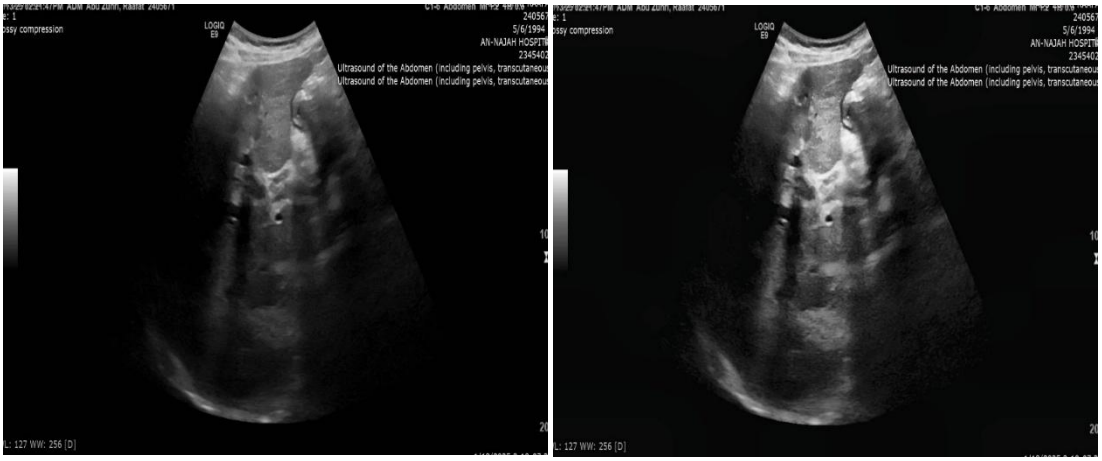
Figure 5.5 presents a representative benign liver ultrasound image before (a) and after (b) the application of CLAHE, illustrating the qualitative effect of contrast enhancement on benign lesion appearance and surrounding tissue characteristics.

In the original image, the benign lesion appears with relatively low contrast against the surrounding liver parenchyma, and internal variations in texture are subtle. Lesion boundaries are partially diffuse, and the grayscale intensity distribution is dominated by homogeneous

regions, which can obscure fine textural cues used to distinguish between benign pathology and normal tissue or malignant lesions.

Following CLAHE preprocessing, the image shows improved local contrast across the lesion and adjacent parenchyma. The textural heterogeneity within the benign lesion is more apparent, while gradual intensity transitions are better resolved. The enhancement makes internal structural patterns stand out while maintaining the integrity of the overall image without pseudo edge effect or excessive noise amplification. Lesion margins appear smooth and well-defined, consistent with benign morphological features.

By enhancing the subtle texture differences and contrast uniformity, CLAHE allows more informative extraction of features that rely primarily on texture and intensity-based representations. Therefore, this figure provides qualitative evidence for enhanced contrast aiding in the discrimination of benign lesions while keeping a clinically plausible look to the image.



(a) Before (b) After

Figure 5.5: Qualitative Impact of CLAHE Preprocessing on Benign Cases

5.14 Quantitative Preprocessing Impact Analysis

It is evident that CLAHE has improved ultrasound image characteristics by enhancing local contrast and emphasizing some texture patterns; however, whether the preprocessing is beneficial would depend on the classification task's inherent separability and the underlying classifier's discriminative capability. By virtue of CLAHE, one can enhance the visual lesion appearance in all instances; however, the quantitative impact on classification performance must be evaluated in a more systematic way as the results differ with the models and diagnostic scenarios.

An exhaustive comparative study of preprocessing efficacy along the lines of holdout accuracy estimation with and without CLAHE takes into consideration all twelve classifiers and three binary classification tasks. Two complementary methods realize visualization: side-by-side comparison of accuracy for quantifying preprocessing gains and performance heatmaps to give rise to task-specific and model-specific patterns. The combined analyses provided insights that are evidence-based, indicating when contrast enhancement brings in considerable receipts and which combinations of classifiers-tasks are most sensitive to the strategies employed in preprocessing.

5.14.1 Accuracy Comparison Across Classification Tasks

Holdout accuracy was compared under both preprocessing conditions for all twelve models across the three binary classification tasks. The model performance comparison such as this stems the quantitative evidence for preprocessing task-specific benefits and discloses different CLAHE effectiveness based on complexity of classification.

Figure 5.6, there is juxtaposition between the holdout overall accuracy of the algorithms with (brown bars) and without (pink bars) CLAHE preprocessing, which is also reflected across the three binary classification problems. Benign malignancy classification (left) is the only one of the three wherein CLAHE improved accuracy on 11 of 12 models, the most significant being Random Forest, which gained 5.7 percentage points, from 0.880 to 0.937. Malignant-normal classification (center) showed very close-to-calculate performance in both conditions, most models exceed 0.95 in accuracy, and hence the preprocessing effect is slight because baseline performance is already very high. Benign-malignant classification (right panel) has effects of preprocessing that were both negative and positive, as some models improved slightly whereas Random Forest went down by -2.2%, suggesting that there are instances when CLAHE would be useful, which can be task-dependent.

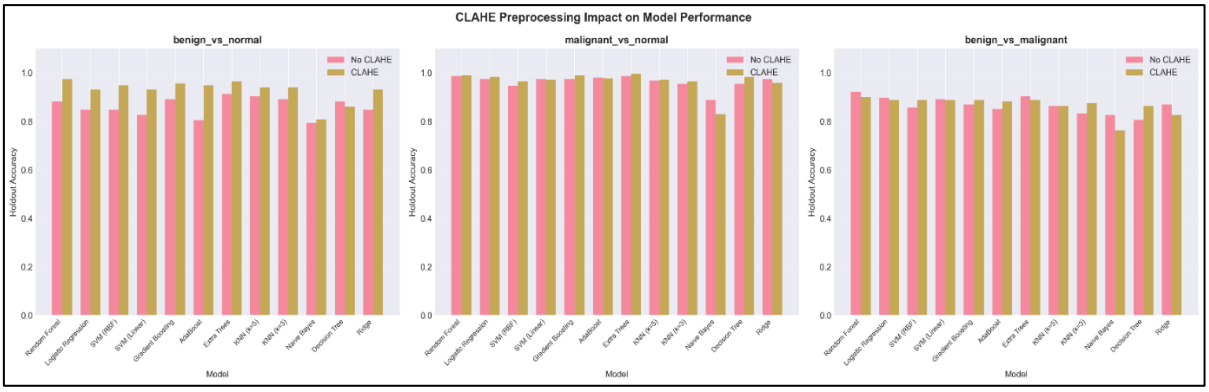


Figure 5.6: CLAHE Preprocessing Impact on Holdout Accuracy Across Tasks

5.14.2 Performance Distribution Analysis Using Heatmaps

Its processing performance is thus arranged to be evaluated from all dimensions in the experimental matrix through color-coded heatmaps for both preprocessing conditions, with dark green representative of high accuracy values themselves. This representation rapidly locates the model-task combinations that benefit most from enhancement.

Figure 5.7 presents a side-by-side comparison of performance distribution under CLAHE preprocessing (left panel) and without CLAHE (right panel). Under CLAHE preprocessing, Clear task-dependent stratification was observed after CLAHE preprocessing: malignant-normal classification scored uniformly high accuracy (0.93-0.99), while benign-normal classification was more variable (0.76-0.97), and benign-malignant classification was moderately accurate (0.76-0.90). Among the various classification methods, tree-based ensembles (Random Forest, Extra Trees, Gradient Boosting) have consistently outperformed all others on all tasks, while Naive Bayes showed poor discrimination.

Comparison between the two panels shows that benign-normal classification benefited most substantially from CLAHE, particularly for ensemble methods. In contrast, malignant-normal classification maintained high performance regardless of preprocessing, suggesting performance saturation. Benign-malignant classification showed mixed preprocessing effects, with some models achieving better accuracy without CLAHE, indicating that contrast enhancement may not universally benefit differential diagnosis tasks.

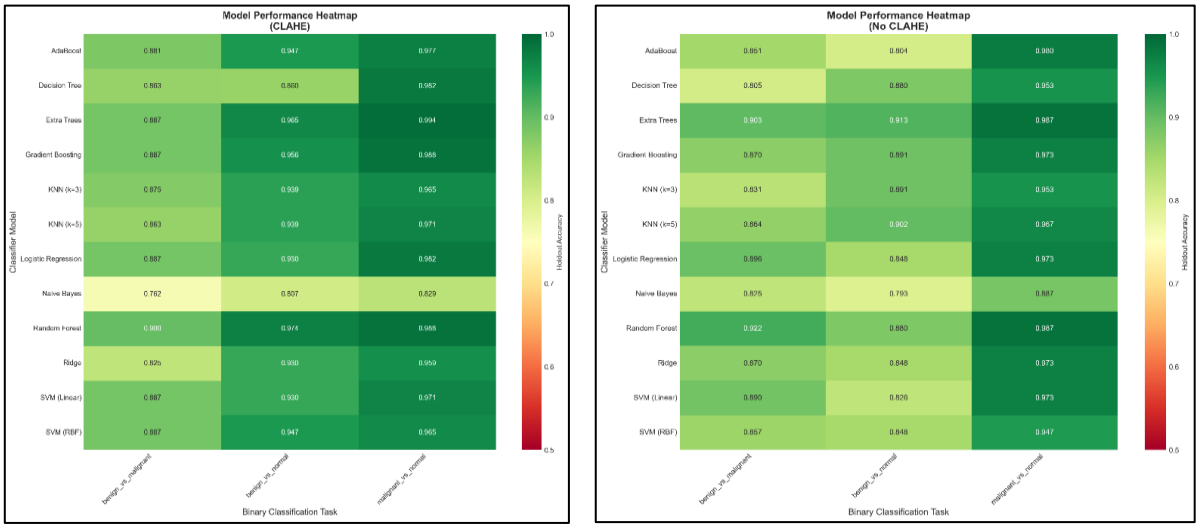


Figure 5.7: Comparative Model Performance Heatmaps: CLAHE Preprocessing (left) vs. Standard Preprocessing (right).

According to these visualizations, the preprocessing strategy does not apply uniformly but should align with the task, in which case CLAHE consistently benefits tasks regarding detection, but its use is limited in differential diagnosis.

This chapter presented the experimental results for liver lesion classification across three binary tasks: benign–normal, malignant–normal, and benign–malignant. It reported classifier performance under standard preprocessing and CLAHE-based contrast enhancement using cross-validation, test set evaluation, and holdout validation. The results showed that ensemble-based models consistently achieved the most stable and accurate performance, particularly for malignant–normal and benign–normal classification, while benign–malignant classification remained the most challenging task. Overall, the chapter demonstrated the task-dependent impact of CLAHE and provided a detailed basis for interpretation in the discussion chapter.

6 Chapter Six

Discussion and Conclusion

This work assessed the functioning of classical machine-learning methods for the classification of liver lesions in ultrasound images regarding three clinically of interest tasks: benign versus normal liver tissue; malignant versus normal tissue; and benign versus malignant lesion differentiation. The analysis included cross-validation testing, independent testing, and some holdout evaluation, together with ROC analysis and qualitative inspection of contrast enhancement effects. Together, these results provide pertinent knowledge about the strengths and weaknesses of machine-learning-based liver ultrasound analysis and help explain the observed patterns of performance from clinical and signal-processing perspectives.

6.1 Relative Difficulty of the Classification Tasks

Throughout all stages of evaluation, an unchanging array of task difficulty has been observed. The malignant-normal task has achieved the highest accuracy and stability; next, benign-normal classification, and finally, benign versus malignant was the most difficult.

This behavior nicely follows the known ultrasound imaging characteristics. Compared to the normal liver parenchyma, malignant liver lesions typically show significant internal heterogeneity, irregular margins, and disrupted echotexture. The visual differences lead to strong intensity and texture contrasts, easily captured by ML algorithms. On the other hand, benign lesions are more inclined to show smooth margins and a rather heterogeneous echogenicity, which thus overlaps tremendously with that of normal tissue and low-grade malignant lesions. Such overlapping creates severe limitations in terms of separability while using grayscale ultrasound alone.

Similar observations were made in the previous ultrasound studies. As reported by Wei et al. 2024, malignant-normal classification accuracies exceed 95 percent, whereas benign-malignant classification is usually lower except for contrast-enhanced ultrasound or advanced feature representations. Therefore, the current results reflect known diagnostic challenges rather than ones specific to the dataset.

6.2 Strength of Tree-Based Ensemble Models

When compared to all other methods across several tasks and evaluation splits, tree-based ensemble methods always gave birth to the best and the most reliable result across all tasks and evaluation splits; Random Forest, Extra Trees, Gradient Boosting. These models achieved accuracy with a fair balance between sensitivity and specificity, as well as very large AUC values, especially in malignant-normal and benign-normal comparisons.
The reason why ensemble models are very successful is that they are capable of modeling the nonlinear relationships and interactions amongst the features. Ultrasound images are characterized by various kinds of noise, which include speckle, local intensity fluctuations, and nonuniform texture patterns. Decision tree ensembles are well suited to this situation since they perform adaptive partitioning of feature space and combine multiple decision rules, hence reducing variance and improving robustness.

Extra Trees are often similar, sometimes even better than Random Forest performance, especially during holdout evaluations. The extra randomness introduced by randomizing during split selection seems to decrease sensitivity to minor intensity variation, which is common in ultrasound images. Advantages of ensemble methods, like that mentioned above, have also been observed in studies using ultrasound imagery of the liver and breast in studies (Md et al., 2023; Zhao et al., 2023).

6.3 Influence of CLAHE Preprocessing

CLAHE had an effect on classifier behavior, which varied according to the classification tasks. In general, across cross-validation, testing, and holdout sets, CLAHE indeed improved the malignant-normal and benign-normal sensitivity and AUC. Qualitative inspection showed clearer lesion boundary, enhancing internal texture variation and contrast between the lesions and surrounding parenchyma.

By means of the processing, CLAHE is avoided in overstaining in giving more contrast locally. This is relatively good with ultrasound-diagnostically relevant patterns are obscured by low dynamic range and speckle noise. By enhancing local intensity distribution, CLAHE permits classifiers to more effectively exploit texture and edge-related cues.

These outcomes agree with those from previous studies (R. R. Priyah & Kamalakkannan, 2025), which reported improved lesion detectability following adaptive histogram equalization in ultrasound images. Additionally, (Qasrawi et al., 2025) demonstrated that CLAHE-enhanced liver ultrasound images led to higher sensitivity for malignant lesion detection using ML approaches.

It was variable in the benign-malignant classification with respect to the benefits of CLAHE. Sensitivity usually improved without a similar degree of increase in specificity but occasionally declined. This suggests that enhancement in contrast may enhance common textural characteristics between benign and malignant lesions and increases false-positive rates. Analogous effects have been described in the breast and thyroid ultrasound studies. Contrast enhancement improved detection but reduced the fine-grained differentiation of lesions(P. Singh et al., 2020b).

6.4 Performance of Linear, Kernel-Based, and Probabilistic Models

Their performances, being relatively stable but limited, ran a lot more inferiority compared with ensemble methods. This means that with rather good results in malignant versus normal classification, it can be said that the task is somewhat linearly separable, but the limitations of linear methods are exposed in benign versus malignant discrimination, which is dominated by nonlinear texture patterns.

Some situations, the sensitivity of SVM with RBF kernel was enhanced for the serene-malignant cut, but often at the expense of specificity. Such a sensitivity-oriented behavior suggests a tendency of classifying an ambiguous case as malignant, a situation undesirable in clinical settings, where the detrimental consequences of treating false positives are unresolved follow-ups. Similar trade-offs have been reported in ultrasound-based SVM studies (Virmani et al., 2013).

Gaussian Naive Bayes has generally been the weakest performer among all tasks and all possible preprocessing strategies. This is because the assumption of independence of features does not deal well with image-derived features, which are highly correlated. As a result, the performance is very sensitive but very specific-less pattern which has been widely documented in the medical imaging literature (P. Singh et al., 2020b).

6.5 Generalization and the Role of Holdout Validation

The independent holdout set further secured the findings. The performance pattern observed during cross-validation and testing remained largely intact in holdout evaluations, especially

with respect to the ensemble-based classifiers. For malignant–normal classification, near-perfect generalization suggests tremendous robustness for the task.

The performance of the benign-malignant holdout, on the other hand, shows a very clear drop as compared to test results. Given that this raises the danger of being overoptimistic from merely performing cross-validation, it will again emphasize the need for independent validation, which is often discussed.

6.6 Relation to Deep Learning Studies

In recent years, due to deep learning, performance in differentiating benign from malignant lesions has generally improved in liver ultrasound, especially with methods relying on convolutional neural networks and transfer-learning (Kim et al., 2021; Virmani et al., 2013).

Many of these studies rely on large datasets, extensive augmentation, or pretraining on natural image databases. In contrast, this study shows that carefully designed classical ML pipelines, supported by appropriate preprocessing and rigorous evaluation, can achieve competitive performance for malignant–normal and benign–normal discrimination. In this work, ground-truth labels for the local clinical cases were based on confirmed clinical diagnosis supported by radiology reports, while labeled public cases were incorporated from an open-source dataset. Model performance was then validated using stratified cross-validation, independent testing, and a fully isolated holdout set to assess generalization.

Future improvements in benign–malignant classification may require richer feature representations, such as deep-learned features, multi-scale texture descriptors, or multimodal integration (e.g., Doppler ultrasound, elastography, and clinical metadata). These directions could help address the long-standing challenge of distinguishing visually overlapping lesion types in grayscale ultrasound.

6.7 Clinical and Methodological Implications

From a clinical perspective, the results indicate that ultrasound-based ML systems are highly reliable for distinguishing malignant lesions from normal liver tissue and may support screening and triage workflows. However, robustness was lower for benign–malignant classification, which remained the most challenging task across evaluation stages. This limitation is expected due to the substantial visual and textural overlap between benign and malignant lesions in grayscale ultrasound, making this distinction less consistent than malignant–normal discrimination. Therefore, while the results support potential clinical utility, benign–malignant differentiation should be interpreted with caution and may require richer feature representations

or multimodal information to improve reliability. Ensemble-based models appear particularly suitable due to their balance of performance and robustness.

From a methodological perspective, task-specific preprocessing and validation strategies are essential in the present findings. The variable effect of CLAHE shows that contrast enhancement should thus be applied differentially according to the diagnostic purpose, rather than uniformly applied.

6.8 Conclusion

With major emphasis on clinical practice, the work elaborated on the application of some classical ML techniques for the classification of liver lesions in ultrasound images under three separate diagnostic scenarios: benign lesions versus normal liver tissue; malignant lesions versus normal liver tissue; benign versus malignant lesions. The study took into account preparation of data, several validation stages, and qualitative visual analysis in order to gain a better understanding of how these models behave and how generalizable they are to unseen data.

The results indicate the potential of ML algorithms in liver ultrasound applications, especially in distinguishing malignant lesions from normal liver parenchyma. Among various algorithms, tree ensemble algorithms, including Random Forest, Extra Trees, and Gradient Boosting, produced the most reliable and stable results irrespective of cross-validation, test, or holdout datasets. Their noise tolerance and capacity to capture highly complex texture patterns become highly suitable for ultrasound images, which are usually disturbed by speckle and random local intensity variations.

Image preprocessing has also played a crucial role. Contrast enhancement by CLAHE augmented the visibility of the lesion, allowing several models to gain higher sensitivity and better class separation of malignant–normal and benign–normal classification. The visual examples showed that after contrast enhancement, the lesion boundaries appeared sharper and the internal texture was more accentuated. However, the positive effects of CLAHE were not uniformly observed across all tasks; in benign–malignant classification, for instance, contrast enhancement led occasionally to higher sensitivity but did not consistently lead to higher specificity. This reflects the strong visual overlap between benign and malignant lesions on the grayscale ultrasound.

The most difficult task among the three was to differentiate benign versus malignant lesions. Even strong classifiers coupled with perfect preprocessing could not overcome the limitations imposed by their common textual and morphological characteristics. Such setbacks show an intrinsic limitation of conventional ultrasound and suggest that further improvement may be workable only with the assistance of other sources of information, that is, advanced feature representations, temporal data, or complementary imaging modalities.

However, while cross-validation set down useful cues for model development, the results on the holdout gave the clearer picture with regard to how models perform on truly unseen data. This was particularly important for the case of the more difficult benign-malignant task, where performance dropped further when tested under stricter conditions.

In a clinical context, ultrasound imaging-based ML models show great promise in assisting the detection of malignant liver lesions and differentiating abnormal tissue from normal liver parenchyma. Such tools can be useful in screening and triage settings, especially in the absence of access to advanced imaging modalities. However, the findings also indicate that benign-malignant differentiation remains a complex problem and needs to be approached carefully prior to clinical implementation.

This research investigation should be enhanced in future work incorporating three-dimensional and longitudinal ultrasound data for hybridization with hand-crafted data and deep-learned features, including irregular parameters in terms of additional clinical or imaging parameters such as Doppler or elastography. Critical translation of these methodologies into routine practice would then require validation on much larger and more diverse datasets and prospective clinical studies.

In general, this thesis demonstrates that well-crafted classical machine-learning pipelines, underpinned by the right preprocessing and an appropriate validation scheme, could achieve strong performance when applied to the analysis of liver ultrasounds. The results provide insights into what can practically be achieved with existing techniques, while also pointing to clear directions for future improvements.

6.9 Future Works

Future research can build on the findings of this thesis to further improve ultrasound-based liver lesion classification and enhance clinical applicability. First, future work should incorporate larger multi-center datasets collected from diverse ultrasound machines and patient populations to improve robustness and reduce the risk of center-specific bias. Second, patient-level splitting and evaluation should be emphasized whenever patient identifiers are available, in order to ensure strict separation between training and testing data and minimize leakage risk in frame-based datasets. Third, future studies may explore advanced feature representations, including hybrid approaches that combine handcrafted radiomic features with deep feature extraction, to better capture subtle differences between benign and malignant lesions. Fourth, integrating additional ultrasound modalities, such as Doppler imaging or elastography, may improve lesion characterization beyond grayscale B-mode imaging alone. Finally, prospective clinical validation and workflow-oriented evaluation are recommended to assess model reliability in real-world settings and determine its potential role as a decision-support tool for radiologists.

This chapter discussed the key findings of the thesis in relation to the study objectives and existing literature. It explained the relative difficulty of the evaluated classification tasks and highlighted the strong and consistent performance of tree-based ensemble models. The chapter also interpreted the role of CLAHE preprocessing and its task-dependent benefits, and it emphasized the importance of holdout validation for assessing real-world generalization. Finally, it summarized the main conclusions, limitations, and recommendations for future research in ultrasound-based liver lesion classification.

References

1. Abdullah, M. O., Altun, Y., Ahmed, R. M. & Abdullah Sr, M. O. (2024). Leveraging Artificial Neural Networks and Support Vector Machines for Accurate Classification of Breast Tumors in Ultrasound Images. Cureus, 16(11).
2. Acharya, U. R., Koh, J. E. W., Hagiwara, Y., Tan, J. H., Gertych, A., Vijayananthan, A., Yaakup, N. A., Abdullah, B. J. J., Fabell, M. K. B. M. & Yeong, C. H. (2018). Automated diagnosis of focal liver lesions using bidirectional empirical mode decomposition features. Computers in Biology and Medicine, 94, 11–18.
3. Alivar, A., Daniali, H. & Helfroush, M. S. (2014). Classification of liver diseases using ultrasound images based on feature combination. 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE), 669–672.
4. Al-Karawi, D. (2019). Texture Analysis based Machine Learning Algorithms For Ultrasound Ovarian Tumour Image Classification within Clinical Practices.
5. Barnard Giustini, A., Ioannou, G. N., Sirlin, C. & Loomba, R. (2023). Available modalities for screening and imaging diagnosis of hepatocellular carcinoma—Current gaps and challenges. Alimentary Pharmacology & Therapeutics, 57(10), 1056–1065.
6. Bebis, G., Ghiasi, G., Fang, Y., Sharf, A., Dong, Y., Weaver, C., Leo, Z., LaViola Jr, J. J. & Kohli, L. (2023). Advances in Visual Computing: 18th International Symposium, ISVC 2023, Lake Tahoe, NV, USA, October 16–18, 2023, Proceedings, Part I (Vol. 14361). Springer Nature.
7. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I. & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263.
8. Chen, J., Wen, Z., Yang, X., Jia, J., Zhang, X., Pian, L. & Zhao, P. (2024). Ultrasound-Based radiomics for the classification of Henoch-Schönlein Purpura nephritis in children. Ultrasonic Imaging, 46(2), 110–120.
9. Chou, R., Cuevas, C., Fu, R., Devine, B., Wasson, N., Ginsburg, A., Zakher, B., Pappas, M., Graham, E. & Sullivan, S. D. (2015). Imaging techniques for the diagnosis of

hepatocellular carcinoma: a systematic review and meta-analysis. Annals of Internal Medicine, 162(10), 697–711.

10. Dana, J., Agnus, V., Ouhmich, F. & Gallix, B. (2020). Multimodality imaging and artificial intelligence for tumor characterization: current status and future perspective. Seminars in Nuclear Medicine, 50(6), 541–548.
11. der Medizinischen Fakultät, V., Baues Sc, M. M., Zandvoort MAMJ, van & Cpm, R. (2018). Modulating and monitoring the microenvironment in cancer and renal fibrosis. In Journal of Controlled Release (Vol. 282).
12. Flannery, M. M. (2020). Development of a Tissue Factor-Targeted Ultrasound Microbubble for Early Detection of Ovarian Cancer. The Ohio State University.
13. Fleury, E. & Marcomini, K. (2019). Performance of machine learning software to classify breast lesions using BI-RADS radiomic features on ultrasound images. European Radiology Experimental, 3(1), 34.
14. Forner, A., Reig, M. & Bruix, J. (2018). Hepatocellular carcinoma. In The Lancet (Vol. 391, Issue 10127, pp. 1301–1314). Lancet Publishing Group. https://doi.org/10.1016/S0140-6736(18)30010-2
15. García Ocaña, M. I. (2019). Analysis of ultrasound images and clinical data in two clinical scenarios: prediction of failure of induction of labor and risk of preterm delivery.
16. González-Luna, F. A., Hernández-López, J. & Gomez-Flores, W. (2019). A performance evaluation of machine learning techniques for breast ultrasound classification. 2019 16th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), 1–5.
17. Harikumar, R. & Karthikamani, R. (2021). Performance Analysis of classifiers in Detection of Abnormalities from Ultrasonic Images-A Review. IOP Conference Series: Materials Science and Engineering, 1084(1), 012030.
18. Imran, M. A., Ghannam, R., Abbasi, Q. H. & Wiley, J. (2021). Engineering and technology for healthcare. Wiley Online Library.
19. James, D. N., Sirsi, S., Brown, K., Tamil, L. & de Gracia Lux, C. (2023). DEVELOPING A PRESSURE ESTIMATION SYSTEM USING PHASE-CHANGE CONTRAST AGENTS AS PRESSURE SENSORS.

20. Jangra, A., Narang, M., Tiwari, S., Gupta, V. & Naqvi, N. (2022). Comparative Analysis of Machine Learning Techniques for Breast Tumor Classification on Balanced and Unbalanced Datasets. Proceedings of the International Conference on Innovative Computing & Communication (ICICC).
21. Jha, S. (2021). Modeling of Driver Attention in Real World Scenarios Using Probabilistic Salient Maps. The University of Texas at Dallas.
22. Kazi, I. A., Jahagirdar, V., Kabir, B. W., Syed, A. K., Kabir, A. W. & Perisetti, A. (2024). Role of imaging in screening for hepatocellular carcinoma. Cancers, 16(19), 3400.
23. Kim, T., Lee, D. H., Park, E.-K. & Choi, S. (2021). Deep learning techniques for fatty liver using multi-view ultrasound images scanned by different scanners: Development and validation study. JMIR Medical Informatics, 9(11), e30066.
24. Kirk, G. D., Nsibirwa, S., Aizire, J. K., Assefa, R. L., Bandala-Jacques, A., Orem, J., Maponga, T., Seydi, M., Wandeler, G. & Mohareb, A. (2025). Pragmatic Approach to Hepatocellular Carcinoma Diagnosis in High-Incidence, Resource-Limited Settings in Africa. JCO Global Oncology, 11, e2400592.
25. Koseoglu, F. D., Alıcı, I. O. & Er, O. (2023). Machine learning approaches in the interpretation of endobronchial ultrasound images: a comparative analysis. Surgical Endoscopy, 37(12), 9339–9346.
26. Llovet, J. M., Kelley, R. K., Villanueva, A., Singal, A. G., Pikarsky, E., Roayaie, S., Lencioni, R., Koike, K., Zucman-Rossi, J. & Finn, R. S. (2021). Hepatocellular carcinoma. In Nature Reviews Disease Primers (Vol. 7, Issue 1). Nature Research. https://doi.org/10.1038/s41572-020-00240-3
27. M Kumar. (2022). Machine Learning Algorithms, Models and Applications Edited by Jaydip Sen.
28. Martínez-Más, J., Bueno-Crespo, A., Khazendar, S., Remezal-Solano, M., Martínez-Cendán, J.-P., Jassim, S., Du, H., Al Assam, H., Bourne, T. & Timmerman, D. (2019). Evaluation of machine learning methods with Fourier Transform features for classifying ovarian tumors based on ultrasound images. PLoS One, 14(7), e0219388.
29. Maryam, A.-H., Sultan, L. R., Sagreiya, H., Cary, T. W., Karmacharya, M. B. & Sehgal, C. M. (2022). Machine learning improves early detection of liver fibrosis by quantitative ultrasound radiomics. 2022 IEEE International Ultrasonics Symposium (IUS), 1–4.

30. Md, A. Q., Kulkarni, S., Joshua, C. J., Vaichole, T., Mohan, S. & Iwendi, C. (2023). Enhanced preprocessing approach using ensemble machine learning algorithms for detecting liver disease. Biomedicines, 11(2), 581.
31. Mishra, A. K., Roy, P., Bandyopadhyay, S. & Das, S. K. (2021). Breast ultrasound tumour classification: A Machine Learning—Radiomics based approach. Expert Systems, 38(7), e12713.
32. Moghassemi, H., Rabiei, R., Moghaddasi, H., Shabani, M. & Ataee, R. (2018). Computer-Aided Diagnosis Approaches to Fatty Liver Disease According to Sonographic Images Based on Wavelet Transform: A Review Study. Journal of Clinical Review & Case Reports.
33. Mohammed, A. B., Garelnabi, M., Alamin, A. & Ali, M. A. M. A. (2023). Characterization of Primary and Malignant Liver Lesions using Texture Analysis. International Journal of Biomedicine. https://doi.org/10.21103/Article13(1)_OA15
34. Nieniewski, M. & Chmielewski, L. J. (2020). Study of classification of breast lesions using texture GLCM features obtained from the raw ultrasound signal. Image Analysis and Stereology, 39(2), 129–145.
35. Priyah, R. & Kamalakkannan, S. (2025). Hybrid contrast-limited adaptive histogram equalization and Deep Learning techniques for improving liver tumor detection. Future Technology, 4(3), 67–75. https://doi.org/10.55670/fpll.futech.4.3.7
36. Priyah, R. R. & Kamalakkannan, S. (2025). Hybrid contrast-limited adaptive histogram equalization and Deep Learning techniques for improving liver tumor detection. Future Technology, 4(3), 67–75.
37. Putra, R. A., Pradipta, G. A. & Ayu, P. D. W. (2025). Gallbladder Disease Classification from Ultrasound Images Using CNN Feature Extraction and Machine Learning Optimization. Journal of Electronics, Electromedical Engineering, and Medical Informatics, 7(4), 1089–1111.
38. Qasem, M. E. A. A. (2018). Right Ventricular Structure and Function in Elite Athletes in Relation to Pre-Participation Screening. Liverpool John Moores University (United Kingdom).
39. Qasrawi, R., Daraghmeh, O., Thwib, S., Qdaih, I., Issa, G., Polo, S. V., Owienah, H., Al-Halawa, D. A. & Atari, S. (2025). Advancing breast cancer detection in ultrasound images

using a novel hybrid ensemble deep learning model. Intelligence-Based Medicine, 11, 100222. https://doi.org/https://doi.org/10.1016/j.ibmed.2025.100222

40. Qurashi, M., von Wagner, C. & Sharma, R. (2024). Optimising Surveillance in Hepatocellular Carcinoma: Patient-Defined Obstacles and Solutions. Journal of Hepatocellular Carcinoma, 1597–1605.
41. Renard, F., Guedria, S., Palma, N. De & Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. Scientific Reports, 10(1), 13724.
42. Roman, R. A. & Detti, L. (2019). Ultrasound evaluation of uterine cavity remodeling after in vitro fertilization. https://doi.org/10.13140/RG.2.2.21020.64641
43. Ronot, M., Pommier, R., Dioguardi Burgio, M., Purcell, Y., Nahon, P. & Vilgrain, V. (2018). Hepatocellular carcinoma surveillance with ultrasound—cost-effectiveness, high-risk populations, uptake. The British Journal of Radiology, 91(1090), 20170436.
44. Sathish Kumar, N., Kasiselvanathan, M. & Vimal, S. P. (2020). Performance comparison of various machine learning algorithms for ultrasonic fetal image classification problem. In Advances in Smart System Technologies: Select Proceedings of ICFSST 2019 (pp. 61–69). Springer.
45. Shen, H., Lv, G., Lin, H., Huang, N., Wu, Y., Cheng, H. & Yang, S. (2020). Development of an ultrasound prediction model to discriminate between malignant and benign liver lesions. Ultrasound in Medicine & Biology, 46(4), 952–958.
46. Simkó, A., Garpebring, A., Jonsson, J., Nyholm, T. & Löfstedt, T. (2024). Reproducibility of the Methods in Medical Imaging with Deep Learning. Medical Imaging with Deep Learning, 95–106.
47. Singh, A. P., Moud, R., Hariyani, D. J., Devi, P. & Mod, P. (2025). Machine learning based dysfunction of thyroid disease. IET Conference Proceedings CP920, 2025(7), 153–161.
48. Singh, N. & Jindal, A. (2012). A segmentation method and comparison of classification methods for thyroid ultrasound images. International Journal of Computer Applications, 50(11), 43–49.
49. Singh, P., Mukundan, R. & De Ryke, R. (2020a). Feature enhancement in medical ultrasound videos using contrast-limited adaptive histogram equalization. Journal of Digital Imaging, 33(1), 273–285.

50. Singh, P., Mukundan, R. & De Ryke, R. (2020b). Feature enhancement in medical ultrasound videos using contrast-limited adaptive histogram equalization. Journal of Digital Imaging, 33(1), 273–285.
51. Suganya, Y., Ganesan, S. & Valarmathi, P. (2022). Comparative analysis of ovarian images classification for identification of cyst using ensemble method machine learning approach. ECS Transactions, 107(1), 7407.
52. Vetter, M., Waldner, M. J., Zundler, S., Klett, D., Bocklitz, T., Neurath, M. F., Adler, W. & Jesper, D. (2023). Artificial intelligence for the classification of focal liver lesions in ultrasound - a systematic review. In Ultraschall in der Medizin (Vol. 44, Issue 4, pp. 395–407). Georg Thieme Verlag. https://doi.org/10.1055/a-2066-9372
53. Virmani, J., Kumar, V., Kalra, N. & Khandelwal, N. (2013). SVM-based characterization of liver ultrasound images using wavelet packet texture descriptors. Journal of Digital Imaging, 26(3), 530–543.
54. Wahdan, P., Saad, A. & Shoukry, A. (2014). Comparing classification techniques to detect breast tumor. Proceedings of the International Conference on Biomedical Engineering and Systems, 1–6.
55. Waibel, A. (2019). Multimodal dialogue processing for machine translation. In The Handbook of Multimodal-Multisensor Interfaces: Language Processing, Software, Commercialization, and Emerging Directions-Volume 3 (pp. 577–620).
56. Wan, K. W., Wong, C. H., Ip, H. F., Fan, D., Yuen, P. L., Fong, H. Y. & Ying, M. (2021). Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study. Quantitative Imaging in Medicine and Surgery, 11(4), 1381.
57. Yap, M. H., Edirisinghe, E. A. & Bez, H. E. (2009). A comparative study in ultrasound breast imaging classification. Medical Imaging 2009: Image Processing, 7259, 598–608.
58. Zhang. (2020). UNIVERSITY OF CALIFORNIA, IRVINE Machine Learning in Clinical Application of Medical Imaging for Lesion Detection.
59. Zhang, H., Wang, Y., Hu, B., Song, B., Wen, Z., Su, L., Chen, X., Wang, X., Zhou, P. & Zhong, X. (2024). Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases. Scientific Reports, 14(1), 28575.

60. Zhao, T., Zeng, Z., Li, T., Tao, W., Yu, X., Feng, T. & Bu, R. (2023). USC-ENet: a high-efficiency model for the diagnosis of liver tumors combining B-mode ultrasound and clinical data. Health Information Science and Systems, 11(1), 15.
61. Zheng, R., Guo, Q., Gao, C. & Yu, M. (2019). A hybrid contrast limited adaptive histogram equalization (clahe) for parathyroid ultrasonic image enhancement. 2019 Chinese Control Conference (CCC), 3577–3582.

دراسة مقارنة منهجية لخوارزميات التعلّم الآلي التقليدية لتصنيف آفات الكبد في صور الموجات فوق الصوتية

اعداد: عائشه لؤي ابراهيم انعيرات

المشرف: الأستاذ الدكتور رضوان القصراوي

الملخص

يُعدّ التصوير بالموجات فوق الصوتية للكبد من أكثر وسائل التصوير شيوعاً نظراً لتوفره وسلامته وانخفاض تكلفته، إلا أن التمييز بين الآفات الكبدية الحميدة والخبيثة لا يزال يمثل تحدياً بسبب الضجيج، وضعف التباين، والتداخل الكبير في الخصائص البصرية بين أنواع الآفات. تهدف هذه الرسالة إلى تقييم أداء وقدرة التعميم لتقنيات التعلم الآلي التقليدية في تصنيف آفات الكبد اعتماداً على صور الموجات فوق الصوتية ضمن ثلاث مهام تصنيف ثنائية ذات أهمية سريرية: (الحميد مقابل الطبيعي)، (الخبيث مقابل الطبيعي)، و(الحميد مقابل الخبيث).

تم دمج مجموعة بيانات سريرية محلية مع مجموعة بيانات عامة متاحة على منصةZenodo ، مما أسفر عن مجموعة أولية تضم 6,791 صورة موجات فوق صوتية. ولتقليل التكرار وضمان نزاهة البيانات، تم حذف الصور المكررة والمتشابهة بدرجة عالية، ليصبح عدد الصور الفريدة في المجموعة المنقّحة 2,387 صورة. كما تمت مقارنة المعالجة المسبقة الافتراضية مع تحسين التباين باستخدام تقنية المعادلة التكيفية للمدرج التكراري محدودة التباين(CLAHE) ، وتم اختبار مجموعة من المصنفات التقليدية، بما في ذلك نماذج التجميع(Ensemble) ، وآلات متجه الدعم، والنماذج الخطية، والمصنفات المعتمدة على الجوار، والنماذج الاحتمالية. وقد تم تقييم أداء النماذج باستخدام التحقق المتقاطع الطبقي، والاختبار المستقل، ومجموعة تحقق معزولة بالكامل (Holdout) لضمان تقييم أكثر دقة لقابلية التعميم.

أظهرت النتائج أن المصنفات المعتمدة على التجميع قدّمت أداءً أكثر ثباتاً وتفوقاً، لا سيما في مهمتي تصنيف (الخبيث مقابل الطبيعي) و(الحميد مقابل الطبيعي). في المقابل، كانت مهمة (الحميد مقابل الخبيث) الأكثر صعوبة بسبب التشابه البصري الكبير بين نوعي الآفات. كما ساهمت تقنية CLAHE في تحسين الحساسية وزيادة قابلية الفصل بين الآفات في عدد من النماذج، مع اختلاف الفائدة وفقاً لطبيعة المهمة التصنيفية.

بصورة عامة، تشير النتائج إلى أن خطوط المعالجة القائمة على التعلم الآلي التقليدي، عند دعمها بمعالجة مسبقة مناسبة وتحقق صارم، قد تكون فعّالة في دعم تصنيف آفات الكبد باستخدام صور الموجات فوق الصوتية، كما يمكن أن تشكل أساساً لأعمال مستقبلية تهدف إلى تطوير وتحسين هذه المنهجيات.

الكلمات المفتاحية: الموجات فوق الصوتية للكبد؛ تصنيف آفات الكبد؛ التعلّم الآلي التقليدي؛ المعادلة التكيفية للمدرج التكراري محدودة التباين(CLAHE) ؛ التعلّم التجميعي؛ آلات متجه الدعم(SVM) ؛ التحقق المتقاطع؛ التحقق باستخدام مجموعة معزولة (Holdout).