Classification of Fatty Liver Disease Using Deep Learning Technique

A thesis submitted in partial fulfillment of the requirements fort the Degree of

Master of Science in Biomedical Engineering

By:

Hume Degebassa Dida

Centre of Biomedical Engineering

Addis Ababa Institute of Technology

Addis Ababa University

Advisor: Dawit Assefa Haile (PhD)

Co-Advisor: Million Molla (MD, MPH)

August 2023

Addis Ababa, Ethiopia

Declaration

I, the undersigned, declares that this MSc thesis is my original work, has not been presented for fulfillment of a degree in this or any other University, and all sources and materials used for the thesis have been acknowledged.

Name: Hume Degebassa Dida

Signature: _____

Date: _____

This MSc. thesis has been submitted for examination with my approval as the thesis advisor.

_____

Dawit Assefa Haile (PhD)

Addis Ababa University

School of Graduate Studies

Certificate of Examination

This is to certify that the thesis prepared by Hume Degebassa Dida entitled *Classification and Grading of Fatty Liver Disease using Deep Learning Technique* submitted in partial fulfillment of the requirements for the degree of Master of Science in Biomedical Engineering (Bioinstrumentation and Imaging) complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the examining committee:

Examiner _____Signature _____ Date _____
(Internal)

Examiner _____Signature _____ Date _____
(External)

Advisor Dawit Assefa Haile (PhD) Signature _____ Date _____

_____

Associate Director, Post Graduate Programs

Acknowledgments

First, I would like to thank God. Without his assistance, this task would not have been possible. Then, my sincere gratitude goes to my advisor, Dr. Dawit Assefa Haile, for his valuable suggestions, directions, and encouraging guidance. I also want to express my sincere appreciation to my clinical co-advisor Dr. Million Molla from St. Peter Specialized Hospital for his consultation in clinical aspects of the thesis. Many thanks to the Center of Biomedical Engineering staff at the Addis Ababa Institute of Technology for their encouragement throughout this research. Thanks, should also be goes to my brother and friends for their helpful advices and feedback. Last but not least, my deepest gratitude goes to my family for their unconditional support in looking after my new born baby while I was working through this thesis. Thank you all for your support and encouragement.

Abstract

Fatty liver disease (FLD) also termed steatosis liver disease can be classified into two broad spectrums: Alcoholic fatty liver disease (AFLD) and Non-alcoholic liver disease (NAFLD). AFLD is a type of liver disease associated with alcohol intake of 40 to 80 grams of ethanol/day for males and of 20 to 40 grams/day for females. NAFLD is a disorder characterized by excess accumulation of fat (in the form of triglycerides) with an amount >5% in the hepatocytes. The progression of the disease is manifested by ongoing inflammation and consequent steatosis grade: G1 (mild steatosis), G2 (moderate steatosis), and G3 (severe steatosis). If no early strategies are adopted, FLD can progress to steatohepatitis (SH) which is a risk factor for liver fibrosis, cirrhosis, and hepatocellular carcinoma (liver cancer). Cirrhosis is currently the 11th most common cause of death globally and liver cancer is the 16th leading cause of death. In Ethiopia, cirrhosis was the 7th leading cause of mortality accounting for 24 deaths per 100,000 populations in 2019. Hence, early diagnosis is important to avoid advanced stages of FLD. In order to diagnose and grade FLDs, Ultrasound imaging is the most well-known technique. The visual assessment of these Ultrasound images is not only time taking and labor-intensive task but it is also prone to inter-observer variability. That calls for the development of an automated system that overcomes subjectivity and inconsistency in the screening process. Different studies in the literature have proposed the classification of FLDs using deep learning and machine learning techniques. However, most of these techniques are limited to binary classification and don't consider the severity level. The current study aims to develop an automatic fatty liver classification and grading scheme using deep learning techniques. A total of 550 liver ultrasound images were collected from Zenode repository and image pre-processing was employed on the acquired images before the images were used to train the deep learning model. The effect of important deep network hyper parameters including batch size, learning rate, regularizer and epoch was investigated. Different pre-trained deep learning networks were tested for their classification efficacy including VGGNet, MobileNetV2, ResNet and Xception based on useful performance matrices. Accordingly, the Xception model was found to outperform the rest for multiclass (normal, mild, moderate, and severe) fatty liver classification. It offered a precision of 96.25%, recall of 97.75%, F1 score of 96.75%, and overall accuracy of 96.36% showing the great promises of the model.

Keywords: Cirrhosis, HCC, liver steatosis; ultrasound image**,** FLD, Xception

Table of Contents

List of Figures

List of Tables

Acronym

AI	Artificial Intelligence
ANN	Artificial Neural Network
ADAM	Adaptive Moment Estimation
CLAHE	Contrast Limited Adaptive Histogram
CAD	Computer Aided Diagnostic
CNN	Convolutional Neural Network
FLD	Fatty Liver Disease
GLCM	Gray Level Concurrence Matrix
GUI	Graphical User Interface
HCC	Hepatocellular Carcinoma
HRI	Hepatorenal l Index
ROI	Region of Interest
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
US	Ultrasound

CHAPTER ONE

INTRODUCTION

1.1 Background

The liver is the heaviest internal organ and the largest gland in the human body consisting of the largest reticulon-endothelial cell network. Physically the liver is found above the stomach and intestines in the upper right-hand portion of the abdominal cavity, and below the diaphragm as shown in Figure 1.1. Its radish cone-shaped organ is approximately 15cm (6 inches) in width and weighs 1.5 Kg (3.3 lb) [1]. The liver is connected to two distinct sources of large blood vessels: the hepatic artery and the portal vein. The hepatic artery carries oxygen-rich blood from the aorta via the celiac trunk whereas the portal vein carries blood rich in digested nutrients from the entire gastrointestinal truck and also from the spleen and pancreas.

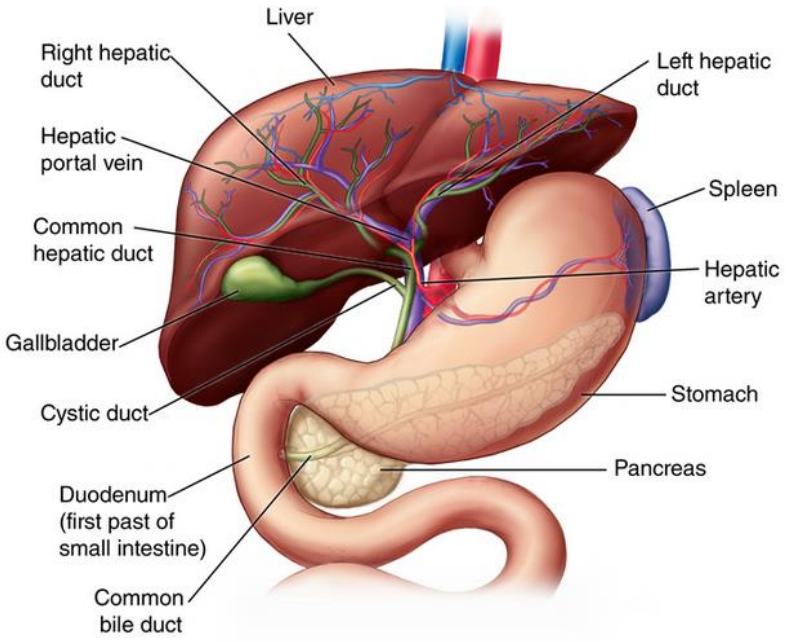


Figure: 1.1: Anatomy of the Liver [2].

The liver is responsible for separating more than 5,000 bodily functions including separation of nutrients and waste as they move through the digestive system. It is also the primary organ for maintaining the chemicals like glucose, fat, vitamins, cholesterol and hormones. Liver problems can be caused by a

variety of factors and the major ones identified worldwide are virus (hepatitis B, C, and D) infections, alcohol use, fatty liver disease (FLD) and drugs [3]. Fatty liver disease (FLD), also termed steatosis liver disease, can be classified into two broad spectrums: Alcoholic fatty liver disease (AFLD) and Non-alcoholic liver disease (NAFLD) [4]. AFLD is a type of liver disease associated with alcohol intake of 40 to 80 grams of ethanol/day by males and of 20 to 40 grams of ethanol/day for females [5]. NAFLD is defined as a condition caused by too much fat deposited in the liver and when the individual consumes little or no alcohol. The disorder is characterized by excess accumulation of fat (in the form of triglycerides) with amount >5% in the hepatocytes, resulting in being overweight, diabetes, and raised blood lipids including cholesterol, triglycerides and high blood pressure. The progression of the disease is manifested by ongoing inflammation and consequent steatosis grades: G1 (mild steatosis), G2 (moderate steatosis) and G3 (severe steatosis). If no treatment is adopted, FLD can progress to steatohepatitis (SH) which is a risk factor for liver fibrosis, cirrhosis, and hepatocellular carcinoma (liver cancer) as also indicated in Figure 1.2 [6]. It contributes to the most frequent cause of chronic liver diseases (CLD) and become one of the leading causes of cirrhosis.

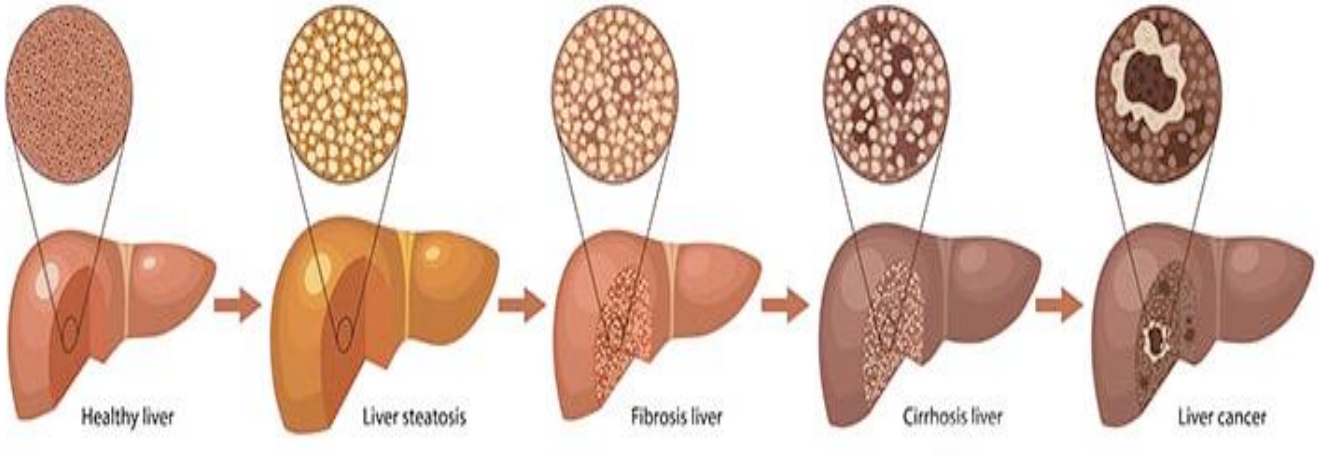


Figure 1.2: Different stages of liver damage [7].

CLD is a leading cause of morbidity and mortality worldwide, and poses a significant burden on health systems. It is a serious health issue that affects approximately 800 million peoples across the globe. FLD is a leading cause of CLD with a prevalence of 25% worldwide. They are associated with a high risk of hepatic cirrhosis and hepatocellular carcinoma [8]. Around 2 million people die from liver disease each year; 1 million of these deaths are related to cirrhosis complications, and 1 million to hepatocellular carcinoma [9]. Collectively Cirrhosis and Liver cancer accounts for 3.5%

of all deaths worldwide. Currently, cirrhosis is the 11th most common cause of death globally and liver cancer is the 16th leading cause of death [9] According to the 2019 Global Burden of Disease study, cirrhosis was responsible for 24 deaths per 100000 people in Ethiopia, ranking as the seventh most common cause of mortality in the country [3].

About 2 billion adults worldwide are overweight and over 400 million have diabetes; both of which are at higher risk of developing FLD [10]. It's also found to be a well-established risk factor for chronic kidney disease, and cardiovascular diseases (CVDs). Similarly, insulin resistance, type 2 diabetes Melitus (T2DM), and obesity are correlated with complications of FLD [11]. For these reasons, it is crucially important to obtain an early diagnosis in order to optimize the management of FLD.

In order to determine the diagnosis and grading of FLD, liver biopsy still remains to be the gold standard. However, it cannot be used frequently for evaluation and follow-up of FLD since it is an invasive procedure with a significant risk of serious complications. Non-invasive techniques for the diagnosis of FLD include ultrasonography (US), computed tomography (CT), magnetic resonance imaging (MRI), and magnetic resonance-based fat quantification techniques. Because it is widely accessible and inexpensive, conventional B-mode US is often used for FLD screening and initial assessment. However, on US, FLD is identified based on a few suggestive criteria, including hyperechoic structure (also known as "bright liver"), deep beam attenuation, liver-to-kidney contrast, and bright artery walls [6].

The traditional method of diagnosing FLDs is characterized by its subjectivity, labor intensiveness, complex sample preparation requirement, and need for invasive procedures. Effective as well as on-time diagnoses are highly important for better management of the disease. In that regard, the current research work aims to create an automated procedure that could be used to detect and correctly classify stages of liver steatosis developing an artificial intelligent (AI) system that takes useful liver bio-markers as input. The method that can process US images of different liver patients and classify them into four categories: normal steatosis, mild steatosis, moderate steatosis, and severe steatosis. This method provide a suitable knowledge to physicians in their decision-making while examining liver cases.

1.2 Fatty Liver

Fatty liver, commonly referred to as steatosis, is a prevalent condition characterized by the accumulation of fat that can range from benign to a more severe form that leads to liver inflammation and fibrosis.

Steatosis is a reversible form of liver disease, whereas disease symptoms are not noticeable until the disease progresses to a critical stage [12]. Therefore, an accurate grading of the fat in the liver is important for timely intervention and prevention of disease progression.

1.3 Grading of Fatty Liver

Clinically, the level of steatosis is graded in four different categories depending on the percentage of fat cells that make up the liver's overall weight. These are:

Grade-0 (healthy): Here the weight of the liver is made up of 0–5% fat cells. It refers to similar echogenicity with the right kidney characterized by the normal appearance of the diaphragm and portal vein [13].

Grade-1 (mild): This is the condition when the weight of the liver is made up of 5–33% fat cells. It is a slight increase in hepatic echogenicity accompanied by normal appearance of the diaphragm and periportal vessel [13].

Grade-2 (moderate): The liver weight is made up of 34-66% fat cells and is depicted as an increase in hepatic echogenicity with imperceptible periportal echogenicity but without obscuration of the diaphragm [13].

*Grade-3 (severe)***:** Fat cells comprise more than 66% of the liver's overall weight. It is depicted as an increase in hepatic echogenicity with both imperceptible periportal echogenicity and obscuration of the diaphragm with no visualization of the portal vein wall, diaphragm, and posterior part of the right liver lobe [13]. Figure 2.1 presents typical pictorial appearances of the liver in its different states as seen on an US scan.

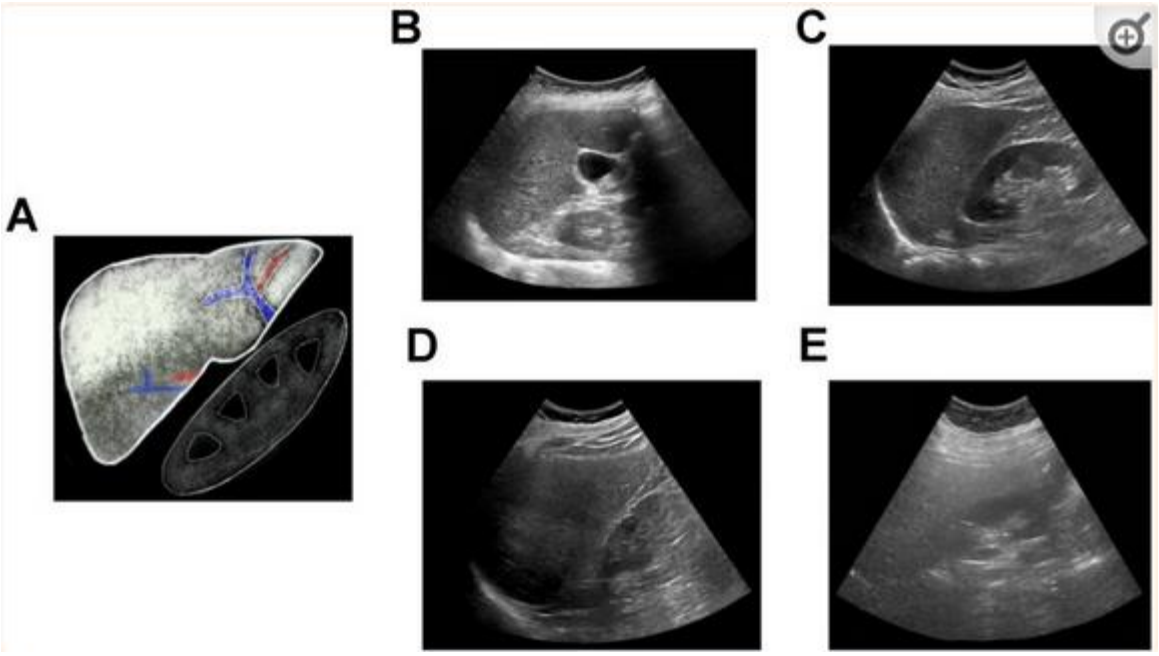


Figure 1.3: *Qualitative assessment of liver US image*: *(A) A schematic showing fatty liver increased echogenicity compared to the right kidney, blurring of intrahepatic vessels and posterior beam attenuation, (B) normal, (C) mild, (D) moderate, and (E) severe fatty liver* [14].

Accurate grading of the fat is crucial for patients with liver steatosis for tracking the progression of liver disease and treatment effect. A higher steatosis grade (2 or 3) can lower the chance of successful transplant. Increased steatosis is indicative of more aggressive disease progression in hepatitis C virus liver disease [15]. The level of liver steatosis is related to metabolic syndrome and cardiovascular risk [13].

The severity of steatosis is also strongly associated with SH and advanced fibrosis which are less irreversible forms of liver disease. Recent studies reported 25% of people with simple fatty liver will develop SH [16]. SH is a silent liver disease that can lead to liver cirrhosis and hepatocellular carcinoma (HCC). Approximately 30–50 % of SH patients may progress to cirrhosis within 10 years [17]. Cirrhosis is a severe form of liver disease in which the liver is permanently damaged and its tissue is replaced by scar tissue. Hence an accurate assessment of liver steatosis is essential for the management of FLD. Figure 2.2 demonstrate the different stage of FLD.

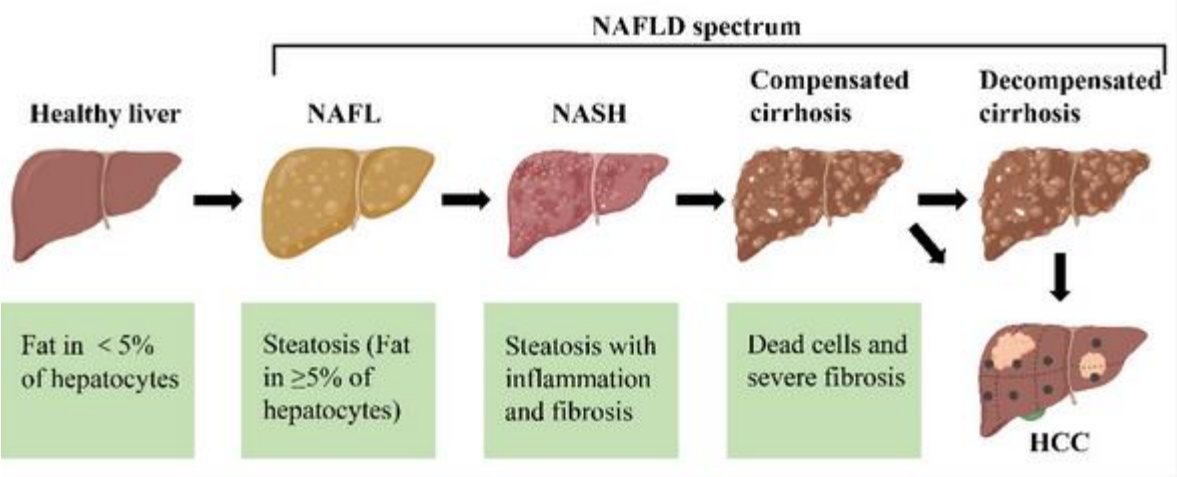


Figure 1.4: Different stages of FLD [18].

1.4 Fatty Liver Screening and Diagnosis Method

Different procedures are used to diagnose fat in the liver. These tests include blood tests, liver biopsy as well as tests used to visualize the appearance of the liver including ultrasound (US), computed tomography (CT) and magnetic resonance imaging (MRI) [19].

A. Blood Tests: Aspartate aminotransferase (AST), alkaline phosphatase (ALP), alanine aminotransferase (ALT), and gamma-glutamyl transpeptidase (GGT) levels are liver enzyme function testes included in blood test. Bilirubin, international normalized ratio (INR**)**, Blood urea nitrogen (BUN**)**, creatinine (Cr) and albumin are all biochemical markers that are initially useful in the diagnosing FLDs [20]. But they are all unspecific. The upper normal limit of these parameters varies with tests performed in different laboratories depending on technical reasons and different reference populations. Further, the test results are influenced by factors including gender, obesity, and the different components of metabolic syndrome.

Many predictive models have been developed to facilitate the diagnosis of FLDs. A method called Fatty liver index is capable of predicting liver steatosis based on body mass index (BMI), serum triglyceride, gamma-glutamyl transferase (GGT) levels, and waist circumference [21]. Another method, Steato Test, is a logistic regression model consisting of 12 indicating parameters: ALT, GGT levels, cholesterol, α2-macrogol-Bulin (A2M), total bilirubin, apolipoprotein A1 (ApoA1), haptoglobin, triglycerides, glucose, age, gender and BMI [22]. Another prediction method based on the laboratory test includes six parameters: ALT, white blood cell count, triglyceride, hemoglobin A1C (HbA1c), high-density

lipoprotein cholesterol and the presence of hypertension. This model is developed to screen FLDs in the general population [23]. However, the use of these current surrogate indicators has not been extensively studied and well documented. Further, the concentration of the liver enzyme can be normal in most of patients with FLD and correlate poorly with the histological severity [24].

B. Liver Biopsy

Liver biopsy is the gold standard for determining the amount of liver steatosis in hepatic tissue samples**.** Histological steatosis is graded on a semi-quantitative scale based on the number of hepatocytes containing microscopically discernible cytoplasmic fat droplets: 0 (<5% hepatocytes), 1 (5–33% hepatocytes), 2 (33–66% hepatocytes), and 3 (>66% hepatocytes) [29]. Techniques used for grading liver steatosis include the Knodell histological activity index (HAI), Scheuer, Ishak, and Metavir, systems [25].

Even though liver biopsy is the gold standard to diagnose and grading of steatosis it has many drawbacks. It is an expensive, invasive, and operator dependent procedure with potential complications such as pain, bleeding, infection, and even death [33]. It also depends on the subjective evaluation of the pathologist who counts the affected hepatocytes manually, instead of measuring the lipid volume in the sample. Moreover, liver biopsy can miss some areas of the liver due to the small size of the sample is analyzed [4]. As a result, liver biopsy is not considered to as the best and easiest technique to assess and track the progression of the liver disease. Therefore, Noninvasive alternatives for screening and detecting hepatic steatosis are highly needed.

C) Image based Visualization

Visualizing the appearance of the liver through imaging is another procedure used to diagnose fatty livers. US, CT and MRI are the commonly used imaging modalities in this regard. US refers to high-frequency sound waves that have frequencies ranging from 20 kHz to 10 GHz and propagate within a medium as pressure waves. In the medical context, US typically operates within the frequency range of 1 MHz to 30 MHz. US is the most preferred modality for screening steatosis because of it is radiation-free, relatively low cost and widely available. Radiologists study liver US images on the basis of echogenicity. If the echogenicity is uniform all across the image, then the image is assumed a normal liver and if there is a variation in echogenicity in the liver, then it is considered as fatty liver [26].

The US imaging modality has a low image quality compared to other modalities. The images are often affected by speckle noise, shading, blurring and other artifacts that can cause misdiagnosis in image interpretation [27]. The image quality depends on how well the speckle noise is reduced. The blurring and other artifacts also need to be addressed. Many methods have been proposed to reduce speckle noise in US and enhance visual clarity for better diagnoses. Nevertheless, in spite of the improved performance, these techniques still have a number of drawbacks because they are prone to image blurring and depend on the selected area.

A common method to reduce speckle noise in US images is to apply complex algorithms, statistical models and image processing techniques to measure the fat content in the liver more accurately. Some of the widely used imaging-based tools for liver steatosis screening are the hepatorenal index (HRI) and the gray-level co-occurrence matrix (GLCM) features [28]. The HRI is the ratio of the mean brightness level of the liver and the kidney cortex. It compares the liver echogenicity with that of the right kidney cortex. When steatosis is present, the liver tissue is brighter than the right kidney, while in normal conditions, they have similar echogenicity. Generally, the HRI tends to increase with the steatosis level. Figure 2.3 shows US images and the regions of interests (ROIs) used to compute the HRI by calculating the mean brightness of the liver and the right kidney. As could be seen on the figure, for the normal liver case where the steatosis level is low, the HRI is small. As the steatosis level increases, the HRI also behaves same way.

The GLCM technique is the most popular algorithm that makes the use of texture analysis for identifying liver steatosis [28]. GLCMs offer helpful information regarding the spatial dependencies of gray levels in the image. The texture patterns in US images are a result of backscattered US waves interfering with tissue microstructures. This textural variation in the image has been found as a useful feature to identify normal and fatty liver.

However, there are several limitations related to both the HRI and GLCM based methods. Notably, the operator has to choose ROIs in the liver and kidney that are as uniform as possible, excluding blood vessels, ducts and other structures. The ROI selection affects the calculation of the GLCM-based features and the HRI parameter. Therefore, these methods rely heavily on the skill of the physicians performing the examination and selecting the ROIs [29].

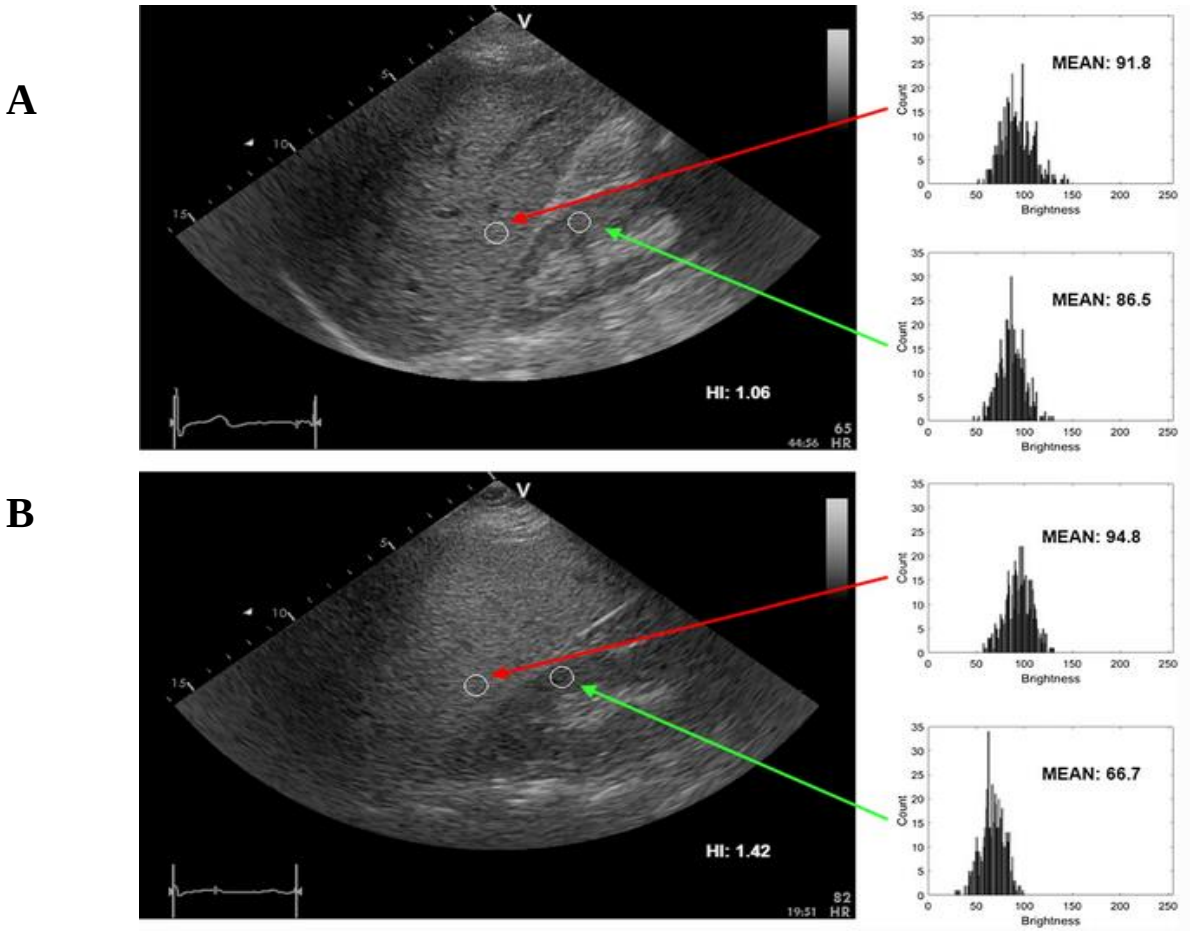


Figure 2.3: Liver US images and the ROIs for HRI calculation, (A) steatosis level of 3% and (B) steatosis level of 25% [30]

In addition to diagnosing steatosis, US can be used to grade the severity of steatosis by scoring the degree of liver brightening and/or blurring of vessels and diaphragm. It performs the best in grading of liver steatosis when there is no other related liver disease. However, it has a limitation in detecting mild steatosis. Using histology as the reference standard, US has a sensitivity and specificity of 80–89% and 87–90% for detecting mild to severe steatosis, respectively [31]. When all levels of FLD is considered the sensitivity and specificity drop to 65% and 81%, respectively [31]. The US imaging results are dependent on the time-gain settings of the machine and operator thus leading to failure in the identification of fatty and non-fatty liver images. Therefore, with rapid advancement of Computer Aided Diagnosis (CAD), it is critical to improve the capacity to interpret 2D imaging for the diagnosis of FLDs.

Another US-based method is called Transient Elastography (TE), which uses the so-called Controlled Attenuation Parameter (CAP) to simultaneously quantify liver stiffness and detect steatosis. According

to a previous study, CAP successfully identified steatosis with area under receiver operating characteristic (AUROC) values of 0.80 for mild, 0.86 for moderate, and 0.88 for severe steatosis. Poorly defined cut-off values for hepatic steatosis classification is the major limitation of CAP. Furthermore, there is the possibility of variations in skin-to-capsule ratio that could impair the method's ability to estimate the amount of hepatic steatosis [32]

CT can measure and assess moderate to severe steatosis, but it is not accurate for mild steatosis [40]. Hepatic steatosis can be detected by unenhanced and contrast-enhanced CT when the liver's absolute attenuation is below 40 Hounsfield units (HU), especially when the liver's fat content is more than 30% [41]. Unenhanced CT is more accurate than enhanced CT for evaluating hepatic steatosis because the contrast enhancement in the liver and spleen differs in the kinetics. However, CT is not often used as the main method to diagnosis hepatic steatosis. This is because CT HU values depend on tissue density. Small amounts of hepatic fat may have a minor effect on attenuation, making it invisible by CT. Also, attenuation values are influenced by high body mass index (BMI), hepatic iron accumulation and the existence of underlying fibrosis while scanning factors like the voltage, tube current, and calibration of HU vary based on the type of scanner and the manufacturer [33]. Furthermore, the use of CT for continuous FLD monitoring is hampered by concerns about repeated exposure to ionizing radiation.

MRI-Proton Density Fat Fraction (MRI-PDFF) is a more recent technique that makes use of chemical-shift imaging (CSI) to calculate liver fat. The precession of water protons is about 3.5 parts per million (ppm) quicker than that of fat protons [34] In order to achieve in-phase imaging (IP) or out-of-phase imaging (OOP) signals from protons in fat and water can be added or subtracted. The differences between IP and OOP signal images can be used to determine the signal intensity originating from fat protons. Many literatures suggest that MRI-PDFF has a better performance than MRI due to its ability to analyze multiple ROIs simultaneously. The drawback of MRI-PDFF includes the images being constructed from the indirect calculation of CSI measurement, still making the direct assessment of hepatic fat using MRI more accurate. Further, the use of MRI-based imaging techniques for diagnosis of FLDs is further limited by their high cost, limited availability and restrictions for patients with implanted metallic devices [35]. Table 2.1 presents a comparison between the different liver diagnosis methods.

Diagnosis Method	Strengths	Drawbacks
Liver Biopsy	- Examination with pathogenic value	- An invasive method which leads to bleedi and other complications - It's user dependent
Blood tests	- Initial method for assessing steatosis - High sensitivity for the level of AST and ALT	- Not specific - Doesn't related with the severity of steatosis
Magnetic resonance imaging	- The fat content can be quantified spectroscopy - Results accurate quantification of fat	- Very expensive - Inaccurate steatosis quantification in the existence of high iron concentration - Not suitable for patients with implanted Metallic devices
Computed Tomography	- characterization of fat with a decrease in CT attenuation -Quantitative measurement	- low sensitivity for mild steatosis - Ionizing radiation - Dependent on the machine setting
Ultrasound	- A primary modality for liver fat diagnosis - High specificity for fat content greater than 33% - Effective, low cost and non-invasive	- Inappropriate in the case of high BMI - Intraoperative variability /operator dependent

Table 2.1: Summary of liver steatosis diagnosis method.

1.5 Computer-Aided Diagnosis (CAD) of Fatty Liver Disease

CAD has been used extensively as a technique to manipulate medical images. It helps medical professionals in a wide range of medical imaging tasks like disease detection, object localization, disease grading, and treatment planning. Images with CAD offer a promising result to reduce differences in imaging performance [36]. They are computationally fast, have high accuracy, and are reliable as compared to the manual procedures of image classification. CAD systems are implemented using machine learning, deep learning, transfer learning and hybrid techniques.

Machine learning is a subset of artificial intelligence (AI), which allows a machine to learn meaningful patterns from data directly with handcrafted features. Often this technique is applied on US images,

particularly in imaging the liver. It is essential to increase the objectivity of the diagnosis of liver steatosis and the level of steatosis. However, the strength of learning depends on the handcrafted feature. In addition, it cannot learn or handle new data due to issues that include the problem of gradients converging to zero and failing to local maxima in that it can't be dynamically optimized when the data set changes [37].

Deep learning is a method that uses non-hand-crafted features to solve problems. It enables the creation of complex networks with deep layers to extract features from input data/images [38]. The last layer or classifier of the deep learning model may result in over fitting if less data is fed to the model. This leads to unnecessary usage of resources and complex computational power.

Transfer learning is a deep learning-based technique that involves using a pre-trained neural network as a starting point for a new or related task rather than training a neural network from scratch. It leverages the knowledge and feature representations learned from one task (the source task) to help improve the performance of a model on a different but related task (the target task). Transfer learning is especially useful when there is a limited amount of data for the target task and when training a deep neural network from scratch would be computationally expensive.

Hybrid learning is a technique that is constructed by combination of machine learning and deep learning that is done by taking advantage of both to efficiently solve the problem. In a previous study, this was done using Convolutional Neural Network (CNN) to extract features from unstructured data and the classical machine learning approach used as a classifier. This was experimentally proven to be a good approach to boost performance [38]. However, the choices of these techniques depend on the specific task requirement of the researcher.

1.6 Deep Learning

Deep learning is a subset of machine learning, which is a type of representation learning in which no feature selection is used. Instead, the algorithm learns which features are best for classifying the data on its own. This allows features to be extracted automatically at low- to high levels via training neural networks including the CNN. Its structure is similar to the human brain with a functional group of neural networks. The input layer, hidden layer and output layers make up the artificial neural network [39]. The input layer takes the input data and pass it to the next hidden layer. The hidden layer is made up of a set of layers (convolutional layers in the case of CNN) located between input and output layers. The

output layer is the final layer which results a specific output [47].The general architecture of a neural network is shown in Figure 2.4.

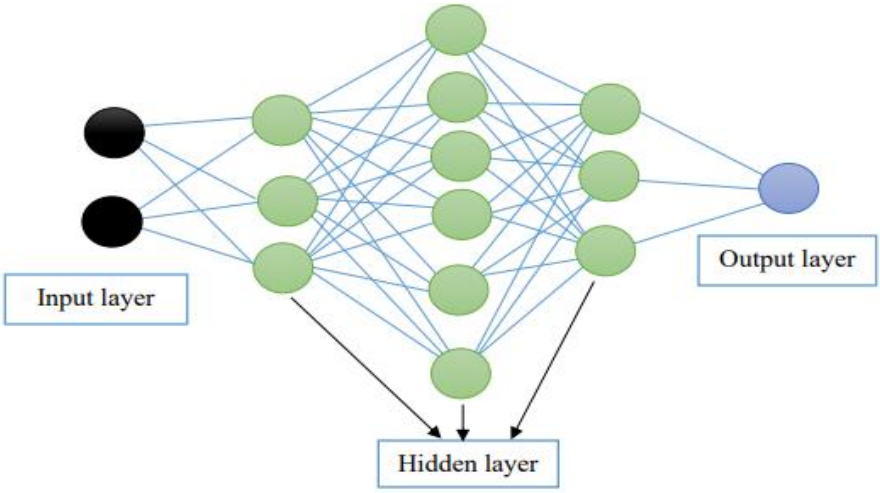


Figure 2.4: Architecture of a neural network [39].

CNN is an artificial neural network made of neurons with learnable weights and biases. It is composed of convolution, pooling, activation, and a fully connected layer. The convolution layer is a central part of CNN used for feature extraction from an input image and the image is convolved with a filter. The down-sampling operation or reducing the size of an input image is done by the pooling layer. The activation layer is used to increase nonlinearity in output to improve the expression ability of the network. A fully connected layer is at the end of CNN used to reduce feature maps as a classifier [40]. The most commonly used convolutional networks include VGGNet, MobileNetV2, ResNet, and Inception.

1.7 Statement of the Problem

According to recent studies, fatty liver disease (FLD) is emerging as the fastest growing cause of mortality due to liver-related complications. It is also emerging as a significant factor contributing to advanced liver conditions, primary liver cancer, and the need for liver transplants, imposing a substantial economic burden on healthcare systems. FLD is now the most prevalent form of CLD with its occurrence ranging from 13.5% people in Africa to 31.8% people in the Middle East. Particularly in Western industrialized nations, it is increasingly prevalent, especially among individuals with central obesity, Type 2 diabetes, dyslipidemia, and metabolic syndrome. It affects approximately 70% of overweight individuals, 70% of those with diabetes, and even up to 90% of severely obese individuals. Despite its

growing significance, FLD is often underestimated as a chronic disease, and there is a lack of comprehensive national strategies or policies addressing its management [8].

The incidence of liver diseases is increasing globally. In 2001, an estimated number of deaths from cirrhosis was 771,000 people worldwide which is currently estimated to be around 1 million. In sub-Saharan African nations, liver disease-related deaths doubled between 1980 and 2010 [41]. In Ethiopia, liver diseases accounted for 11.4% of all medical records according to the report made in 2021 [42]. Notably, in high-income countries, approximately 40% of liver diseases are detected at early stages, allowing for interventions with the potential for cure. In contrast, 95% of patients in sub-Saharan Africa are diagnosed with advanced or terminal liver disease, making it impossible to reverse the disease [43]. Consequently, there is an urgent need for early detection methods and tools for indication of the progression of the disease is highly needed.

Conventional diagnostics methods for liver diseases require centralized laboratories, experienced personnel and bulky equipment where the overall procedures are time-consuming and sophisticated. The diagnosis of the liver requires an extensive amount of information ranging from details of clinical symptoms to various types of blood tests. It could be serological, biochemical, or hematologic. Although liver biopsy is the most accurate method to diagnose and assess liver steatosis, it is not the best option for routine screening and diagnosis since it is invasive procedure. Sampling variability is also significant problem due to the fact that only one portion of the liver is sampled [44]. This can prove to be quite problematic resulting in a wide margin for sampling errors and potentially over or underutilization of therapy.

In the realm of imaging modalities, particularly 2-dimensional (2D) US, holds significant importance in diagnosing FLDs due to its non-invasive nature and simple operation. Unfortunately, its effectiveness is hindered by an unequal fat distribution, limitations in the interpretation of grayscale, and capabilities of the US equipment itself. Additionally, using 2D US results in a 20% dismiss of hepatocytes steatosis that have been verified by pathology [45]. These methods often exhibit reduced sensitivity in detecting mild steatosis.

The visual differentiation of liver images is susceptible to inter-observer variability, along with demanding significant time and labor resources. The diagnosis of FLD depends on radiologists' evaluations of texture attributes like homogeneity and echogenicity, often leading to substantial discrepancies. Literature suggests that visual assessment achieves an approximate diagnostic accuracy of around 72%, which is low [26].

Most of the available studies on liver tissue characterization rely on texture analysis, a method susceptible to variations in US machine settings, such as time gain compensation, frequency, and amplifier gain, making their characterization both observer and machine-dependent. This makes liver diagnosis quite difficult affecting its effective treatment. This calls for the development of other automatic techniques, deep learning being one, that could take diverse and relevant liver US images and classify them into categorized outputs (normal or classifying the degree of steatosis as G1, G2, or G3).

1.8 Significance of the Study

This research will have a significant impact on improving the current approach to the diagnosis and grading of FLDs. An automatic expert system will be developed targeting robust and accurate diagnosis and classification of fatty liver. It will enable professionals to easily differentiate normal liver from fatty liver and to decide treatment options for fatty liver depending on the severity level of the disease. The work is expected to encourage early liver disease detection, reducing consequences that could develop as a result of diagnostic constraints. The scheme should provide suitable knowledge to physicians in their decision-making when assessing liver steatosis by reducing time, subjectivity, and workload.

1.9 Scope of the Study

The study will use US liver image patient data with known clinical outcomes to develop and test the proposed detection and classification algorithm. The research is aimed to identify normal cases from fatty liver and assess the level of steatosis using a deep learning scheme as a decision support system in the area of diagnosis and treatment. Clinical implementation of the method is not addressed in the current study.

1.10 Objectives

1.10.1 General Objective

The main objective of this thesis research is to build a deep learning system for use in the automated detection and grading of FLDs.

1.10.2 Specific Objectives

- To implement pre-processing techniques on the acquired liver US image;

- To compare pre-trained models and identify the best model that could be used to classify liver steatosis grades;
- To test the performance of the algorithm using known evaluation measures based on useful quantitative matrices;

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORKS

2.1 Ultrasound Based Fatty Liver Diagnosis using CAD – Review

There are various image processing methods developed in the literature to evaluate the degree of steatosis. As mentioned earlier, the GLCM and the HRI are the most well-known and straightforward image-based screening methods for liver steatosis [28]. These techniques, however, heavily rely on the ability to select a good ROI and the expertise of the doctors doing the examination. Recently, several deep learning-based approaches have been introduced in the literature to overcome the issues and challenges associated with US image quality and operator dependency for the diagnosis of liver diseases.

To support the classification of liver diseases, Andrea et al. [46] suggested a CAD system using feature extraction. The proposed method is based on co-occurrence matrix, first order statistics, run length matrix and fractal dimensions. Three different classifiers namely artificial neural network (ANN), a support vector machine (SVM), and a k-nearest neighborhood (KNN) algorithm were used for the evaluation of specific features. Based on the results reported by the authors, their CAD system offered 79.77% accuracy, which may not necessarily be sufficient for automated clinical applications. In another study, a shallow CNN was suggested by Zhu et al. to assess the degree of liver steatosis [47]. The researchers considered that the US image of a normal liver is uniform, whereas a fatty liver's image is non-uniform in appearance. The original liver US pictures obtained from commercial scanners are very dark and shaded to correctly identify liver steatosis using a shallow CNN based model. The proposed model achieved an accuracy of 92%. The paper acknowledges the issue of limited data, and it's essential to consider potential biases or imbalances in the dataset, which may affect the generalizability of the model to different populations or clinical scenarios. Additionally, the reliance on pixel-level features may lead to interpretability challenges, as the model might capture patterns that are not clinically meaningful.

Another approach using three image-processing methods was examined by Cao et al. where a total of 240 images consisting of normal, mild, moderate, and sever fatty liver cases were examined by envelope signal, grayscale values, and a neural network [48]. The authors compared the performance of the three methods by drawing ROC curve and calculating the area under the curve (AUC) among the four groups.

They were able to achieve the highest value of AUC when using deep learning technique which is CNN. The calculate AUC values for deep learning, envelop signal, and gray scale value methods were 0.933, 0.859, and 0.857, respectively. Their comparison demonstrates that the deep learning approach had the highest ability to figure out the how severe FLDS were. However, they used a shallow network (not deep network) architecture consisting of only three convolutional layers and two fully connected layers. There are obviously rooms for improvement regarding the overall performance of their method. There is also a need to validate their method on a larger dataset to enhance the generalizability of their model.

Inception-ResNetV2, GoogLeNet, AlexNet, and ResNet101 were the four pre-trained networks utilized by Zamanian et al. to extract features from the initial data proposing a new scheme [49]. Features extracted from the four networks were summed and SVM was used to categorize them. The outcome of their proposed method was evaluated with that of the four individual networks. The accuracy offered from Inception-ResNetV2, GoogLeNet, AlexNet, ResNet101 and their proposed algorithm was 81.08%, 94.60%, 99.32%, 99.32% and 98.64%, respectively. The experimental findings reveal that the individual pre-trained networks are more accurate than the suggested algorithm merging the four networks. Specifically, AlexNet and ResNet101 produced better results. The study was, however, limited to binary classification (normal liver and fatty liver) without assessing the severity of the fatty liver.

Different literatures have proposed fatty liver classification using transfer learning approach. In one of the previous studies, researchers used pre-trained InceptionV3 and VGG-16 using fine-tuning [6]. The top layers of the pertained models were replaced by newly connected layers with random parameters. After augmentation and rescaling, the models were trained and validated with a total of 629 images. A comparison was done between InceptionV3 and VGG-16 and the accuracy obtained through these methods was 93.23% and 90.70%, respectively. However, a larger data set is required to validate the generalizability of the proposed method. A similar study classified B-mode US liver images of patients using HRI, GLCM, and CNN algorithms, with liver biopsy as a reference. The results showed that the Inception-ResNetV2 pre-trained model was efficient and operator-independent, with higher accuracy (96.3%) than HRI and GLCM. It is difficult to draw the conclusion that the pre-trained CNN outperforms the HRI and GLCM based classification methods considering the performances of the HRI and GLCM are significantly influenced by the selection of ROIs [30].

On the other hand, a quantitative tissue characterization technique (QTCT) has been proposed based on the classification of fatty liver using radio frequency (RF). Their algorithm generates two images

estimated from the RF signal generated by the US probe: a de-speckled image containing the anatomic and echogenic information of the liver and another image containing only the speckle used to compute textural features. The intensity features from both images were calculated and related with two texture features. The classification process was applied to ten steatosis images and 10 non-steatosis images. Then depending on three distinct classifiers (k-NN, Bayes and SVM), the highest accuracy they were able to achieve was 93.54% [50].

In another study, a voting-based classifier and machine learning algorithms were used to classify liver tissues as being fatty or normal based on the attributes extracted US images. Liver image classification as normal or fatty was accomplished without a segmentation phase. GLCM and First-Order Statistics (FOS) were used to extract 26 features from US images. The vote classifier was used to finally identify the type of liver tissue. A comparison was performed using voting-based classifier and J48 algorithm on a dataset. The researchers obtained a classification accuracy of 95.71%; and 93.12% when utilizing the voting-based classifier and the J48 method, respectively [51]. Despite an increasing performance these techniques still have a number of drawbacks. The limited number of images in their dataset was one issue. Also, the study considered only binary classification (normal and fatty liver) while grading of the diseases is critically important. It should be noted that specialized radiologists find it challenging to identify these levels of liver steatosis by eye. Furthermore, there is a challenge in class imbalance. For example, liver cirrhosis usually only affects a small number of people in comparison to other liver diseases such as liver steatosis, liver cancer, and liver fibrosis. This could be a source of class imbalance when developing a fatty liver grading scheme.

Another study proposed an automated liver steatosis diagnosis using a Cascaded deep neural network. The liver and kidney (L-K) are first semantically segmented on parasagittal US pictures using transfer learning, and the L-K area is subsequently cropped from the original US images. The second neural network involves semantic segmentation by checking the L-K region from the original US images and testing for the existence of a ring that is typically present in the kidney. The final neural network, Steatosis Net, uses these clipped L-K sections to assess the severity of steatosis. The study offer an accuracy of 99.91% which is comparable to annotations done by specialists [52]. This approach outperforms the aforementioned studies. However the segmentation method utilized was computationally complex.

2.2 Research Gap

Even though the aforementioned works achieved promising results, still improvements are needed to fill the gaps. These gaps include US image degradations caused by existence of speckles that obscure the details of the tissue being imaged. It notably reduces image quality and hinders the identification of intricate features. Therefore, it is highly needed to mitigate this speckle interference prior to any application. Furthermore, there is little work previously reported in areas of classification of fatty liver and the grading of steatosis. It should be noted that these classes of steatosis are difficult to recognize by radiologists through visual methods.

CHAPTER THREE

RESEARCH METHODS

This chapter focuses on the methodology followed in this study starting from data collection to model deployment.

3.1 Research Design

The research aims to develop an automated detection mechanism for fatty liver disease and to classify the liver steatosis stages by making use of AI. The proposed system aimed to classify liver cases into G0 (normal) G1 (mild), G2 (moderate), and G3 (severe). In this study, experimental research methodology is followed. It enables the researcher to observe the effect of an independent variable on the dependent variable. The independent variables are the acquired US images and the dependent variables include the performance metrics.

The data set used in this study is composed of US images of the liver. The data set was split into training (used to construct the model), validation and test set (used to evaluate the performance of the constructed model). Prior to classification, different pre-trained models including VGG16, ResNet50, InceptionV3, Xception, and MobileNetV2 models are compared for their efficacy in classifying the liver cases using different model hyper-parameters in order to pick the best performing model.

3.2 Data Collection and Preparation

The proposed study assess the degree of steatosis for confirmed fatty liver images. For classification tasks, US image data was acquired from Zenode repository, a publicly available database. The dataset consists of B-mode liver US images which are used in the current study as ground truth information. The non-important features (like white points, and reading bars) in the data set were removed prior to further processing. Contrast limited adaptive histogram equalization (CLAHE) and Gaussian filters were applied during pre-processing for better performance of the classification models. The data set was split into three subsets to serve as the training, validation, and testing sets. As a result, the whole data set was split into 90% for training and out that 10% was used for validation and the remaining 10% for testing. During model training different data augmentation techniques applied to provide diverse data to the models.

Once the dataset is preprocessed and partitioned, pre-trained CNN models are established to train and select the best pre-trained model for the specific purpose. This is done by optimizing the parameters in the top layer only. In the current thesis, parameter tuning was applied through a random search, where the number of epochs, the learning rate, the number of top layers in the CNN models, and the dropout probability were varied to find the optimal values. The other parameters were kept at their default settings to maintain the robustness of the original CNN models.

3.3 Implementation Tool Selection for Graphical User Interface Development

The study being proposed includes two models that have been developed to achieve the best performance for two cases of classification. First, a binary class model was developed that is able to differentiate between abnormal (Fatty liver) and normal liver cases. Then a four-class model was also computed to classify normal controls and mild, moderate, and severe steatosis cases**.**

The final model was selected to be the best model and used in the user graphical user interface (GUI) after measuring the performance of the models using useful quantitative matrices. The GUI was developed on the Gradio library. Gradio is the fastest way to demonstrate a deep learning model with a friendly web interface. The user interface was developed through continuous guidance and feedback that has been collected from clinical collaborators and advisors. Such interfaces make life of professionals using such classification systems (in our case say people working in the gastroenterology department) very easy.

3.4 Performance Evaluation

In this study, several evaluation metrics were computed to allow comprehensive comparisons against previous studies. A test dataset was utilized to evaluate the performance of the models developed for both binary and multi class classification tasks. Different metrics were used to evaluate the models performance including confusion matrix, accuracy, precision, recall, F1 score, and receiver operating characteristics (ROC) and the corresponding area under the ROC (AUC) value. The confusion matrix provides a tabular report of the actual and predicted number of test images, forming the basis for calculating the other metrics such as accuracy, recall, Fl score, and precision. The ROC curve is a graphical method used to assess the model’s performance at different threshold values.

3.5 Materials Used in This Study

In this thesis work *Google Collaborator* was used as the code editor to write and execute the codes used for the classification algorithm. It is a cloud-based computing platform that provides a free GPU with 12 GB of memory for up to 12 hours. The system uses Python 3.8.8 with various modules including Keras, Tensor Flow and Scikit-learn. Python is an open source programming language which is easy to use. It is used with different packages that support overall computation which can enhance the system performance. To build the GUI, an open source Python library called Gradio is used. Matlab (v2022) was used during pre-processing of the US images before they were inputted to the deep models. Table 3.1 summarizes the materials used in the current study.

Software	**Hardware**
● Python 3.8.8 ● MATLAB 2022 ● Tensor flow version 2.12 ● Google Colab	● Personal Computer Lenovo ThinkPad CORE i5

Table 3.1: Materials used in the study.

CHAPTER FOUR

RESEARCH METHODOLOGY

4.1 Overview of Liver Steatosis Classification

Figure 4.1 demonstrates the steps followed to develop the final liver steatosis classification system. Four class classification model is developed to classify liver US images in to into normal, mild, moderate, and severe. After developing the model, the performance of five deep pre-trained CNN models, namely VGG16, ResNet-50, MobileNetV2, Xception and InceptionV3 was assessed to pick the best performer. Training of the deep models was preceded by a pre-processing step in order to remove unnecessary parts of the input US images and the pre-processing steps were implemented on a Matlab environment. CLAHE and Gaussian filters were applied as pre-processing techniques. The dataset was split into three subsets: training, validation and testing sets. Finally, the prepared data is fed into the classification model.

4.2 Data Collection

The dataset consisting of 550 US images and the biopsy results, available on the Zenodo repository. Table 4.1 displays the entire amount of data together with its corresponding class and Figure 4.2 represents sample US images taken from the database [53].

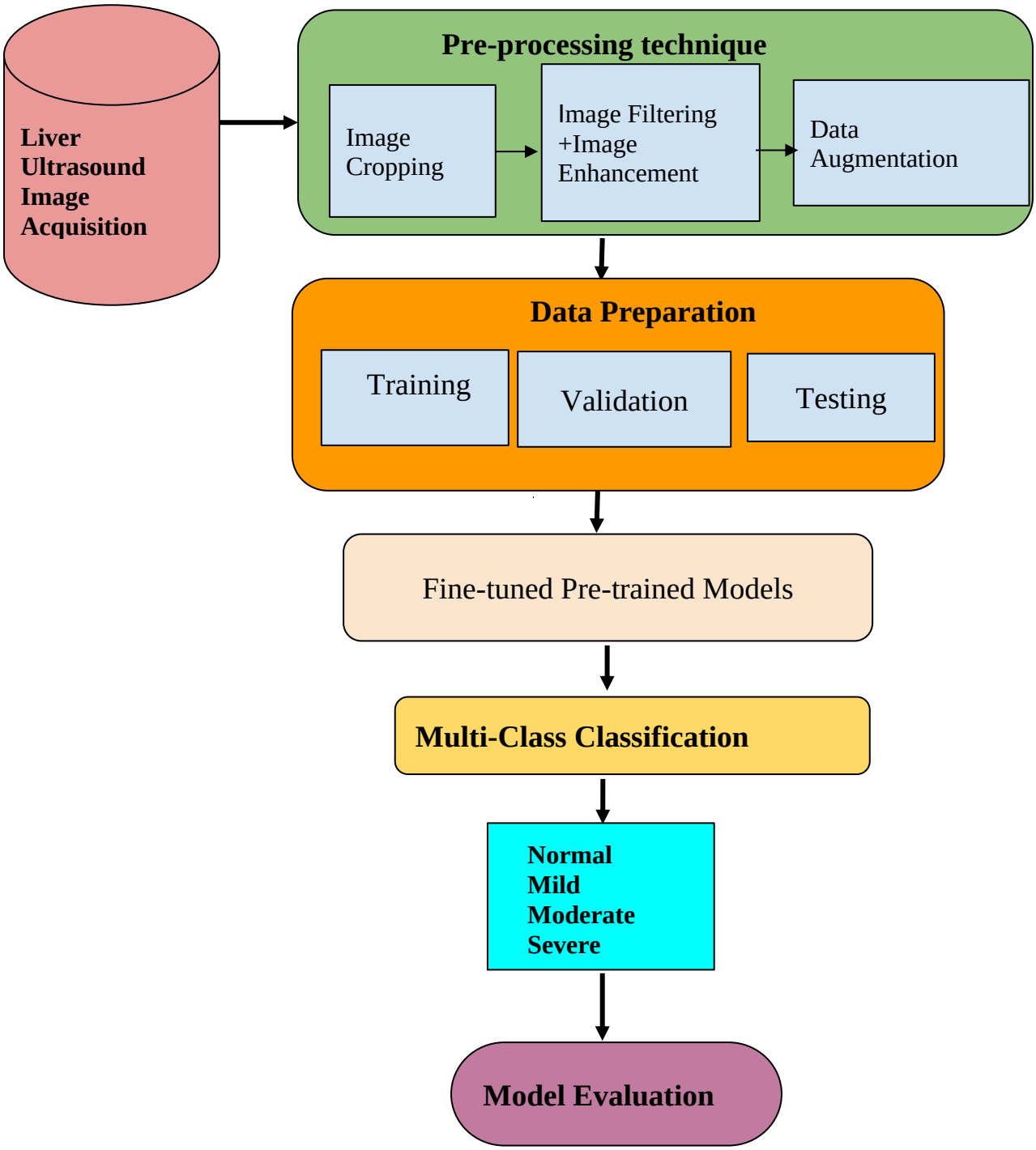


Figure 4.1: General architecture of liver steatosis classification.

Liver Type	Number of Images
Normal	180
Mild Steatosis	190
Moderate Steatosis	80
Severe Steatosis	100

Table 4.1: The total amount of data sets used for the study.

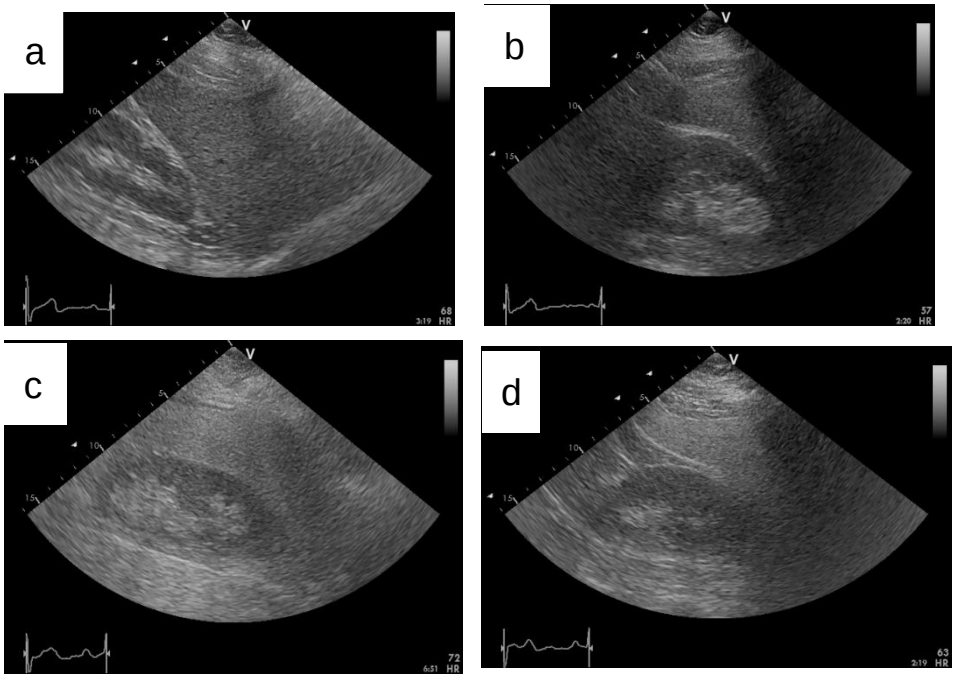


Figure 4.2: Sample images from the dataset: (a) Normal, (b) Mild steatosis (c) Moderate steatosis, and (d) Sever steatosis.

4.3 Image Pre-processing

Pre-processing is mostly used to enhance the image quality. The effectiveness of the classification task is greatly influenced by the quality of preprocessing technique [54]. Pre-processing is carried out by extracting valid regions, enhancing the image for extracting essential information, and removing noises from the image for further analysis.

The common problem in US images is low image quality due to low contrast and speckle and other noises. US images have grayscale values that are susceptible to speckle noise in the image. Speckle is

multiplicative noise that contains echo disorders which result in undesired interference effects on images [55]. The purpose of the pre-processing step is then to suppress such noises and to improve the contrast while retaining important image features for the subsequent image processing tasks.

4.3.1 Image Cropping

Ultrasound images often contain areas outside the ROI that are not relevant to the diagnostic task. Pre-processing is applied on US images using a Matlab software to exclude irrelevant data such as frame number and reading bares. By removing these unnecessary parts of the image, the model can focus its attention on the most relevant and diagnostically significant features within the image. This can also reduce the computational burden on the network, allowing it to process the remaining relevant information more efficiently.

This is done by using the steps below.

1. Collecting the US images and finding the right and left slope of the images that consist of unnecessary parts (bar, frame numbers)
2. To obtain the slope coordinates a Matlab function “impixelinfo” is used. Then the right and left slopes of the image are calculated.

$$C1 = [32,229] \qquad C1 = [603,226] \quad \text{..... (4.1)}$$
$$C2 = [32,229] \qquad C2 = [316,8] \quad \text{.....(4.2)}$$
$$m1 = \frac{y2-y1}{x2-x1} \quad \text{And} \quad m2 = \frac{y2-y1}{x2-x1} \quad \text{.....(4.3)}$$
$$m1 \approx -0.77 \qquad m2 \approx 0.77 \quad \text{.....(4.4)}$$
$$Y1 = m1x + b \qquad Y2 = m2x + b \quad \text{..... (4.5)}$$
$$Y1 = -0.77x + 253.64 \qquad Y2 = 0.77x - 238 \quad \text{.....(4.6)}$$

3. The part of the pixels above the right and left slope is then converted into black (i.e. assigned a value 0).
4. Then image is cropped horizontally (from the bottom) for further removal of unnecessary parts.

4.3.2 Image Filtering

To reduce speckles noise and other noises present in US liver images, different filtering methods including Median filter, Shock filter, Gaussian filter, Butterworth low pass filter and Weiner filter are used. In the current study, Gaussian filtering is used to suppress speckle noise. Gaussian filtering is a smoothing technique that works by convolving the image with a Gaussian kernel. As one of the efffective low-pass filters, it is used to reduce the negative effects of noise by substituting the pixel value

at that point with the weighted average of the pixels in its neighborhood to achieve a better noise reduction at the global scale [55]. The de-noising step is required to improve image quality by reducing noise in ultrasound images and making liver disease diagnosis more accurate.

4.3.3 Image Enhancement

Ultrasound imaging is a non-invasive technique used to diagnose various liver diseases, including fatty liver disease. However, US images can be affected by noise and non-uniform illumination, making it difficult to detect small changes in tissue structure and affecting the performance of the model. Therefore, in order to avoid the problem of image information loss due to the presence of excessive brightness in some regions of the image, the current study used an image enhancement technique called Contrast-Limited Adaptive Histogram Equalization (CLAHE) to improve the image quality. CLAHE is the improved version of Adaptive Histogram Equalization (AHE). CLAHE operates on a small region on a grayscale image by limiting the contrast enhancement of AHE [56]. It was developed to prevent the overall amplification of noise that adaptive histogram equalization can cause. One advantage is that the portion of the histogram that exceeds the clip limit is redistributed equally among all histogram bins rather than being eliminated. The CLAHE approach is shown in Figure 4.3 and the technique follows three steps. The tile generation, histogram equalization, and bilinear interpolation. The input image is first divided into sections. Each section is called a tile. The input image shown in the figure is divided into four tiles. Histogram equalization is then performed on each tile using a pre-defined clip limit. Histogram equalization consists of five steps: histogram computation, excess calculation, excess distribution, excess redistribution, and scaling and mapping using a cumulative distribution function (CDF). The histogram is computed as a set of bins for each tile. Histogram bin values higher than the clip limit are accumulated and distributed into other bins. CDF is then calculated for the histogram values. CDF values of each tile are scaled and mapped using the input image pixel values. The resulting tiles are stitched together using bilinear interpolation, to generate an output image with improved contrast.

4.4 Data Augmentation

Deep learning algorithms for image classification face significant challenges due to limited amount of data. Class imbalance is frequently a problem; while some classes have enough data, others that are equally significant, but are under sampled will have low class-specific accuracy. To overcome this,

image augmentation is one useful technique that can increase the size of the training set without acquiring new images [57]. In this study, augmentation techniques using image rotation at 20°, rotation at -20°, horizontal reflection (X reflection) and vertical reflection (Y reflection) were applied to the original image. Consequently, a significant and substantial amount of data that represent the same class as the original image is reproduced, which enables the model to correctly learn the features of an image. Note that augmentation is performed only to the training dataset.

In addition to this, to ensure conformity with the image classification model and to optimize computational efficiency, the images were scaled to a uniform size. Consequently, all data was resized from their initial dimensions of 636 × 434 pixels to a smaller size of 224 × 224 pixels.

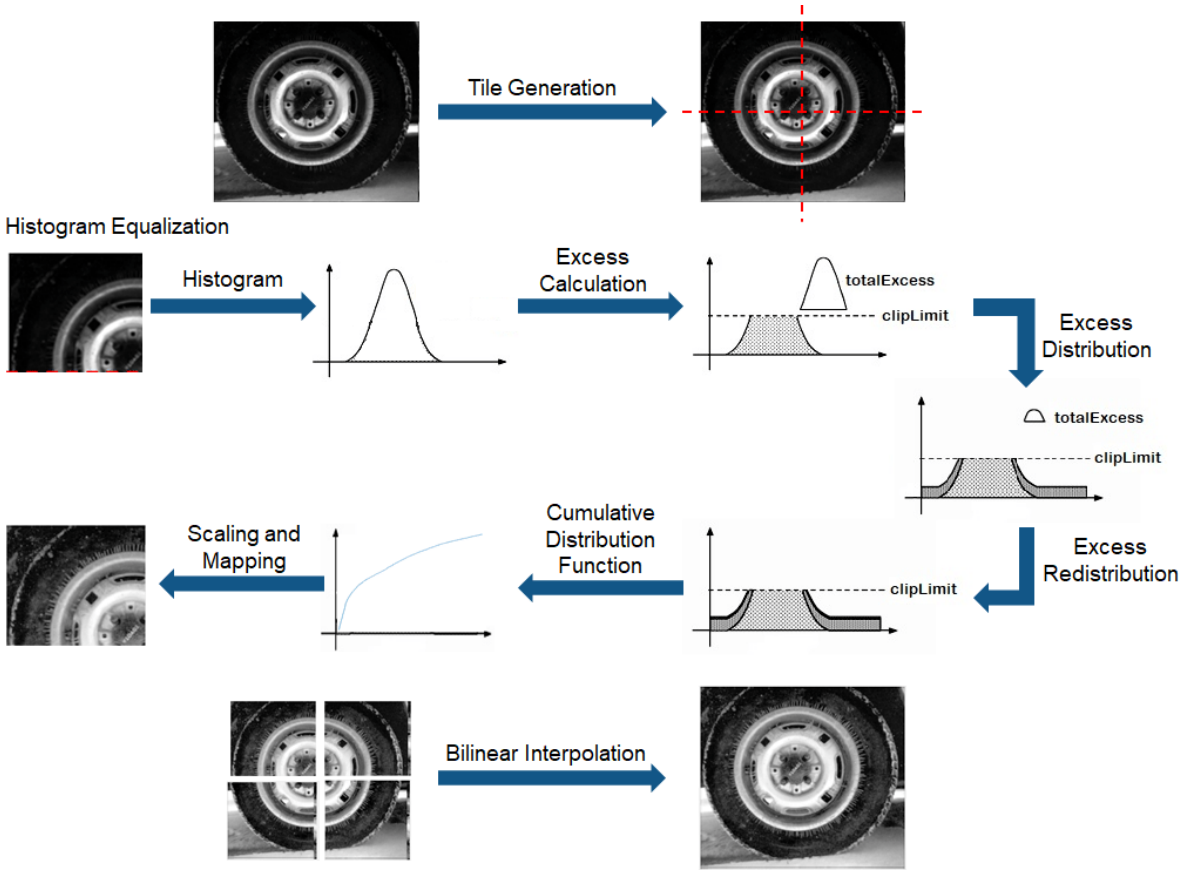


Figure 4.3: Contrast Limited Adaptive Histogram (CLAHE): [58].

4.5 Data Preparation

In this research, a total of 550 images were available in the dataset acquired from Zenodo repository which was partitioned into three: 80% for training, 10% for validation and 10% for testing. Before pre-

processing, the training set was augmented. Figure 4.3 displays the distribution of images and their corresponding classes across each of the four groups.

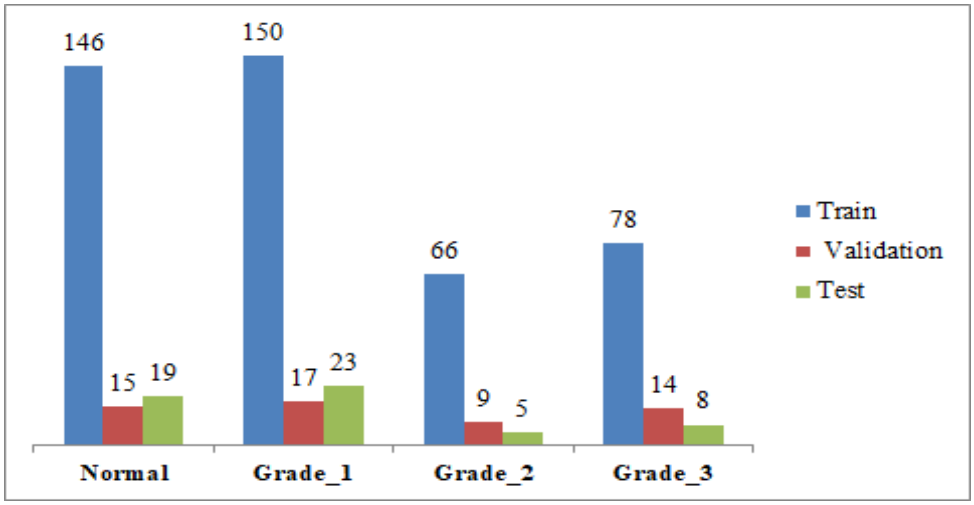


Figure 4.4: Ultrasound image data set.

4.6 Image Classification

The efficiency of CNNs in image processing is well-established and pre-existing CNN models have proven to be particularly useful for image classification. In order to determine the optimal model for fatty liver diagnostic prediction, experiments were done using different pre-trained models such as ResNet50, VGG16, InceptionV3, Xception and MobileNetV2. The pre-trained models were validated using a train-validation-test split of the entire data set. Various optimization model parameters were adjusted during training to test the model's performance. The number of top layers in the CNN models and the dropout probability were varied to find the optimal values. The study utilized RMSProp (Root Mean square propagation) and Adam (Adaptive Moment Estimation) optimizers varying the learning rate between 0.01 to 0.0001, while epoch was varied between 10 and 50 while the batch size was varied between 32 and 256. The primary objective was to determine the most suitable algorithm for accurately predicting FLDs.

4.6.1 VGGNet

VGGNet is a CNN architecture that is primarily concerned with the effect of CNN depth on accuracy. It accepts input image with size 224x224 RGB image and deeper than other models including AlexNet. VGGNet has VGG16 and VGG19 families which are 16 and 19 layers deep, respectively. VGG16 is a

simpler and more straightforward architecture and it has been tested in the current study. However, the VGG network in general result in longer training and inference times and larger network weights, which can limit their predictive performance. The structure of VGG16 is shown in Figure 4.5 [59].

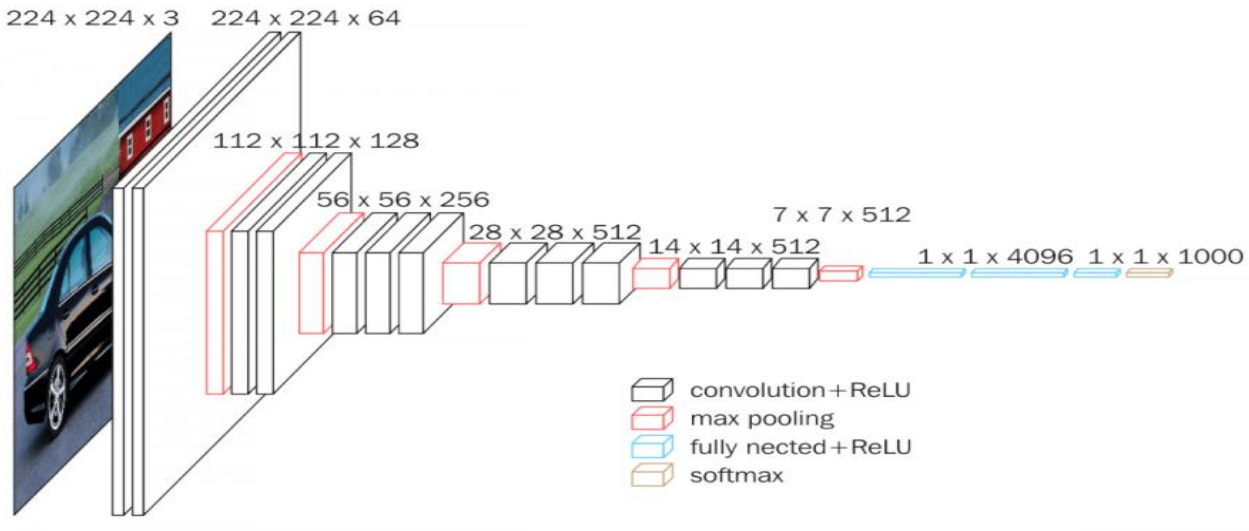


Figure 4.5: VGGNet [59].

4.6.2 ResNet

ResNet (Residual Network) is a CNN that was developed in 2015 by Microsoft researchers. The main idea behind ResNet is to overcome the problem of vanishing gradients in deep neural networks, which can make it challenging to successfully train very deep networks. ResNet overcome this by incorporating a residual connection between the input and the output of each block of the network, allowing the gradients to flow more easily through the network and making it easier to train deep networks [60]. Figure 4.6 represents the residual block of ResNet.

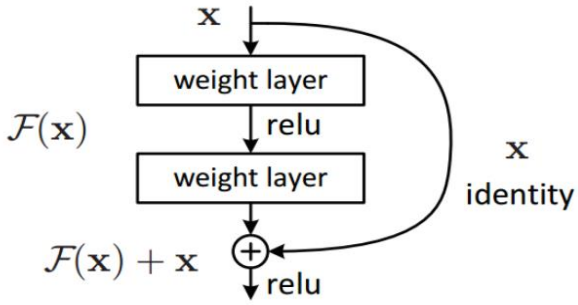


Figure 4.6: Residual learning building block [60].

ResNet architectures typically consist of a series of stacked residual blocks, consisting of one or more densely connected layers for classification. Residual blocks are built using skip connections that allow information to be passed directly from one layer to another, skipping one or more layers in between. Each residual block is made up of two or more convolutional layers, followed by a skip connection that adds the input to the output of the last convolutional layer. The input image is passed through a series of convolutional and pooling layers to extract features, and then through the residual blocks to learn complex representations of the input. The output of the last residual block is then flattened and passed through one or more fully connected layers to produce a class prediction. This residual connection also helps to alleviate the problem of over fitting, by allowing the network to learn more complex and hierarchical representations of the data [60].

ResNet50, the ResNet model tested in the current study, was demonstrated with the aim of simplifying the training process. Compared to other models such as VGG and GoogLeNet, ResNet50 demonstrated superior performance on the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) datasets. ResNet152, which is a deeper version with 152 layers, showed even better predictive performance. However, when compared to InceptionV3 and Xception using the ILSVRC-2012 data set, the ResNet152 network underperformed [60]

4.6.3 Inception

Inception is a family of pre-trained models, initially introduced by Google in 2014. The Inception family of models has evolved over the years, with the latest iteration being InceptionV4, which was introduced in 2016 [61]. The key feature of the Inception models is their ability to perform multi-level feature extraction, which enables them to identify objects at various scales and orientations within an image. This is achieved through the use of multiple convolutional layers with different filter sizes, which allows

the model to capture both high and low-level features within an image. Inception models also make use of various pooling techniques, such as average pooling and max pooling, to further enhance their ability to capture features at different scales [61]. Figure 4.7 indicates a rough architecture of the original Inception model.

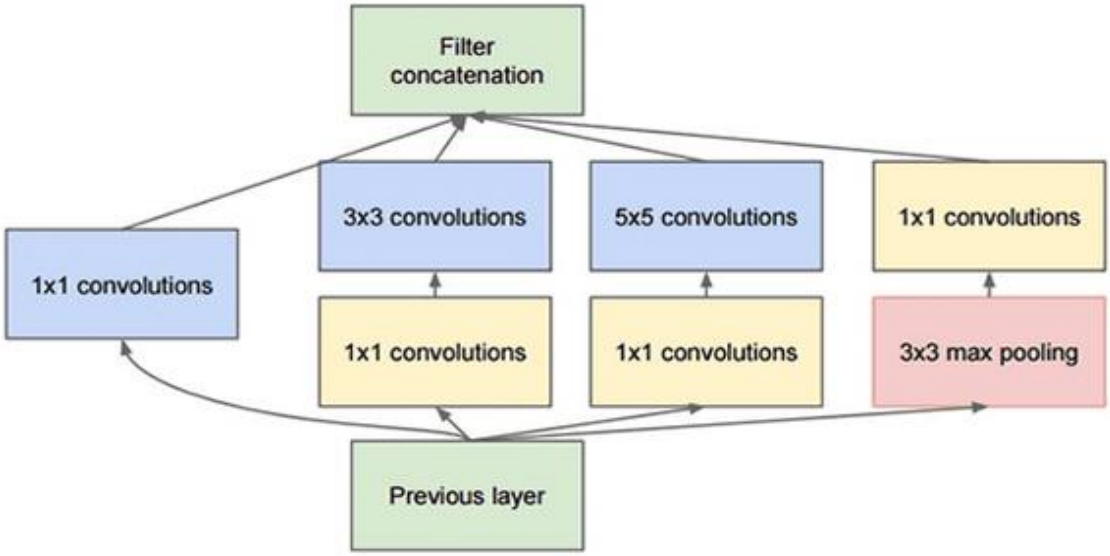


Figure 4.7: The original inception model [61].

Each inception architecture consists of four operations in parallel. These are the 1x1 convolutional layer, 3x3 convolutional layer, 5x5 convolutional layer, and max pooling. The 1x1 convolutional blocks seen in yellow are used for dimensionality reduction. The results from the four parallel techniques are then combined depth-wise to form the Filter Concatenation block (in green).

There are several families of Inception models each with its unique characteristics and improvements over its predecessor. These are InceptionV1, InceptionV2, InceptionV3 and InceptionV4. In terms of performance, InceptionV4 is considered the most accurate of the Inception models, with InceptionV3 being the second one. However, InceptionV4 also has the most significant computational cost, with over 40 million parameters. InceptionV3 was the Inception model that was tested in the current study. Overall, the choice of the Inception model depends on the specific task requirements, computational resources, and desired accuracy.

4.6.4 Xception

Xception is a CNN that was developed by Fracois Chollet in 2017 [62]. It is the most recent CNN model with 71 deep layers, which is one of the smallest weighted models. Xception, which stands for "Extreme Inception" is built on the same principle as the Inception family but differs significantly from it in terms of its architectural design. The Xception model separable convolutions are illustrated in Figure 4.5.

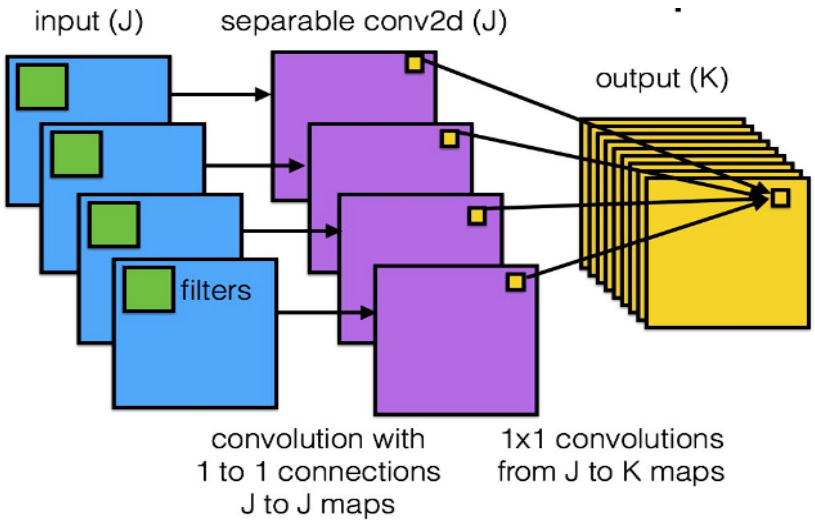


Figure 4.8: Xception architecture [63].

In the Xception model, the convolution operation is split into two separate operations: depth-wise separable convolution and point-wise convolution. Depth wise separable convolution performs the convolution on each channel separately where as a standard convolution applies convolution across all channels. Point wise convolution conducts a 1*1 convolution over all channels. The major goal is to reduce the number of multiplication operations involved because it is more expensive than addition. This enhances the computational power and minimize the number of parameters in the convolutional operation [62]. Unlike that of Inception, the Xception model applies depth-wise separable convolution followed by point wise convolution which removes the presence of non-linearity. This approach reduces the number of parameters in the model while maintaining high accuracy.

The Xception model has several advantages over the Inception models, such as improved accuracy and faster training times. Xception also has a smaller memory footprint compared to Inception models, which is essential when working with limited resources. In terms of performance, Xception outperforms InceptionV3 and InceptionV4 on many image recognition tasks [64]. Overall, Xception is a powerful

pre-trained model for deep learning that offers significant improvements over the Inception family of models, with high accuracy and faster training times, making it an attractive option for many image classifications. The Xception model was one of the deep models tested in the current study.

4.6.5 MobileNetV2

MobileNetV2 is a lightweight model with 53-layer deep CNN, designed to be used in mobile applications. It is based on an inverted residual connection where the shortcut connections are between bottlenecks. It uses depth wise separable convolution followed by point wise convolution [65]. It also has different versions and one of those, MobileNetV2 is shown in Figure 4.9.

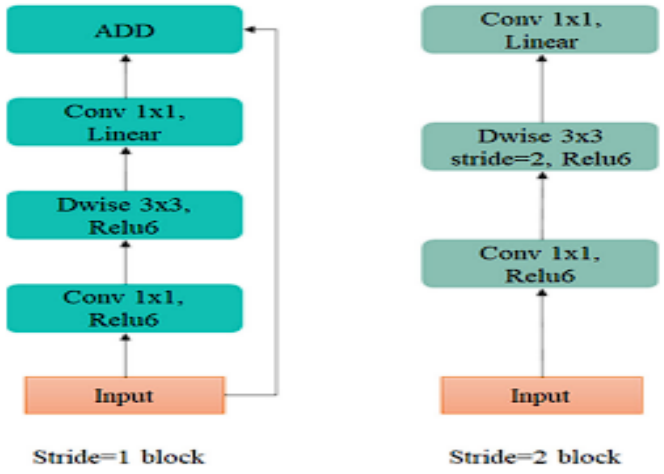


Figure 4.9: The convolutional blocks in MobileNetV2 [66]

4.7 Regularization Techniques

Overfitting is a typical problem in neural networks that reduces a models capacity to generalize. This is a situation when a model trains successfully on a training data but loses its capacity to predict on new data. To handle this, a variety of methods including early stopping, L1, L2, and dropout techniques have been proposed in the current work. In this research, to smooth the parameter distribution, L2 layer weight decay regularizer was utilized. As a result, the model's overfitting is decreased and its accuracy is boosted.

4.8 Performance Evaluation Metrics

Assessing the performance of a trained model is crucial to determine its effectiveness in various aspects. There are several widely used performance metrics, including the confusion matrix, accuracy, precision, recall, F1 score, and ROCAUC plot. To generate a classification report, the Scikit-learn Python library is used in this study.

4.8.1 Confusion Matrix

A confusion matrix is a tabular representation of a classifiers performance on a given data set. It presents a summary of the actual and predicted values for each class in the data set, allowing us to evaluate the effectiveness of the classifier. To evaluate the models' effectiveness, a test set of 55 images that represented 10% of the total image dataset was used. A confusion matrix was used to evaluate a models performance on an unseen dataset. For illustration, four-class classification evaluation metrics is shown in Table 4.2 with labels A, B, C, and D. In the matrix, the diagonal cells reflect occurrences that have been successfully classified, whereas the off-diagonal cells represent instances that have been incorrectly classified. By analyzing the confusion matrix, we can identify the types of errors being made by the classifier, such as false positives or false negatives. Additionally, performance metrics such as accuracy, precision, recall, and F1 score can be calculated based on the values in the confusion matrix, providing further insights into the model's performance.

It was calculated from confusion matrices as [67]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4.7)$$

$$Recall = \frac{TP}{TP+FN} \quad (4.8)$$

$$Precision = \frac{TN}{TN+FP} \quad (4.9)$$

$$Specificity = \frac{TP}{TN+FP} \quad (4.8)$$

$$F-measure = \frac{2*Recall*precision}{Recall+Precision} \quad (4.10)$$

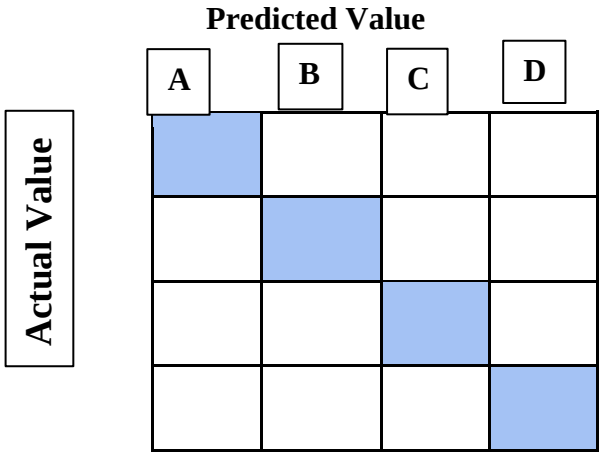


Table 4.2: Confusion matrix.

4.8.2 Receiver Operating Characteristics (ROC) and Area under ROC (AUC)

AUC ROC plot visualizes how well the model can distinguish between classes. The ROC curve can be plotted with two known methods for binary and multi-class classification which include one versus one class and one versus rest classes, respectively. Additionally, the AUC score can be calculated for each class.

4.9 User Interface (UI)

In this study, the Gradio platform is used for creating user interfaces. It is an open-source Python library customizable for deep learning models. Gradio provides customizable components to build UI which include checkboxes, buttons, canvas, message boxes and text boxes. The overall system architecture is shown in Figure 4.10 showing the layout for classification and grading of steatosis.

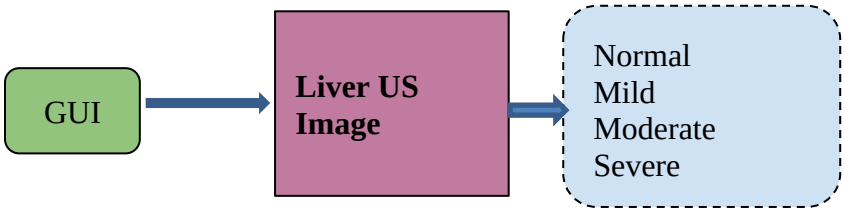


Figure.4.10: The overall system architecture including the GUI.

CHAPTER FIVE

RESULTS AND DISCUSSSION

5.1 Liver Ultrasound Dataset

In this study, the dataset contains 550 images classified into four classes. Among these, 180 images are Normal liver (G0), 190 images of Mild fatty liver (G1), 80 images of moderate Fatty liver (G2), and 100 images of severe fatty liver (G3). Figure 5.1 presents the US images data set of these four classes.

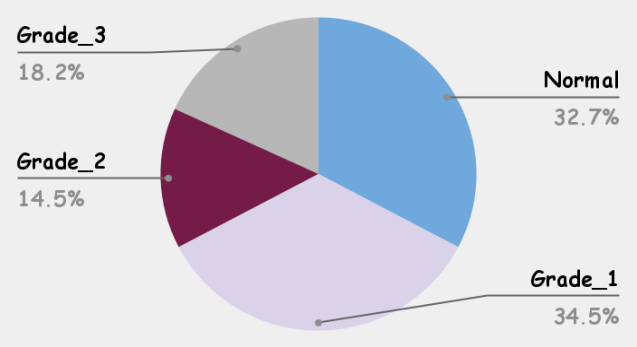


Figure 5.1: Liver ultrasound image data set.

The effectiveness of the model is affected by data imbalance. The class balance was done using the augmentation techniques 20°, -20°, 'X Reflection' and 'Y Reflection' on the image data set. Figure 5.2 presents different degrees of augmentation techniques used in this study. Figure 5.3 represents the proportion of augmented images used for training. In order to further enhance the data imbalance issue, an over sampling technique was utilized.

5.2 Pre-processing of Liver Ultrasound Images

The first step of fatty liver classification was pre-processing the US images. The non-relevant features are removed using Matlab and resized to 224x224 to be appropriate for the pertained models. A sample preprocessed (cropped and resized) image is shown in Figure 5.4.

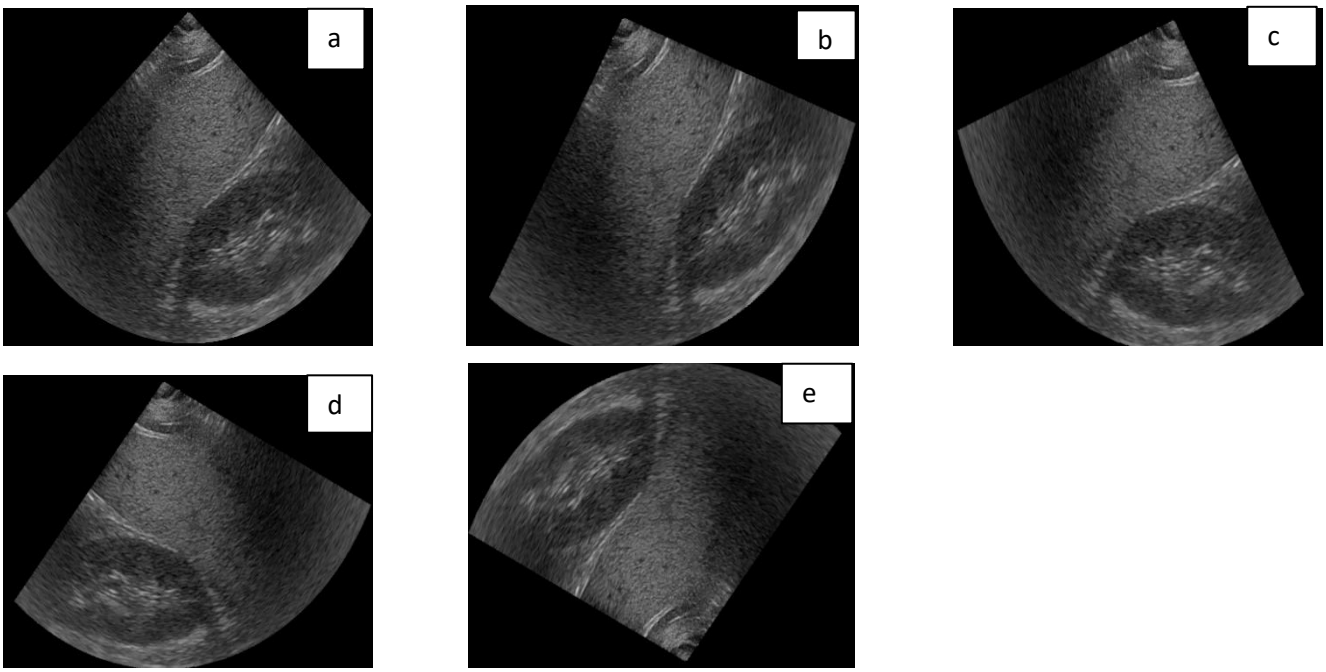


Figure 5.2:Results of Augmentation: (a) Original image, (b) 20° rotated image, (c) -20° rotated image, (d) X Reflection, and (e) Y Reflection.

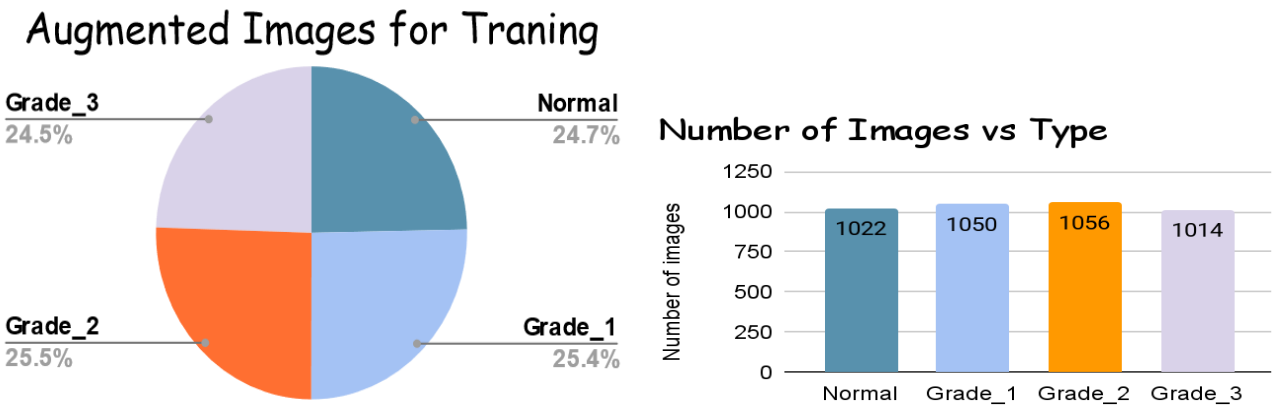


Figure 5.3: Balanced image data set after augmentation and over sampling.

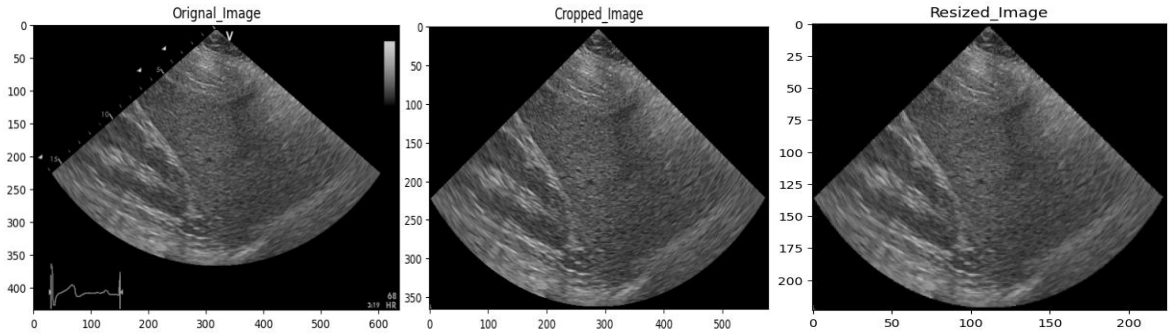
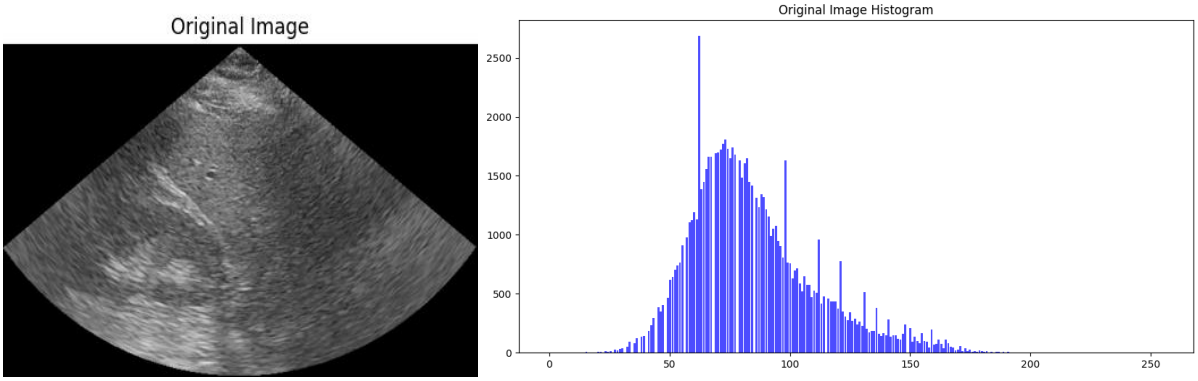


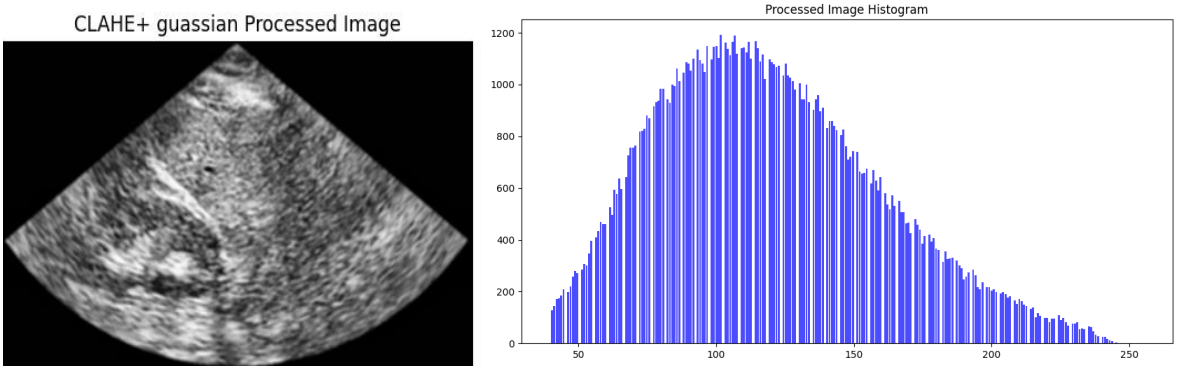
Figure 5.4: A typical pre-processed Ultrasound image.

5.2.1 Image Enhancement and Filtering

A Gaussian filtering and CLAHE was applied to the original images for use in image enhancement before they are inputted to the deep learning model for training. The contrast enhancement technique, CLAHE, has been effectively used in many image processing application. Same is true with the Gaussian filtering in dealing with US noise with focus on speckle noise being the major one. Figure 5.5 shows the effect of applying the Gaussian filtering and CLAHE on a template US image and its effect on the histogram. Clearly, the dynamic range of the image has significantly increased and the noise level has been smoothed out. Visually, the output image is much more enhanced than the input image.



(a) (b)



(c) (d)

Figure 5.5: Image pre-processing results: (a) Original image, (b) The Histogram plot, (C) Image after applying Gaussian filtering and CLAHE, and (d) The resulting Histogram plot.

5.3 Training Results for Multiclass Classification

The five pre-trained models, VGG16, ResNet50, MobileNetV2V2, Xception, and InceptionV3, were further tested for their efficacy in grading fatty liver cases into mild, moderate and severe (and normal cases used as controls) to pick the best model that could be used for FLD diagnostic and grading prediction. The CNN models were then retrained for parameters at the top layer to extract features of B-mode US images. The results again showed that the Xception model was found to be the best deep learning model based on the evaluation metrics. Training results for the five different pre-trained models are shown in Figure 5.6 - 5.10.

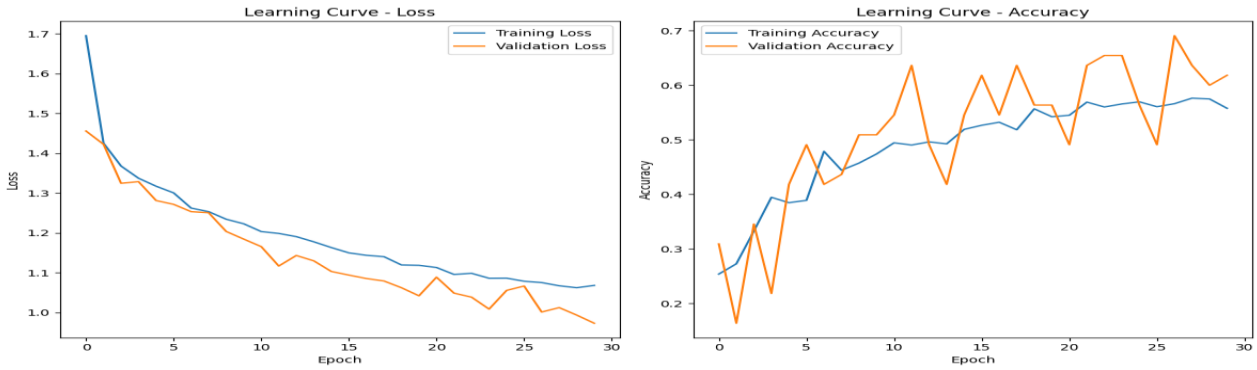


Figure 5.6: Training result of ResNet50.

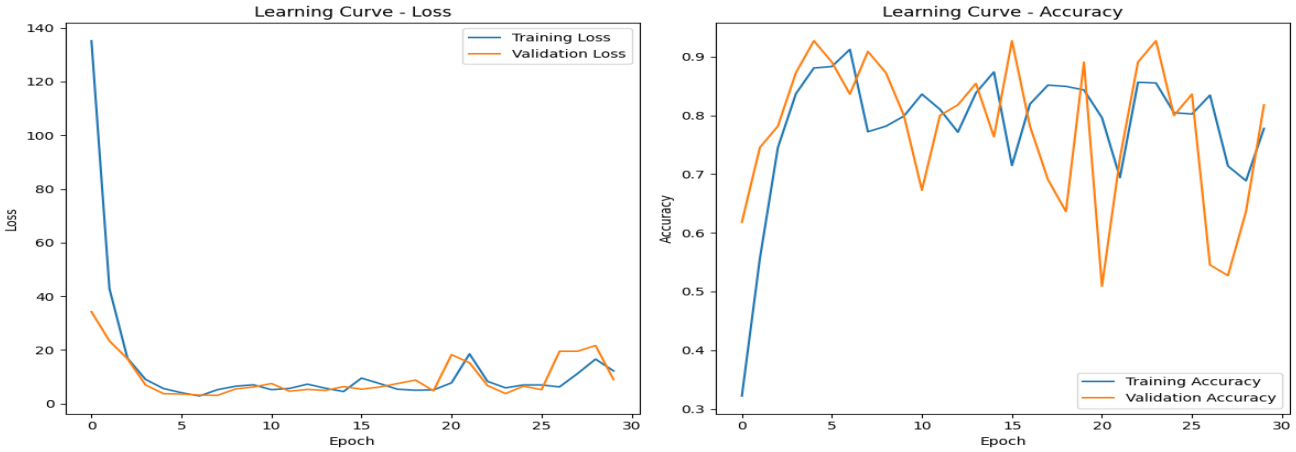


Figure 5.7 Training result of MobileNetV2

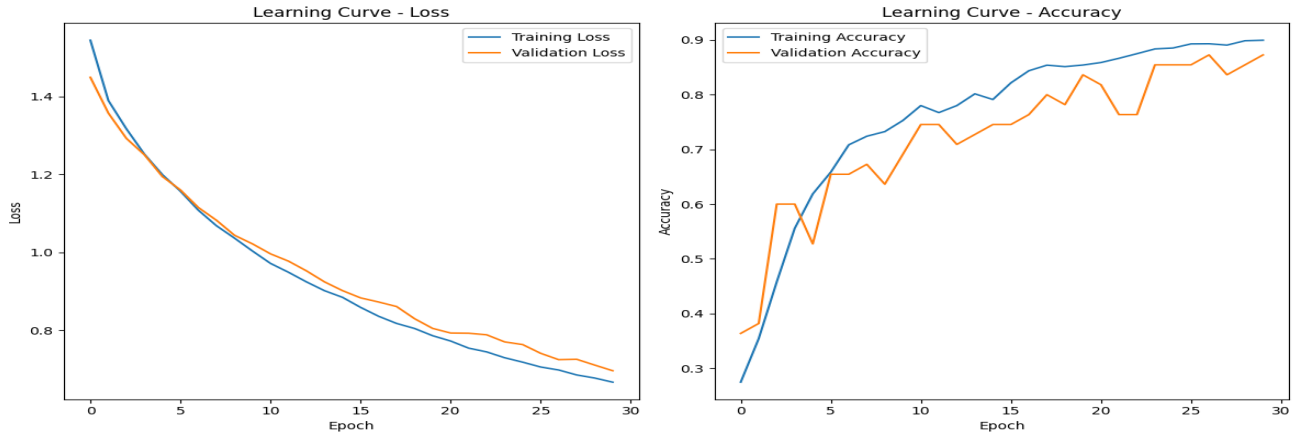


Figure 5.8: Training result of VGG16.

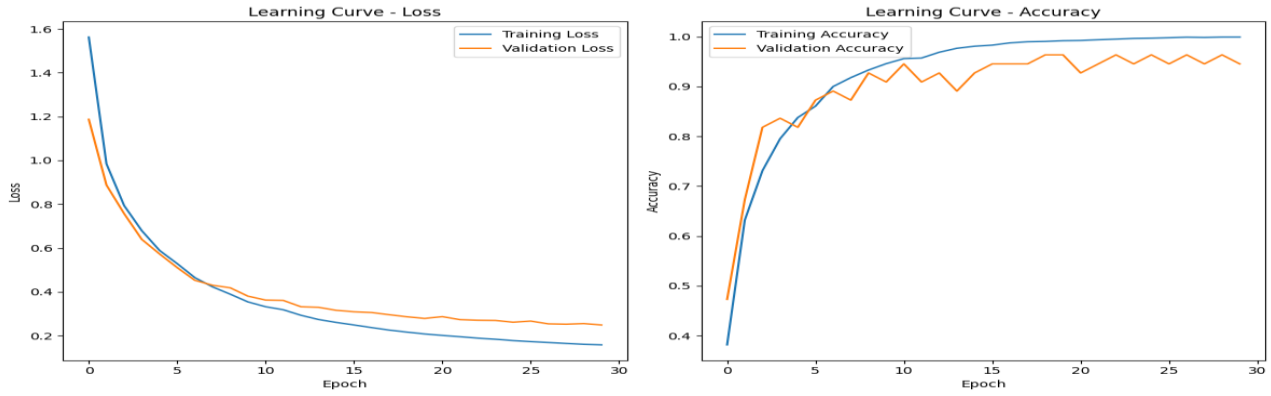


Figure 5.9: Training result of InceptionV3

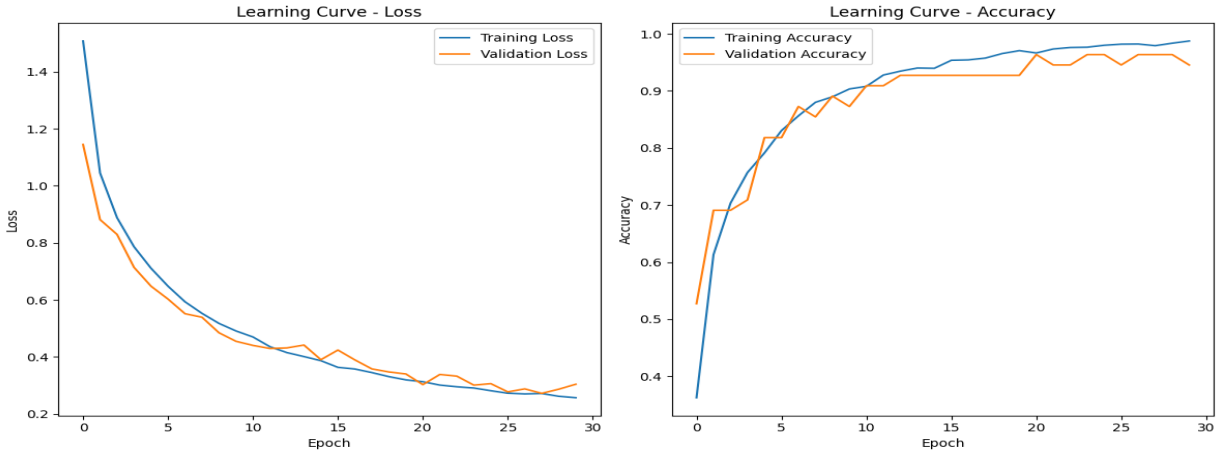
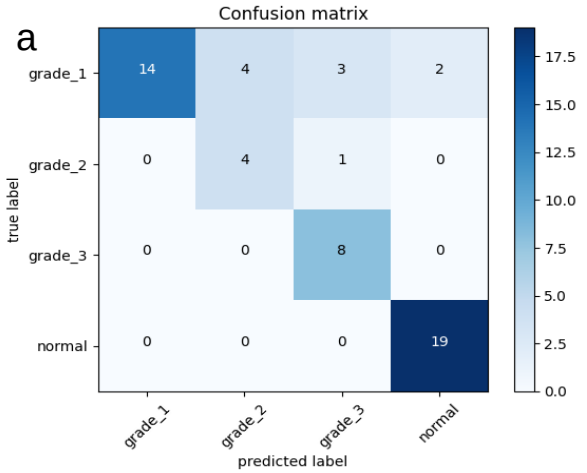


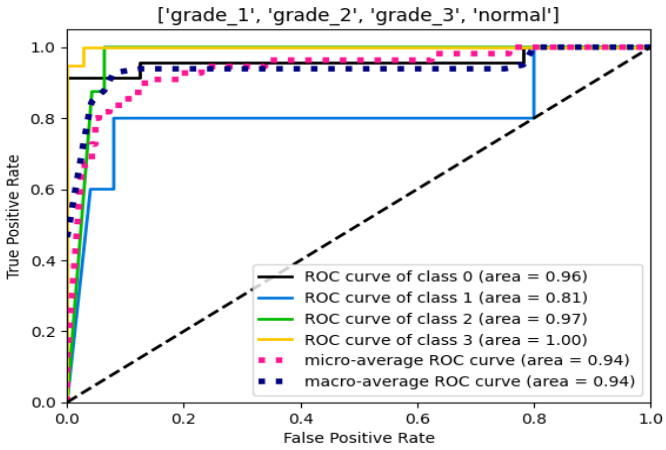
Figure 5.10: Training result of Xception.

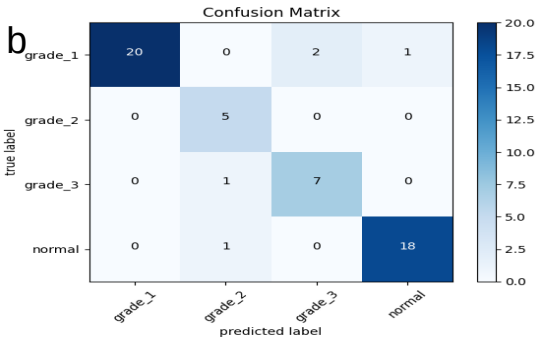
A multi classification model is constructed to classify US liver images into normal, mild, moderate and severe and fatty liver. The performance of the five deep models (VGG16, ResNet50, MobileNetV2, Xception, and InceptionV3) in differentiating normal and fatty livers was checked with different learning rates, batch sizes and epochs. Adams and RMSProp were checked for their optimization efficiency.

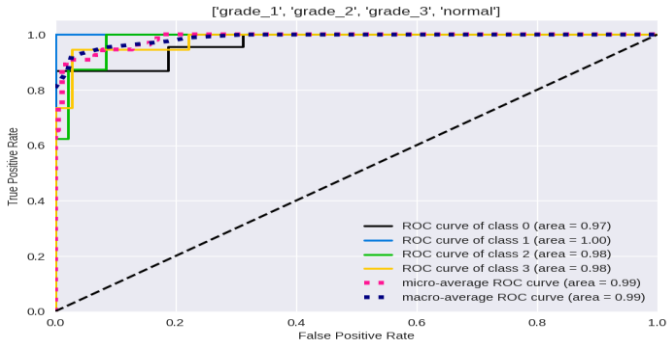
The results obtained for the MobileNetV2 model shows a training accuracy was 77.79% and validation accuracy was 81.82%. The validation accuracy is greater than the training accuracy which is a sign of under fitting. On the other hand, VGG16 attained a training and validation accuracy of 89.86% and 85.45%, respectively. Likewise, the InceptionV3 model trained for 30 epochs with batch size 256 offered a training accuracy and validation accuracy of 99.93% and 94.55% respectively. The worst performance was attained when using the ResNet50 showing an unsatisfactory generalization ability of the deep model. Overall, the pre-trained Xception model was found to be the best to classify fatty liver images. The model, offered a training accuracy of 98.74% and a validation accuracy of 94.55% showing a comparable performance with the InceptionV3 model.

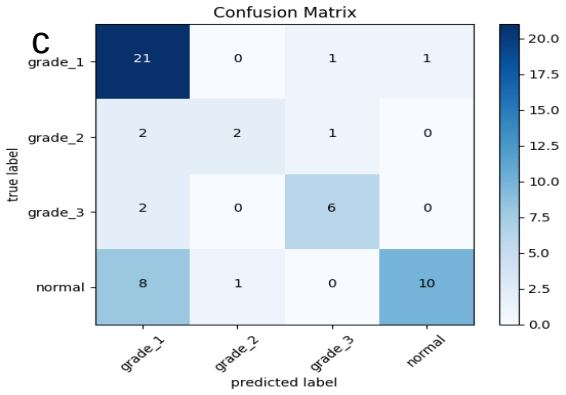
5.4 Test Results for Multiclass Classification

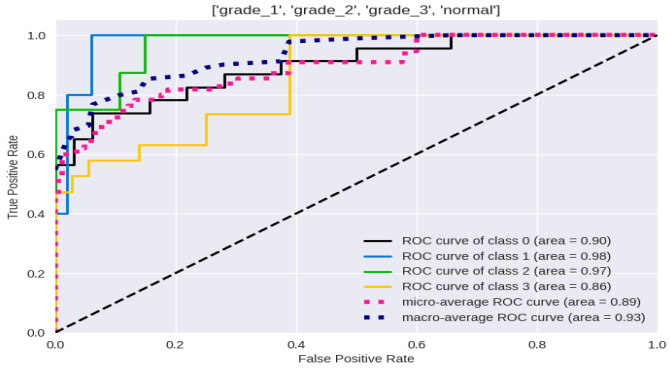












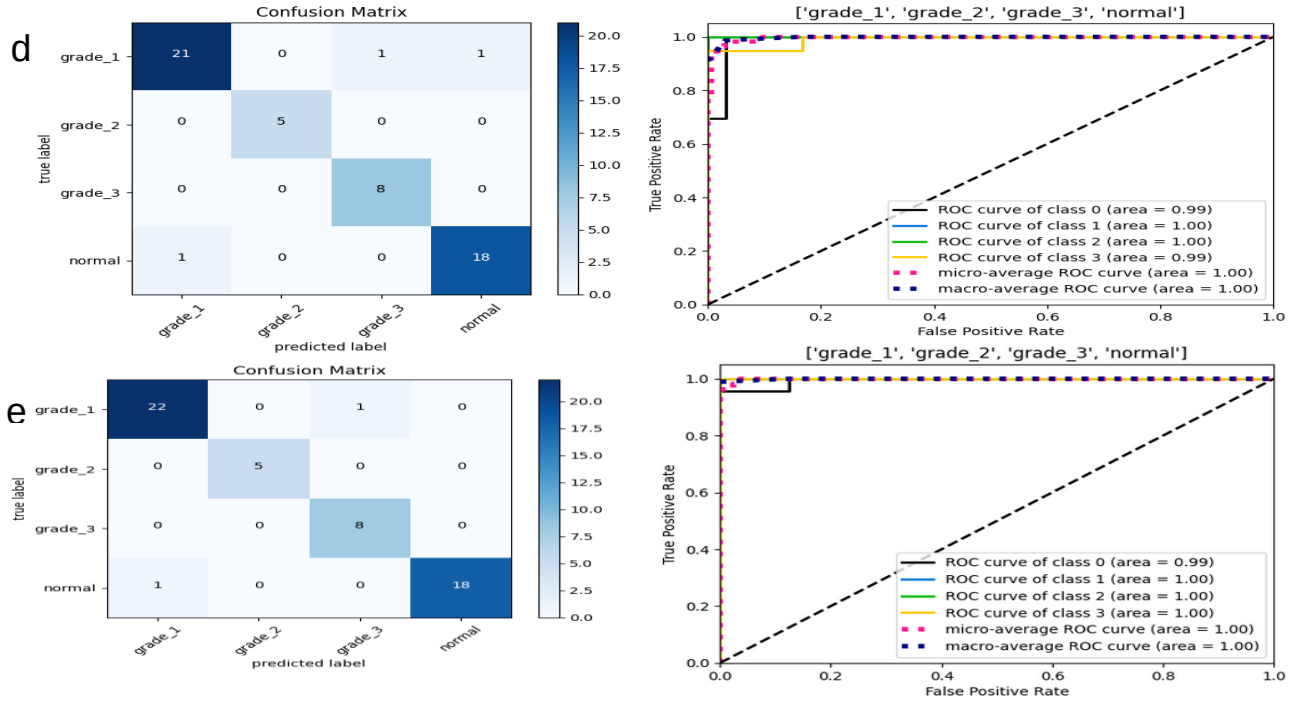


Figure 5.11: Comparison of confusion matrix and ROCAUC curves using different pertained models for multiclass classification: (a) MobileNetV2, (b) VGG16, (c) ResNet50, (d) InceptionV3 and (e) Xception and

	precision	recall	f1-score	support
grade_1	0.96	0.96	0.96	23
grade_2	1.00	1.00	1.00	5
grade_3	0.89	1.00	0.94	8
normal	1.00	0.95	0.97	19
accuracy			0.96	55
macro avg	0.96	0.98	0.97	55
weighted avg	0.97	0.96	0.96	55

Figure 5.12: Results obtained using the Xception model based on different metrics.

Figure 5.11 demonstrates the five deep models' final testing results in terms of the ROCAUC curves and confusion matrix. The obtained results support the robustness of the Xception model to detect liver steatosis with high accuracy, precision, recall, and F1-score. These results are presented in Figure 5.12. Table 5.1 presents a comparison between the five deep models in terms of their overall accuracy in correctly differentiating the three stages of FLDs (and the normal controls as well). As can be seen on the Table, the classification performance of the Xception model (96.36%) is the best than the rest four.

The second and third best performance were obtained when using InceptionV3 and VGG16. While the worst performance was obtained when using the ResNet50 deep model.

Model	Optimizer	Learning Rate	Batch Size	Performance
VGG16	Adam	0.0001	256	90.91%
ResNet50	Adam	0.0001	256	70.91%
InceptionV3	Adam	0.0001	256	94.55%
Xception	Adam	0.0001	256	96.36%
MobileNetV2	Adam	0.0001	256	81.82%

Table 5.1: Comparison of predictive performances of the different deep learning models.

5.5 Graphical user Interface (GUI)

To classify fatty liver and assess the severity level a GUI is designed for ease of use of the model. Figure 5.13 represent the layout of the GUI.

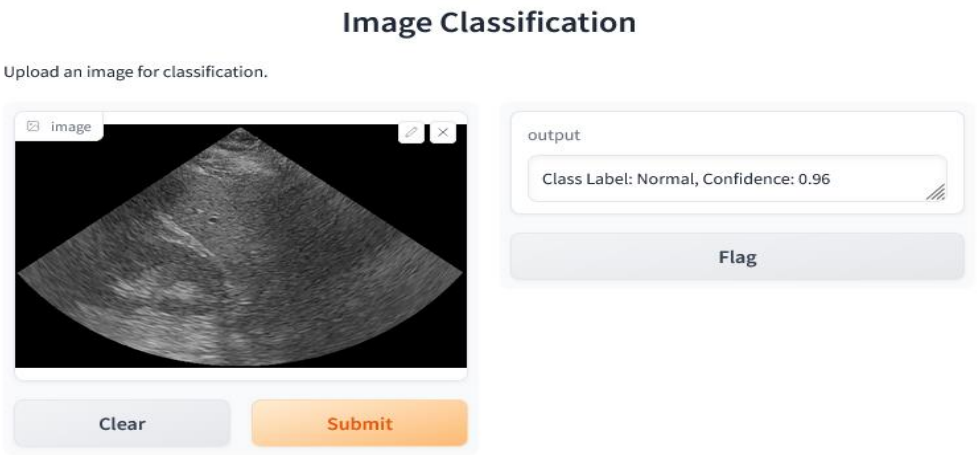


Figure 5.13: A representation of the developed Graphical User Interface.

5.6 Discussion

This study aims to classify FLDs and to grade the severity of FLDs using deep learning techniques. The study developed four-class prediction model to assess the severity level of fatty livers mild (G1),

moderate (G2), and severe (G3) steatosis including normal controls (G0). Certain pre-processing methods were applied to all images to have the best model of classification. Non-relevant parts were removed before applying CLAHE and Gaussian filtering. CLAHE is applied to enhance the image intensity to effectively mitigate the problems of contrast inconsistencies existing on the images. Further Gaussian filtering is applied to remove speckle noise while preserving the edges.

The four-class classification was done by comparing the five popular deep learning models (VGG16, ResNet50, InceptionV3, MobileNetV2 and Xception families). These deep learning models consist of CNN with a series of layers (convolutional, pooling, and fully connected) that can perform end-to-end supervised learning in order to identify patterns. In order to acquire competent accuracy, the current study applied data augmentation techniques and utilized oversampling to correct the issue of data imbalance. Further to improve the performance of the models, hyperparameters like the learning rate, optimizer, batch size, and number of epochs were fine-tuned. The study used a very low learning rate so that the newly connected layers can learn features from already learned convolutional layers. The models were trained with a very low learning rate of 0.0001. The model was trained by flattening the layers before feeding into the fully connected layers. Flatten layers serve as the bridge between these two types of layers by reshaping the multi-dimensional output from convolutional layers into a flat vector that can be fed into dense layers. In order to solve the overfitting problem, L2 regularization technique with weights of 0.01 was applied. The best performance was achieved when using the Xception model with a 96.36% accuracy during model testing. The best hyperparameters combination is found by training the softmax layer of Xception model with learning rate of 0.0001, Adam optimizer, 30 epochs and a batch size of 256. The softmax function is primarily used in multi-class classification problems, where it takes a vector of inputs and normalizes them into a probability distribution over the classes, with each element representing the probability of belonging to a specific class. The model's performance was evaluated using a confusion matrix, accuracy, recall, precision, AUC-ROC and F1 score in the case of both binary and multi-class classifications.

Compared to other results reported by other researchers in the literature, the study reported in this thesis improved the classification accuracy for fatty liver diseases. The work by Andrea et al. [46] that proposed a binary classification of liver images integrating multiple classifiers including ANN, SVM, and k-NN algorithms achieved an accuracy capped at 79.77%, limiting its clinical applications. Another study adopted three methods for FLD severity assessment: neural network, envelope signal and

grayscale values revealing the neural networks superior sensitivity and specificity. The results of a deep learning neural network were not taken into account because the work used a shallow network architecture. [48]. The method offered an accuracy of 92.0%, a result which could be considered quite promising but at the same time inferior to the results reported in the current study. Zamanian et al. [49] proposed a deep learning model by merging Inception-ResNetV2, GoogLeNet, AlexNet, and ResNet101 pre-trained models. The study is limited to binary classification (normal liver and fatty liver) without assessing the severity of the fatty liver. In another study, a transfer learning using two pre-trained networks, InceptionV3 and VGG16, was utilized for fatty liver classification achieving accuracies of 93.23% and 90.70%, respectively. This study was again limited to binary classification of liver US images [6]. In a related study, cascaded deep neural networks using segmentation techniques show great potential but come with computational complexity [52]. Despite the improved performances, challenges involving classification accuracy, steatosis level assessment, reducing speckle noise and class imbalance remain the major bottlenecks.

Compared to the aforementioned methods, the proposed approach demonstrates a promising accuracy of 96.36% when determining the severity of fatty liver disease, which is much higher than the other results reported in the literature. Overall, the methodology proposed in the current study significantly improved the US based FLDs classification performance. The developed GUI is meant for ease of use in implementing the proposed system.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Fatty liver disease (steatosis) is the most common liver disease among all liver aliments. Early diagnosis of FLDs is crucial to prevent advanced stages. Ultrasound imaging is the most well-known technique due to its use no ionizing radiation, low-cost and wide availability. However, the classification of FLDs from liver US images involves subjective visual assessment of texture features like homogeneity and echogenicity. Operators rely on US and pathological findings. However, the interpretation of marginal echogenicity details remains extensively argued among experienced radiologists. Considering the global burden of FLDs, there is an urgent need to have an accurate and operator-independent US diagnosis method.

The current work offered a method to precisely differentiate between fatty and normal livers and to categorize fatty liver severity into mild, moderate, and severe classes using normal samples as controls. The obtained results support the robustness of CNNs using transfer learning to classify liver steatosis with high accuracy and precision. Five popular deep learning models, namely VGG16, ResNet50, MobileNetV2, Xception, and InceptionV3, were tested for their efficiency in accurately predict FLDs. The overall classification performance of Xception proved to be superior to the rest. For multi-classification, precision of 96.25%, recall of 97.75%, F1 score of 96.75%, and overall accuracy of 96.36% was obtained. The effect of different hyperparameters including optimizer, batch size, epoch, and learning rate was investigated to improve the model performance. As a result, Adam optimizer with epoch 30, a batch size of 256, and learning rate of 0.0001 resulted in the best accuracy for the Xception model.

The key contribution of this thesis is utilizing B-mode US images to develop four-class classifications. Clinically, assessing the severity of liver fat is highly important for further treatment of the disease. The proposed model showed promising accuracy in predicting FLDs and their grades which can support physicians in their decision making. The design of an easy to use GUI also

addresses the systems usability. The proposed automated method can significantly contribute to fill some of the gaps in low resource settings, like in rural Ethiopia, where the majority of health facilities struggle with shortage of radiologists.

6.2 Recommendation

- It is strongly advised to assess the performance of the proposed approach using more extensive data set. Ensuring system robustness becomes imperative for potential clinical validation.
- It would also be helpful if the proposed system could be tested on US images acquired locally in hospitals/clinics to check its robustness. Of course, this requires the hospitals/clinics to have good data handling and labeling system for liver US images.
- The required computational infrastructure should also be taken into account if the suggested approach needs to be evaluated on a more extensive dataset. The GPU that can be accessed on the cloud system mostly offer limited space and time to run the algorithm and that could be an issue in the case we have exaggerated training times.
- It is highly recommended a collaborative effort between universities and medical facilities to collect datasets for researches. Establishing partnerships with medical institutions can significantly enrich the dataset by incorporating a wide range of patient profiles, ultrasound equipment variations, and clinical settings. This collaborative approach ensures the inclusion of diverse cases and contributes to the generalizability of the classification and grading model. Moreover, it facilitates access to a larger pool of expertise, allowing for a more comprehensive understanding of the complexities involved in diagnosing and grading fatty liver disease. The mutual exchange of knowledge between academia and medical practitioners will not only enhance the quality of the research but also promote a more holistic and practical application of the proposed classification system within the medical community.

REFERENCES

[1] R. S. Cotran, V. Kumar, N. Fausto, Nelso Fausto, S. L. Robbins, and A. K. Abbas, “Robbins and Cotran pathologic basis of disease | WorldCat.org.” Accessed: Sep. 01, 2023. [Online]. Available: https://www.worldcat.org/title/Robbins-and-Cotran-pathologic-basis-of-disease/oclc/57335913

[2] “Liver: Anatomy and Functions | Johns Hopkins Medicine.” Accessed: Sep. 01, 2023. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/liver-anatomy-and-functions

[3] B. T. Tesfaye, T. M. Feyissa, A. B. Workneh, E. K. Gudina, and M. A. Yizengaw, “Chronic Liver Disease in Ethiopia with a Particular Focus on the Etiological Spectrums: A Systematic Review and Meta-Analysis of Observational Studies,” *Can J Gastroenterol Hepatol*, vol. 2021, 2021, doi: 10.1155/2021/8740157.

[4] C. A. Matteoni, Z. M. Younossi, T. Gramlich, N. Boparai, Yao Chang Liu, and A. J. McCullough, “Nonalcoholic fatty liver disease: a spectrum of clinical and pathological severity,” *Gastroenterology*, vol. 116, no. 6, pp. 1413–1419, 1999, doi: 10.1016/S0016-5085(99)70506-8.

[5] U. Becker *et al.*, “Prediction of risk of liver disease by alcohol intake, sex, and age: A prospective population study,” *Hepatology*, vol. 23, no. 5, pp. 1025–1029, May 1996, doi: 10.1053/jhep.1996.v23.pm0008621128.

[6] E. C. Constantinescu *et al.*, “Transfer learning with pre-trained deep convolutional neural networks for the automatic assessment of liver steatosis in ultrasound images,” *Med Ultrason*, vol. 23, no. 2, pp. 135–139, 2021, doi: 10.11152/MU-2746.

[7] “What are the 5 stages of liver disease?” Accessed: Sep. 02, 2023. [Online]. Available: https://kauveryhospital.com/blog/liver-diseases/5-stages-of-liver-disease/

[8] E. E. Powell, V. W. S. Wong, and M. Rinella, “Non-alcoholic fatty liver disease,” *The Lancet*, vol. 397, no. 10290. Elsevier B.V., pp. 2212–2224, Jun. 05, 2021. doi: 10.1016/S0140-6736(20)32511-3.

[9] A. A. Mokdad *et al.*, “Liver cirrhosis mortality in 187 countries between 1980 and 2010: a systematic analysis,” *BMC Med*, vol. 12, no. 1, Sep. 2014, doi: 10.1186/S12916-014-0145-Y.

[10] S. K. Asrani, H. Devarbhavi, J. Eaton, and P. S. Kamath, “Burden of liver diseases in the world,” *J Hepatol*, vol. 70, no. 1, pp. 151–171, Jan. 2019, doi: 10.1016/J.JHEP.2018.09.014.

[11] G. Abebe, D. Ayanaw, T. Ayelgn Mengstie, G. Dessie, and T. Malik, “Assessment of fatty liver and its correlation with glycemic control in patients with type 2 diabetes mellitus attending Dessie Comprehensive Specialized Hospital, Northeast Ethiopia,” *SAGE Open Med*, vol. 10, 2022, doi: 10.1177/20503121221124762.

[12] “The four stages of Non-Alcoholic Fatty Liver Disease (NAFLD).” Accessed: Sep. 01, 2023. [Online]. Available: https://www.liverhealthuk.com/post/the-four-stages-of-nafld

[13] G. Ferraioli and L. B. S. Monteiro, “Ultrasound-based techniques for the diagnosis of liver steatosis,” *World J Gastroenterol*, vol. 25, no. 40, p. 6053, Oct. 2019, doi: 10.3748/WJG.V25.I40.6053.

[14] A. M. Pirmoazen, A. Khurana, A. El Kaffas, and A. Kamaya, “Quantitative ultrasound approaches for diagnosis and monitoring hepatic steatosis in nonalcoholic fatty liver disease,” *Theranostics*, vol. 10, no. 9, p. 4277, 2020, doi: 10.7150/THNO.40249.

[15] A. Qayyum, M. Nystrom, S. M. Noworolski, P. Chu, A. Mohanty, and R. Merriman, “MRI steatosis grading: development and initial validation of a color mapping system,” *AJR Am J Roentgenol*, vol. 198, no. 3, pp. 582–588, Mar. 2012, doi: 10.2214/AJR.11.6729.

[16] S. K. Han, S. K. Baik, and M. Y. Kim, “Non-alcoholic fatty liver disease: Definition and subtypes,” *Clin Mol Hepatol*, vol. 29, no. suppl, pp. S5–S16, Feb. 2023, doi: 10.3350/CMH.2022.0424.

[17] E. Cobbina and F. Akhlaghi, “Non-alcoholic fatty liver disease (NAFLD) - pathogenesis, classification, and effect on drug metabolizing enzymes and transporters,” *Drug Metab Rev*, vol. 49, no. 2, pp. 197–211, Apr. 2017, doi: 10.1080/03602532.2017.1293683.

[18] X. Guo, X. Yin, Z. Liu, and J. Wang, “Non-Alcoholic Fatty Liver Disease (NAFLD) Pathogenesis and Natural Products for Prevention and Treatment,” *Int J Mol Sci*, vol. 23, no. 24, Dec. 2022, doi: 10.3390/IJMS232415489.

[19] M. E. Rinella, “Nonalcoholic Fatty Liver Disease A Systematic Review,” 2015, doi: 10.1001/jama.2015.5370.

[20] A. J. S. Ribeiro, X. Yang, V. Patel, R. Madabushi, and D. G. Strauss, “Liver Microphysiological Systems for Predicting and Evaluating Drug Effects,” *Clin Pharmacol Ther*, vol. 106, no. 1, pp. 139–147, Jul. 2019, doi: 10.1002/CPT.1458.

[21] G. Bedogni *et al.*, “The Fatty Liver Index: a simple and accurate predictor of hepatic steatosis in the general population,” *BMC Gastroenterol*, vol. 6, p. 33, Nov. 2006, doi: 10.1186/1471-230X-6-33.

[22] A. Kotronen *et al.*, “Prediction of non-alcoholic fatty liver disease and liver fat using metabolic and genetic factors,” *Gastroenterology*, vol. 137, no. 3, pp. 865–872, 2009, doi: 10.1053/J.GASTRO.2009.06.005.

[23] T. C. F. Yip *et al.*, “Laboratory parameter-based machine learning model for excluding non-alcoholic fatty liver disease (NAFLD) in the general population,” *Aliment Pharmacol Ther*, vol. 46, no. 4, pp. 447–456, Aug. 2017, doi: 10.1111/APT.14172.

[24] J. K. Dyson, Q. M. Anstee, and S. McPherson, “Non-alcoholic fatty liver disease: a practical approach to diagnosis and staging,” *Frontline Gastroenterol*, vol. 5, no. 3, pp. 211–218, Jul. 2014, doi: 10.1136/flgastro-2013-100403.

[25] R. G. Knodell *et al.*, “Formulation and application of a numerical scoring system for assessing histological activity in asymptomatic chronic active hepatitis,” *Hepatology*, vol. 1, no. 5, pp. 431–435, 1981, doi: 10.1002/HEP.1840010511.

[26] V. Sharma and K. C. Juglan, “Ultrasound-based Classification of Fatty Liver Disease: A Review,” *J Phys Conf Ser*, vol. 1531, no. 1, p. 012033, May 2020, doi: 10.1088/1742-6596/1531/1/012033.

[27] F. Guan, P. Ton, S. Ge, and L. Zhao, “Anisotropic diffusion filtering for ultrasound speckle reduction,” *Sci China Technol Sci*, vol. 57, no. 3, pp. 607–614, 2014, doi: 10.1007/S11431-014-5483-7.

[28] R. H. Marshall, M. Eissa, E. I. Bluth, P. M. Gulotta, and N. K. Davis, “Hepatorenal index as an accurate, simple, and effective tool in screening for steatosis,” *AJR Am J Roentgenol*, vol. 199, no. 5, pp. 997–1002, Nov. 2012, doi: 10.2214/AJR.11.6677.

[29] S. Strauss, E. Gavish, P. Gottlieb, and L. Katsnelson, “Interobserver and intraobserver variability in the sonographic assessment of fatty liver,” *AJR Am J Roentgenol*, vol. 189, no. 6, p. 1449, Dec. 2007, doi: 10.2214/AJR.07.2123.

[30] M. Byra *et al.*, “Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images,” *Int J Comput Assist Radiol Surg*, vol. 13, no. 12, pp. 1895–1903, Dec. 2018, doi: 10.1007/S11548-018-1843-2.

[31] B. Palmentieri *et al.*, “The role of bright liver echo pattern on ultrasound B-mode examination in the diagnosis of liver steatosis,” *Dig Liver Dis*, vol. 38, no. 7, pp. 485–489, Jul. 2006, doi: 10.1016/J.DLD.2006.03.021.

[32] A. M. Pirmoazen, A. Khurana, A. El Kaffas, and A. Kamaya, “Quantitative ultrasound approaches for diagnosis and monitoring hepatic steatosis in nonalcoholic fatty liver disease,” *Theranostics*, vol. 10, no. 9. Ivyspring International Publisher, pp. 4277–4289, 2020. doi: 10.7150/thno.40249.

[33] P. R. S. Mendonça, P. Lamb, A. Kriston, K. Sasaki, M. Kudo, and D. V. Sahani, “Contrast-independent liver-fat quantification from spectral CT exams,” *Med Image Comput Comput Assist Interv*, vol. 16, no. Pt 1, pp. 324–331, 2013, doi: 10.1007/978-3-642-40811-3_41.

[34] F. H. Cassidy *et al.*, “Fatty liver disease: MR imaging techniques for the detection and quantification of liver steatosis,” *Radiographics*, vol. 29, no. 1, pp. 231–260, 2009, doi: 10.1148/RG.291075123.

[35] A. Tang *et al.*, “Nonalcoholic fatty liver disease: MR imaging of liver proton density fat fraction to assess hepatic steatosis,” *Radiology*, vol. 267, no. 2, pp. 422–431, May 2013, doi: 10.1148/RADIOL.12120896.

[36] H. C. Shin *et al.*, “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning,” *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016, doi: 10.1109/TMI.2016.2528162.

[37] F. M. Alshagathrh and M. S. Househ, “Artificial Intelligence for Detecting and Quantifying Fatty Liver in Ultrasound Images: A Systematic Review,” *Bioengineering*, vol. 9, no. 12, Dec. 2022, doi: 10.3390/BIOENGINEERING9120748.

[38] L. Nanni, S. Ghidoni, and S. Brahnam, “Handcrafted vs. non-handcrafted features for computer vision classification,” *Pattern Recognit*, vol. 71, pp. 158–172, Jan. 2017, doi: https://doi.org/10.1016/j.patcog.2017.05.025.

[39] L. Cai, J. Gao, and D. Zhao, “A review of the application of deep learning in medical image classification and segmentation,” *Ann Transl Med*, vol. 8, no. 11, pp. 713–713, Jun. 2020, doi: 10.21037/ATM.2020.02.44.

[40] J. Wang, H. Zhu, S. H. Wang, and Y. D. Zhang, “A Review of Deep Learning on Medical Image Analysis,” *Mobile Networks and Applications*, vol. 26, no. 1, pp. 351–380, Feb. 2021, doi: 10.1007/S11036-020-01672-7.

[41] S. Vento, B. Dzudzor, F. Cainelli, and K. Tachi, “Liver cirrhosis in sub-Saharan Africa: neglected, yet important,” *Lancet Glob Health*, vol. 6, no. 10, pp. e1060–e1061, Oct. 2018, doi: 10.1016/S2214-109X(18)30344-9.

[42] S. Erkabu, B. Demeke, H. Desallegn, and S. Getachew, “Liver Disease: A Retrospective Hospital Based Study in Addis Ababa-Ethiopia,” *Journal of Spleen and Liver Research*, vol. 1, no. 4, pp. 1–7, Aug. 2021, doi: 10.14302/ISSN.2578-2371.JSLR-21-3912.

[43] C. W. Spearman *et al.*, “Hepatocellular carcinoma: measures to improve the outlook in sub-Saharan Africa,” *Lancet Gastroenterol Hepatol*, vol. 7, no. 11, pp. 1036–1048, Nov. 2022, doi: 10.1016/S2468-1253(22)00041-3.

[44] G. P. Coral, A. dal P. Antunes, A. P. A. Serafini, F. B. Araujo, and A. A. De Mattos, “Liver biopsy: Importance of specimen size in the diagnosis and staging of chronic viral hepatitis,” *Rev Inst Med Trop Sao Paulo*, vol. 58, 2016, doi: 10.1590/S1678-9946201658010.

[45] C. W, A. X, C. L, L. C, Z. Q, and G. R, “Application of Deep Learning in Quantitative Analysis of 2-Dimensional Ultrasound Imaging of Nonalcoholic Fatty Liver Disease,” *J Ultrasound Med*, vol. 39, no. 1, 2020, doi: 10.1002/JUM.15070.

[46] A. Andrade, J. S. Silva, J. Santos, and P. Belo-Soares, “Classifier Approaches for Liver Steatosis using Ultrasound Images,” *Procedia Technology*, vol. 5, pp. 763–770, 2012, doi: 10.1016/j.protcy.2012.09.084.

[47] H. Zhu, Y. Liu, X. Gao, and L. Zhang, “Combined CNN and Pixel Feature Image for Fatty Liver Ultrasound Image Classification,” *Comput Math Methods Med*, vol. 2022, 2022, doi: 10.1155/2022/9385734.

[48] W. Cao, X. An, L. Cong, C. Lyu, Q. Zhou, and R. Guo, “Application of Deep Learning in Quantitative Analysis of 2-Dimensional Ultrasound Imaging of Nonalcoholic Fatty Liver Disease,” *J Ultrasound Med*, vol. 39, no. 1, pp. 51–59, Jan. 2020, doi: 10.1002/JUM.15070.

[49] H. Zamanian, A. Mostaar, P. Azadeh, and M. Ahmadi, “Implementation of Combinational Deep Learning Algorithm for Non-alcoholic Fatty Liver Classification in Ultrasound Images,” *J Biomed Phys Eng*, vol. 11, no. 1, p. 73, 2021, doi: 10.31661/JBPE.V0I0.2009-1180.

[50] R. Ribeiro and J. Sanches, “Fatty liver characterization and classification by ultrasound,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5524 LNCS, pp. 354–361, 2009, doi: 10.1007/978-3-642-02172-5_46.

[51] A. Gaber, H. A. Youness, A. Hamdy, H. M. Abdelaal, and A. M. Hassan, “Automatic Classification of Fatty Liver Disease Based on Supervised Learning and Genetic Algorithm,” *Applied Sciences 2022, Vol. 12, Page 521*, vol. 12, no. 1, p. 521, Jan. 2022, doi: 10.3390/APP12010521.

[52] S. Y. Rhyou and J. C. Yoo, “Cascaded Deep Learning Neural Network for Automated Liver Steatosis Diagnosis Using Ultrasound Images,” *Sensors 2021, Vol. 21, Page 5304*, vol. 21, no. 16, p. 5304, Aug. 2021, doi: 10.3390/S21165304.

[53] P. Kalinowski *et al.*, “Liver Function in Patients With Nonalcoholic Fatty Liver Disease Randomized to Roux-en-Y Gastric Bypass Versus Sleeve Gastrectomy: A Secondary Analysis of a Randomized Clinical Trial,” *Ann Surg*, vol. 266, no. 5, pp. 738–745, Nov. 2017, doi: 10.1097/SLA.0000000000002397.

[54] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, “Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition,” *Procedia Comput Sci*, vol. 116, pp. 523–529, Jan. 2017, doi: 10.1016/J.PROCS.2017.10.038.

[55] Smriti Sahu, Maheedhar Dubey, Mohammad Imroze Khan, and Jitendra Kumar, “Comparative Evaluation of Filters For Liver Ultrasound Image Enhancement | PDF | Medical Imaging | Filter (Signal Processing).” Accessed: Sep. 01, 2023. [Online]. Available: https://www.scribd.com/document/131246966/Comparative-Evaluation-of-Filters-for-Liver-Ultrasound-Image-Enhancement#

[56] S. X. Yang, “(PDF) Contrast Limited Adaptive Histogram Equalization Based Fusion in YIQ and HSI Color Spaces for Underwater Image Enhancement.” Accessed: Sep. 01, 2023. [Online]. Available: https://www.researchgate.net/publication/321879418_Contrast_Limited_Adaptive_Histogram_Equalization_Based_Fusion_in_YIQ_and_HSI_Color_Spaces_for_Underwater_Image_Enhancement

[57] A. Mikołajczyk and M. Grochowski, “Data augmentation for improving deep learning in image classification problem,” *2018 International Interdisciplinary PhD Workshop, IIPhDW 2018*, pp. 117–122, Jun. 2018, doi: 10.1109/IIPHDW.2018.8388338.

[58] “Clipping with CLAHE. Graphs show how CLAHE redistributes the mapped... | Download Scientific Diagram.” Accessed: Sep. 04, 2023. [Online]. Available:

https://www.researchgate.net/figure/Clipping-with-CLAHE-Graphs-show-how-CLAHE-redistributes-the-mapped-intensities-of-the_fig2_12331201

[59] Karen Simonyan∗ & Andrew Zisserman+, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION Karen," *American Journal of Health-System Pharmacy*, vol. 75, no. 6, pp. 398–406, 2018, Accessed: Sep. 01, 2023. [Online]. Available: https://click.endnote.com/viewer?doi=arxiv%3A1409.1556&token=WzQxNTM2OSwiYXJ4aXY6MTQwOS4xNTU2Il0.1ERLUjt9JINMPEYW74pwJFdB7PA

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2016, doi: 10.1109/CVPR.2016.90.

[61] C. Szegedy *et al.*, "Going Deeper with Convolutions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1–9, Sep. 2014, doi: 10.1109/CVPR.2015.7298594.

[62] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1800–1807, Nov. 2017, doi: 10.1109/CVPR.2017.195.

[63] "Xception Architectural Design - coding." Accessed: Sep. 01, 2023. [Online]. Available: https://stephan-osterburg.gitbook.io/coding/coding/ml-dl/tensorfow/ch3-xception/xception-architectural-design

[64] "Review: Xception — With Depthwise Separable Convolution, Better Than Inception-v3 (Image Classification) | by Sik-Ho Tsang | Towards Data Science." Accessed: Sep. 01, 2023. [Online]. Available: https://towardsdatascience.com/review-xception-with-depthwise-separable-convolution-better-than-inception-v3-image-dc967dd42568

[65] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Jan. 2018, doi: 10.1109/CVPR.2018.00474.

[66] K. Dong, C. Zhou, Y. Ruan, and Y. Li, "MobileNetV2 Model for Image Classification," in *Proceedings - 2020 2nd International Conference on Information Technology and Computer Application, ITCA 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 476–480. doi: 10.1109/ITCA52113.2020.00106.

[67] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/JOURNAL.PONE.0118432.

APPENDICES

Appendix A: Implementation of the code using Matlab for Preprocessing

```
myDir = uigetdir('*.m'); %gets directory
myFiles = dir(fullfile(myDir,'*.png')); %gets all files in struct
for k = 1:length(myFiles)
  baseFileName = myFiles(k).name;
  fullFileName = fullfile(myDir, baseFileName);
  fprintf(1, 'Now reading %s\n', fullFileName);
  [imData, Fs] = imread(fullFileName);

%myFiles.name
filename = myFiles.name;
img = imread(filename);
img = im2gray(img); % need to work on the right image
% use the actual image dimensions
[nrows,ncols,~] = size(img);
for row = 1:nrows
    for col = 1:ncols
        if row < (-0.77*col + 253.64)
            img(row,col) = 0;
        end
    end
end
[nrows,ncols,~] = size(img);
for row = 1:nrows
    for col = 1:ncols
        if row < (0.77*col - 238.31)
            img(row,col) = 0;
        end
    end
end
end
cropped=img(5:370,31:609);
imshow(img)
imshow(cropped)
```

Applying Augmentation on the preprocessed Image datasets

```
% Load the input images
images = imageDatastore('train_normal/');

% Define the imageDataAugmenter object with the desired augmentation techniques
augmenter = imageDataAugmenter('RandRotation',[-45 45],'RandXReflection',true,'RandYReflection',true);

% Define the output folder
outputFolder = 'aug_normal_v5';

% Apply data augmentation to each image in the dataset and save the original and augmented images to disk
for i = 1:numel(images.Files)
    % Read the original image
    img = readimage(images, i);

    % Save the original image to disk
    imwrite(img, fullfile(outputFolder, sprintf('normal%d.png', i)));

    % Generate six augmented versions of the original image
    for j = 1:7
        augmentedImage = augment(augmenter, img);
        % Save the augmented image to disk
        imwrite(augmentedImage, fullfile(outputFolder, sprintf('aug_normal%d_%d.png', i, j)));
    end
end
```

Appendix B: Implementation Code for Classification

```
# Importing all the necessary libraries
import PIL
print('Pillow Version:',PIL.__version__)
import matplotlib.pyplot as plt
import numpy as np
import os
import tensorflow as tf
import cv2
from skimage import exposure
from tensorflow import keras
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Activation, Dense, Flatten, BatchNormalization, Conv2D, MaxPool2D
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.metrics import categorical_crossentropy
from tensorflow.keras.preprocessing.image import ImageDataGenerator
from sklearn.metrics import confusion_matrix
import itertools
import warnings
import pickle
warnings.simplefilter(action='ignore', category=FutureWarning)
import pandas as pd
from glob import glob
import sklearn.metrics as metrics
from tensorflow.keras.applications.xception import Xception
from tensorflow.keras import regularizers
from sklearn.metrics import classification_report,roc_curve, roc_auc_score,auc
from sklearn.metrics import roc_curve, auc
# Mounting the google drive
from google.colab import drive
drive.mount('/content/drive')
train_dataset = '/content/drive/MyDrive/mat_grading FL_aug_v5/'
test_dataset = '/content/drive/MyDrive/testv3/'
validation_dataset = '/content/drive/MyDrive/validationv3/'
```

```
# applying image enhancement and filitering
def clahe_preprocessing(image):
    # convert image to uint8
    image = np.uint8(image * 255)

    # apply CLAHE
    clahe = exposure.equalize_adapthist(image, clip_limit=0.03)
    # apply guassian filter
    guassian = cv2.GaussianBlur(clahe, (5, 5), 2)
    return guassian
#  Preparing the dataset so as to feed to the model
train_batches = ImageDataGenerator(
    preprocessing_function=clahe_preprocessing,
).flow_from_directory(
    directory=train_dataset,
    target_size=(224, 224),
    class_mode='categorical',batch_size=256

)

valid_batches = ImageDataGenerator(preprocessing_function=clahe_preprocessing).flow_from_directory(
    directory=validation_dataset,
    target_size=(224, 224),
    class_mode='categorical',batch_size=256

)

test_batches = ImageDataGenerator(preprocessing_function=clahe_preprocessing).flow_from_directory(
    directory=test_dataset,
    target_size=(224, 224),
    class_mode='categorical',
    batch_size=256,
    shuffle=False
)
```

```
 # Visualizing the dataset
imgs, labels = next(train_batches)
def plotImages(images_arr):
    fig, axes = plt.subplots(1, 10, figsize=(20,20))
    axes = axes.flatten()
    for img, ax in zip( images_arr, axes):
        ax.imshow(img)
        ax.axis('off')
    plt.tight_layout()
    plt.show()
plotImages(imgs)
print(labels)
```

```
# Define image size and number of classes
img_width, img_height = 224, 224
num_classes = 4

# Define the Xception model
Xception_model = Xception(weights='imagenet', include_top=False, input_shape=(img_width, img_height, 3))

# Freeze the base model layers
for layer in Xception_model.layers:
    layer.trainable = False
from tensorflow.keras import regularizers

# Create a new model and add the base model as a layer
model = Sequential()
model.add(Xception_model)

# Add dense layers for classification
model.add(Flatten())

model.add(Dense(num_classes, activation='softmax', kernel_regularizer=regularizers.l2(0.01)))
# Compile the model
model.compile(optimizer=Adam(learning_rate=0.0001), loss='categorical_crossentropy', metrics=['accuracy'])
```

```
# Traning the model
history = model.fit(x=train_batches,
    steps_per_epoch=len(train_batches),
    validation_data=valid_batches,
    validation_steps=len(valid_batches),
    epochs=30)
```

```
# Save the traning model
model.save('/content/drive/MyDrive/xception_v100.h5')
# Save the history object using pickle
with open('/content/drive/MyDrive/xception_v100.pickle', 'wb') as file:
    pickle.dump(history.history, file)
# installing scikit to plot the validation and traning curve
! pip install scikit-plot
```

```
def show(pickleFile, insert_model):
  def secondRoc():
 # installing scikit plot
    ! pip install scikit-plot
    import scikitplot as skplt
    y_true_labels = test_batches.classes

    target_names =cm_plot_labels
    skplt.metrics.plot_roc_curve(y_true_labels, predictions,target_names)
    #return  plt.show()


  def curveRoc():
    # Make predictions on your test data using the loaded model
    y_pred = insert_model.predict_generator(test_batches)

    # Convert predicted probabilities to class labels
    y_pred_labels = np.argmax(y_pred, axis=1)

    # Extract true class labels from test_batches
    y_true_labels = test_batches.classes

    # Generate classification report
    class_names = list(test_batches.class_indices.keys())
    print(classification_report(y_true_labels, y_pred_labels, target_names=class_names))

    # Compute the ROC curve and AUC for each class
    fpr = dict()
    tpr = dict()
    roc_auc = dict()
    for i in range(len(class_names)):
        fpr[i], tpr[i], _ = roc_curve(y_true_labels == i, y_pred[:, i])
        roc_auc[i] = auc(fpr[i], tpr[i])
```

```
# Plot the ROC curve for each class
plt.figure()
colors = ['blue', 'green', 'red', 'orange', 'black', 'yellow'] # Change colors as per your need
for i in range(len(class_names)):
    plt.plot(fpr[i], tpr[i], color=colors[i], lw=2,
             label='ROC curve of class {0} (AUC = {1:0.2f})'
             ''.format(class_names[i], roc_auc[i]))

plt.plot([0, 1], [0, 1], 'k--', lw=2) # Plot the diagonal line
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Grade_1,Grade_2,Grade_3,Normal')
plt.legend(loc="lower right")
#return plt.show()

predictions = insert_model.predict(x=test_batches,verbose=0) #verbose=0(to get no output whenever we run the predictions)

cm = confusion_matrix(y_true=test_batches.classes, y_pred=np.argmax(predictions, axis=-1))
```

```
def plot_confusion_matrix(cm, classes,
                          normalize=False,
                          title='Confusion matrix',
                          cmap=plt.cm.Blues):
    """
    This function prints and plots the confusion matrix.
    Normalization can be applied by setting `normalize=True`.
    """
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    print(cm)

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, cm[i, j],
                 horizontalalignment="center",
                 color="white" if cm[i, j] > thresh else "black")

training_history = pickleFile
# Extract the relevant information from the training history
train_loss = training_history['loss']
val_loss = training_history['val_loss']
train_acc = training_history['accuracy']
val_acc = training_history['val_accuracy']
```

```
 # Plot the learning curve
 plt.figure(figsize=(12, 6))

 # Plot the training and validation loss
 plt.subplot(1, 2, 1)
 plt.plot(train_loss, label='Training Loss')
 plt.plot(val_loss, label='Validation Loss')
 plt.xlabel('Epoch')
 plt.ylabel('Loss')
 plt.title('Learning Curve - Loss')
 plt.legend()

 # Plot the training and validation accuracy
 plt.subplot(1, 2, 2)
 plt.plot(train_acc, label='Training Accuracy')
 plt.plot(val_acc, label='Validation Accuracy')
 plt.xlabel('Epoch')
 plt.ylabel('Accuracy')
 plt.title('Learning Curve - Accuracy')
 plt.legend()

 cm_plot_labels = ['grade_1','grade_2','grade_3','normal']


 # return (plt.tight_layout(),
 #          plt.show(),plot_confusion_matrix(cm=cm, classes=cm_plot_labels, title='Confusion Matrix'),plt.ylabel('true label'),
 # plt.xlabel('predicted label'),curveRoc())
 return (plt.tight_layout(),
         plt.show(),plot_confusion_matrix(cm=cm, classes=cm_plot_labels, title='Confusion Matrix'),plt.ylabel('true label'),
 plt.xlabel('predicted label'),curveRoc(),secondRoc())
show(pickleFile, insert_model)
```

Appendix C: Implementation Code for GUI

```
import gradio as gr
import tensorflow as tf
import numpy as np
from skimage import exposure
import cv2
from PIL import Image
import os
!pip install -q gradio torch torchvision

# Load the pre-trained model
model = tf.keras.models.load_model('/content/drive/MyDrive/xception_v100.h5')
class_labels = ["Grade 1", "Grade 2", "Grade 3", "Normal"]

def create_directory_if_not_exists(directory):
    if not os.path.exists(directory):
        os.makedirs(directory)

def clahe_preprocessing(image):
    image = np.uint8(image * 255)
    clahe = exposure.equalize_adapthist(image, clip_limit=0.03)
    guassian = cv2.GaussianBlur(clahe, (5, 5), 2)
    return guassian

def classify_image(image):
    image = Image.fromarray(image)
    image = image.convert("RGB")
    image = image.resize((224, 224))
    image_array = np.array(image)
    preprocessed_image = clahe_preprocessing(image_array)

    image_array = preprocessed_image
    image_array = tf.expand_dims(image_array, 0)
    predictions = model.predict(image_array)
    predicted_class = np.argmax(predictions[0])
    predicted_label = class_labels[predicted_class]
    confidence = predictions[0][predicted_class]

    # Create a string with all class predictions and probabilities
    class_predictions = ", ".join([f"{class_labels[i]}: {predictions[0][i]:.2f}" for i in range(len(class_labels))])
    return predicted_label, class_predictions
```

```python
def classify_images_in_directory(directory):
    full_directory_path = os.path.join('/content/drive/MyDrive', directory)
    output_directory = os.path.join(full_directory_path, 'predictions')
    create_directory_if_not_exists(output_directory)

    for filename in os.listdir(full_directory_path):
        if filename.endswith(".jpg") or filename.endswith(".jpeg") or filename.endswith(".png"):
            image_path = os.path.join(full_directory_path, filename)
            image = cv2.imread(image_path)
            predicted_label, predictions = classify_image(image)

            # Create a subdirectory for the predicted class if it doesn't exist
            class_directory = os.path.join(output_directory, predicted_label)
            create_directory_if_not_exists(class_directory)

            # Save the original image and predictions in the class subdirectory
            cv2.imwrite(os.path.join(class_directory, filename), image)
            with open(os.path.join(class_directory, f"{filename}_predictions.txt"), 'w') as f:
                f.write(predictions)

# Interface
iface = gr.Interface(fn=classify_images_in_directory,
                     inputs="text",
                     outputs="text",
                     title="Image Classification for Directory",
                     description="Enter the name of the directory in your Google Drive containing images for batch classification.")
iface.launch()
```