School of Art and Creative Technologies

University of Greater Manchester

Detection and Classification of Gallbladder Diseases through Lightweight Deep Learning Techniques

Student Name and ID: Bogomil Iliev (2011184)

Supervisor Name: Dr. Naveed Islam

Programme: MSc. Artificial Intelligence

Department: Computing

Institution: University of Greater Manchester

Date: 08/09/2025

University of Greater Manchester, Deane Road, Bolton. BL3 5AB

DECLARATION

This is to certify that this thesis titled the "**Detection and Classification of Gallbladder Diseases through Lightweight Deep Learning Techniques**" is entirely original with no submissions to other educational institutions and the sources utilised in this study have been properly cited and referenced.

ACKNOWLEDGEMENT

I want to extend my heartfelt gratitude to God for his blessings and support throughout this period.

I am sincerely grateful to my supervisor - Dr. Naveed Islam, for providing me with invaluable guidance, motivation, constructive criticism, and the necessary information to accomplish the objectives of this thesis.

A deep thankful wish goes to my fiancée, who supported me and motivated me through the whole studies, my mother and father for believing in me. The inspiration for this work is also coming from my mother as she is suffering from a form of Gallbladder disease, that has been diagnosed in a developed stage. Thus, the successful completion of this work is striving to provide means to help people get diagnosed while their condition is in its early stages of development.

ABSTRACT

Ultrasound (US) is the most utilised imaging modality for gallbladder disease (GBD), but diagnostical accuracy is linked to operator skill, scanner settings and patient factors. The current work investigates whether lightweight deep learning models, that are designed for real-time use on modest hardware, can reliably classify nine gallbladder pathologies from US samples while meeting important validation standards that reflect clinical deployment.

A reproducible pipeline is designed around three algorithms – ResNet-50 (serving as a baseline), TinyVit with 11 million parameters and GhostNet-1.0 (for benchmarking purposes). To improve robustness on the limited GBD US data that is frequently noisy, the input geometry is standardised, z-score normalisation is applied to each image, and QC filtering is applied. This is done together with conservative augmentations including Gaussian blur. Data leakage is controlled by patient-level grouping: a 5-fold stratified cross validation. Training employs linear probe phase followed by phased unfreezing with discriminative learning rates and early stopping on macro-F1. Class imbalances handling is also further tested with a weighted sampler. Model behaviour is examined with Grad-CAM++ and error analysis through confusion matrix and precision-recall curves.

Under strict patient-level splits, all models achieve modest macro-F1 scores, with clear signs of overfitting once deeper layers are unfrozen. Grad-CAM++ reveals attention to US artefacts alongside pathology, which explains the limited generalisation. A deliberately “easy” demonstration aiming at splitting of US data without patient controls, is produced to showcase much higher headline scores. Thus, confirming that leakage can inflate performance and mask true clinical value and difficulty.

The research concludes that on this multi-class task, methodological discipline (e.g. patient-level evaluation and training) is decisive to provide real generalisable results.

Keywords: Gallbladder Disease (GBD), Artificial Intelligence (AI), Deep Learning (DL), Ultrasound (US).

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI - Artificial Intelligence

ANN – Artificial Neural Network

AP – Average Precision

BA – Biliary Artesia

BBFL – Batch-Balanced Focal Loss

BN – Bayesian Network

CAD – Computer Aided Diagnostic

CEUS – Contrast-Enhanced Ultrasound

CH-EUS - Contrast-enhanced Endoscopic Ultrasound

CLAHE – Contrast Limited Adaptive Histogram Equalisation

CNN – Convolutional Neural Network

DL – Deep Learning

DNN -Deep Neural Network

EDA – Exploratory Data Analysis

EUS – Endoscopic Ultrasound

GAN – Generative Adversarial Networks

GB – Gallbladder

GBC – Gallbladder Cancer

GBD - Gallbladder Disease

HRUS – High-Resolution Ultrasound

ICC - Interclass Correlation Coefficient

LASSO - Least Absolute Shrinkage and Selection Operator

LOOCV - Leave-one-out Cross Validation

ML - Machine Learning

NLM - Non-Local Means

NN – Neural Network

PR-AUC - Precision-Recall Area

RNN – Recurrent Neural Network

ROI – Region of Interest

SGD - Stochastic Gradient Descent

SRAD - Speckle-reducing Anisotropic Diffusion

SVM – Support Vector Machine

TAN – Tree-Augmented Naïve Bayes

TUS - Transabdominal Ultrasonography

US – Ultrasound

1. INTRODUCTION

1.1. Background of Study

The gallbladder (GB) is a relatively minute body part that is situated under the liver (**Figure 1**), in the upper right quadrant of the digestive system. It resembles a small sack with a pear-like form. Its main functionality is to stockpile and condense bile, which is a liquid produced by the liver that aids the human body to break down the consumed fats. Thus, the GB introduces the bile into the small bowel to achieve the latter purpose (Housset *et al.*, 2016).

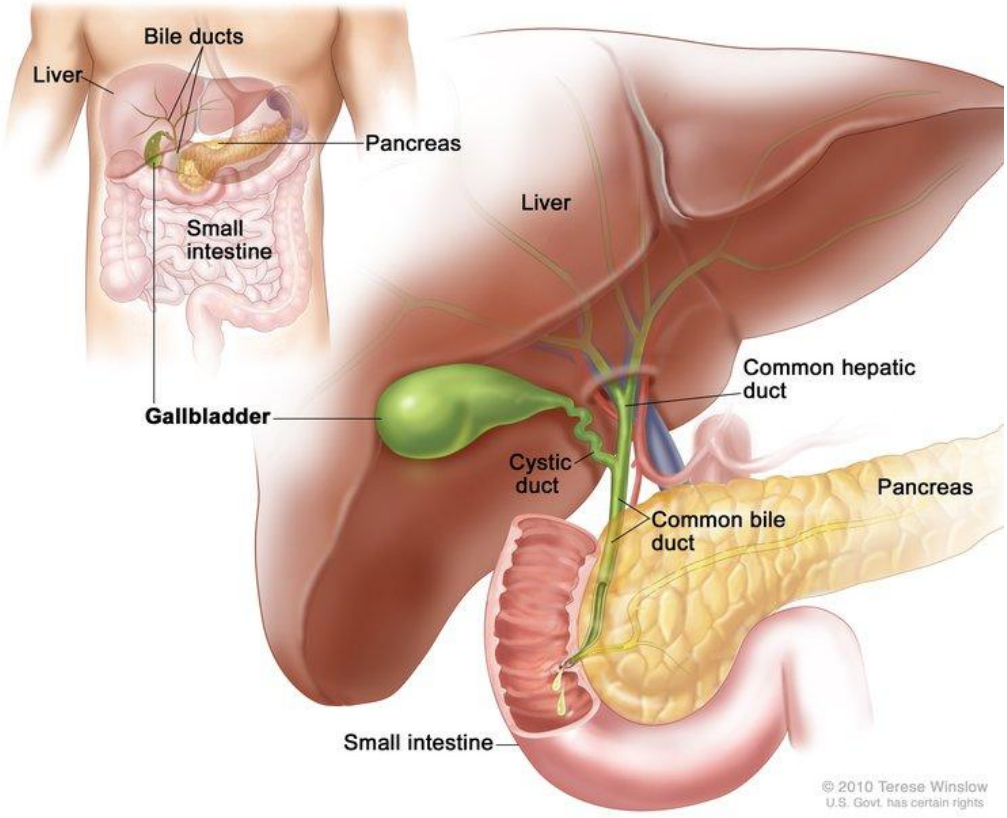


Figure 1 – Gallbladder Location in the Human Body (Winslow, 2010).

Gallbladder disease (GBD) confines a wide range of sub-pathologies that span from frequently encountered gall stones to lethal carcinomas. All of these conditions do exhibit close features in regard to complaints of the patients; however they do differ in the intensity with which they affect the GB's function. A contemporary work summarised that an estimated 6% of the world's population now harbour gall stones. More specifically females in the advancing age group, are highly affected in urbanised areas, with numbers on the rise. Ultrasound (US) is still the "go-to" medical approach for identifying issues with the organ due to its non-invasive nature, lack of screening radiation, ease of use, and cost effectiveness. Nonetheless, its diagnostic accuracy is tightly related to the skilfulness of the examiner and the particular variation of the US technology used. Such include transabdominal ultrasonography (TUS), high resolution ultrasound (HRUS), endoscopic ultrasound (EUS), contrast-enhanced EUS (CH-EUS), and more. TUS is the most commonly used one due to its high availability and non-invasive nature; however it lacks in resolution size and artefacts when compared to the others. HRUS provides better structural imaging definition, nonetheless its diagnostical accuracy is greatly impacted by the operator's skilfulness. The EUS introduces improved spatial resolution and is very useful in diagnosing abnormalities particularly in the cystic duct and neck lesions, which makes it handy in early cancer detection cases. However, it does need to be performed under sedation and by a highly technically trained expert, which prolongs the learning curve. CH-EUS uplifts vascular imaging and exemplifies higher specificity in highlighting malignancies, specifically in wall thickening and polypoid lesions (Wang *et al.*, 2024, 2025; Atlas University Hospital, 2022; Takahashi *et al.*, 2024).

Contemporary developments in the field of Computer Vision have proven effective as deep learning (DL) architectures exemplified accuracies surpassing the 90% mark, which is competitive to the human-level performance of professional sonographers in the field.

However, unlike liver or thyroid US, gallbladder Artificial Intelligence (AI) research remains sparse, and most studies aim at identifying one or two pathologies rather than the full clinical spectrum. These works also involve networks that are computationally heavy, require a lot of resources to train and have doubtful adaptability in existing hospital technological infrastructure. Thus, the scarce research in the field is partially caused by the fact that large enough datasets are not widely available for researchers to delve into the topic. All of the latter reasons outline the multi-class classification of gallbladder diseases as an underdeveloped research topic, especially in consideration of AI models that are lighter and can improve hardware requirements, costs, patient waiting times, reduce experts' learning curves and speed up appropriate patient treatment (Obaid *et al.*, 2023; Bozdag *et al.*, 2025; Wang *et al.*, 2025).

1.1.1. Typical Gallbladder Conditions

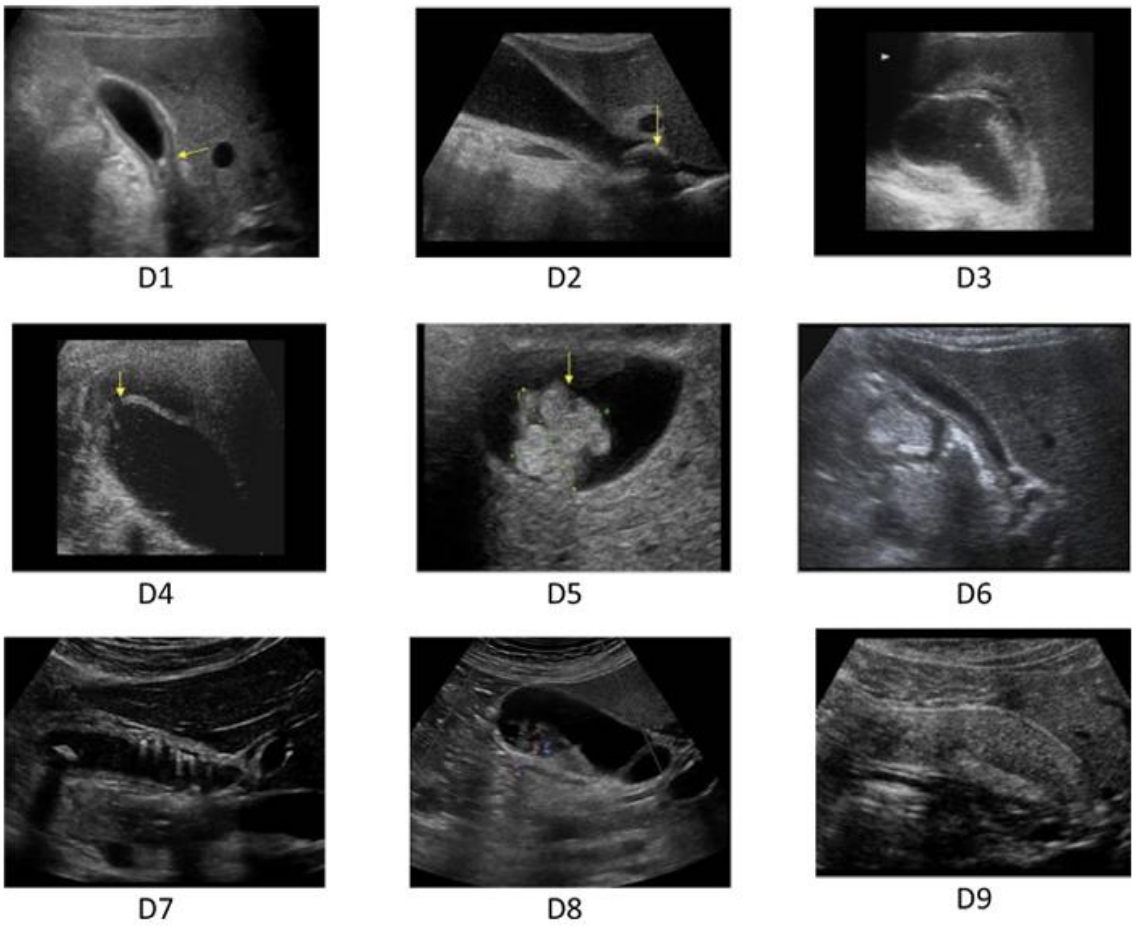
To make it easier for the non-medically educated reader, the GBD conditions (**Figure 2**) that are discussed in the current work are briefly explained below (Turki *et al.*, 2024):

- **Gallstones** – composed of cholesterol or bile salts, combined with calcium. They manifest as minute crystals in the organ, which take the shape of rigid matter formations with variable dimensions. Thus, they may impact the free flow of bile into the stomach by causing an obstruction in the GB duct. Such complications may result in the need of surgical intervention in severe cases.
- **Cholecystitis** – vexation and soreness of the latter body part caused by the occurrence and relocation of stones through the duct. Thus, inducing the bile liquid to be confined inside the organ, unable to reach the stomach when needed. As a result, it is frequent that a bacteria infestation will emerge which inflicts the GB's soreness and vexation.

Usually, such an illness would require immediate therapeutic attention or operative intervention.

- **Gangrenous Cholecystitis** – is a severe form of the previous condition, where the bodily part loses the free flow of blood to it, leading to tissue death. It is life-threatening and more likely in people with diabetes, heart disease, or weak immune functionality.
- **Perforation** – manifests when the organ's wall tears or breaks. Hence, it is one of the most dangerous GBD that may result in death if not speedily managed, especially in people with diabetes and poor immunity.
- **Polyps and Cholesterol Deposits** – appear as an accumulation on the inner layer of the body part, which is usually harmless, however if it spans over 10 mm in size it can turn into cancer.
- **Wall Thickening** – appears after repetitive soreness of the organ. Over time, this can cause deviations in the wall's depth and result in a carcinoma formation, especially when edged stones irritate the area.
- **Adenomyomatosis** – an illness that causes the inner lining of the bodily part to become overgrown and form small pockets in the wall. It is not fully understood as a condition yet but is more common in people over 50 and affects men and women equally.
- **Gallbladder Cancer** – it is a disease of uncommon occurrence. Estimated cases are about 1 in 100,000 people, where the average patient profile is a woman over 70 with pre-existing gallstone condition. Early discovery is crucial as it can migrate to other organs.

- **Intraabdominal and Retroperitoneum Issues** – refers to problems in the surrounding areas of the abdomen that may be linked to GBD, where the carcinoma has migrated.



Gallbladder diseases (D1:D9): (D1) Gallstones, (D2) Cholecystitis, (D3) Gangrenous Cholecystitis, (D4) Perforation of GB, (D5) Polyps and Cholesterol Crystals, (D6) Gallbladder-Wall Thickening, (D7) Adenomyomatosis of the GB, (D8) Carcinoma, (D9) Intraabdominal and Retroperitoneum problems.

Figure 2 – Gallbladder Diseases (Turki et al., 2024).

1.1.2. What is Artificial Intelligence?

Artificial Intelligence (AI) is a branch of computer science that emphasises on the manufacturing of systems that can carry out chores which are native to the human species and involve superior cognitive and reasoning potential. Such duties involve learning,

reasoning, problem-solving, perception, and language comprehension. Contemporary AI applications often include various subfields such as machine learning (ML), robotics, natural language processing, and computer vision to replicate human cognitive functionality (Russell *et al.*, 2022).

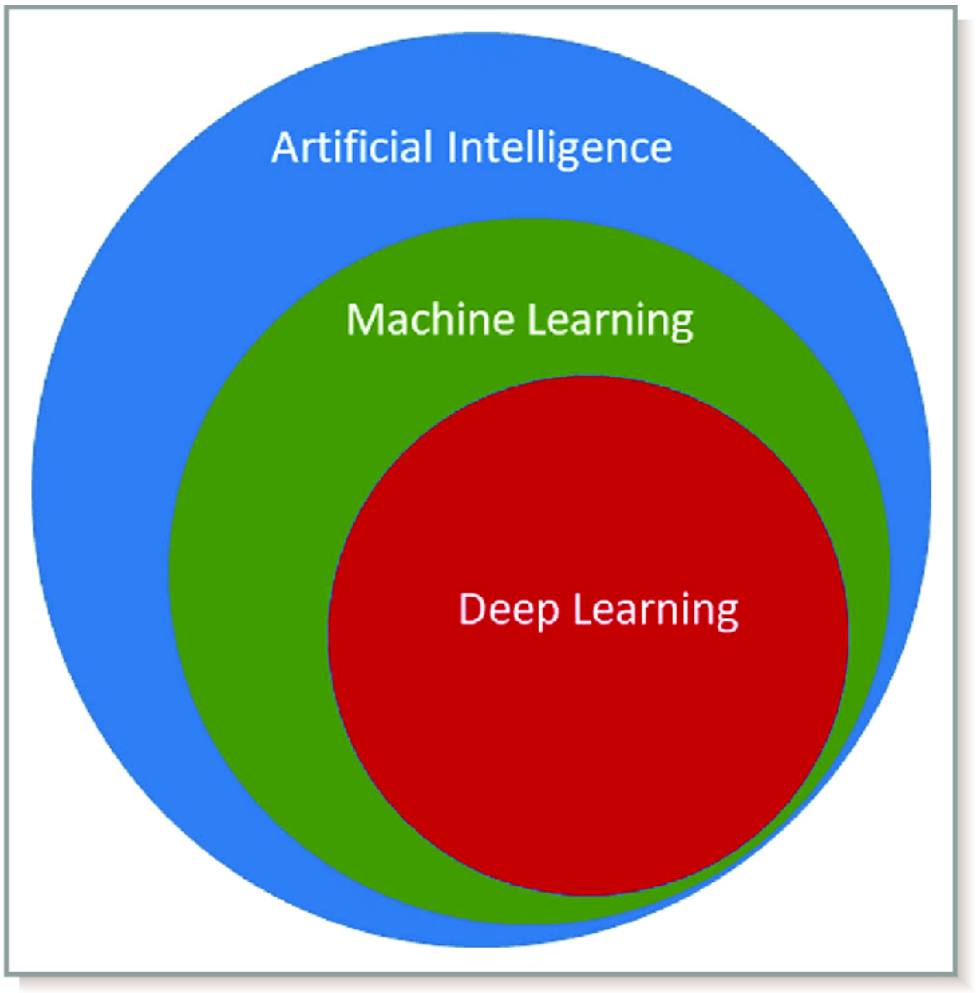


Figure 3 – How AI, ML and DL relate (Sevakula *et al.*, 2020).

1.1.3. What is Machine Learning?

Machine Learning (ML) is a subdomain of AI that allows systems to learn from experience or information without prior provided rules or instructions. ML architectures construct mathematical representations through data that they are provided with and target deliverables. They refine their performance with elapsed training time and as data becomes more available. ML (**Figure 4**) is generally sectioned into three major paradigms (Awad and Khanna, 2015):

- **Supervised Learning –** architectures are taught on labelled information to deliver predictions or classifications.
- **Unsupervised Learning –** models spot hidden structures or patterns in unlabelled data.
- **Reinforcement Learning –** optimises decision-making by interacting with an environment and learning based on a rewards and penalties system, where wrong decisions are penalised and correct ones rewarded.

In healthcare ML has provided pivotal advances in implementations that empower early disease detection, hospital appointment scheduling, analysis of electronic health records, classification of screening imagery, and more. However, the quality of its deliverables is often dependent on the quality of manually engineered features. Thus, its adoption in domains like image recognition can be negatively affected or performance impacted, because of the hierarchical and complex structure of patterns in such data (Lin *et al.*, 2024).

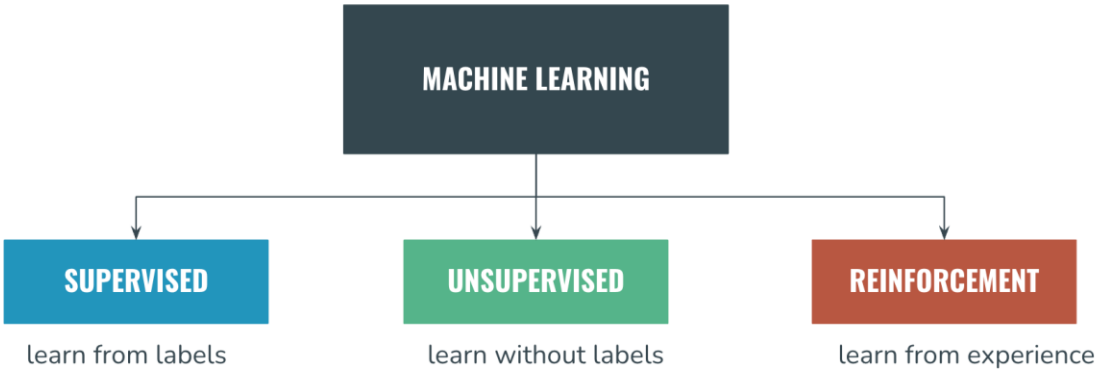


Figure 4 - Main Types of ML (ThinkAutonomous, 2024).

1.1.4. What is Deep Learning?

Deep Learning (DL) is a quickly evolving area of ML. It is based on Artificial Neural Networks (ANN) that consist of multiple layers (**Figure 5**). They are designed to replicate the neuron network of the human brain by employing nodes ordered in these layers. The latter are referred to as "deep layers". The aforementioned algorithms are able to grasp hierarchical and abstract representations from raw information that is fed to them. When compared to traditional ML, that commonly involves thorough feature engineering, DL algorithms are made to autonomously extract features from complex datasets such as images, video, and audio (Goodfellow *et al.*, 2016; Martin, 2019).

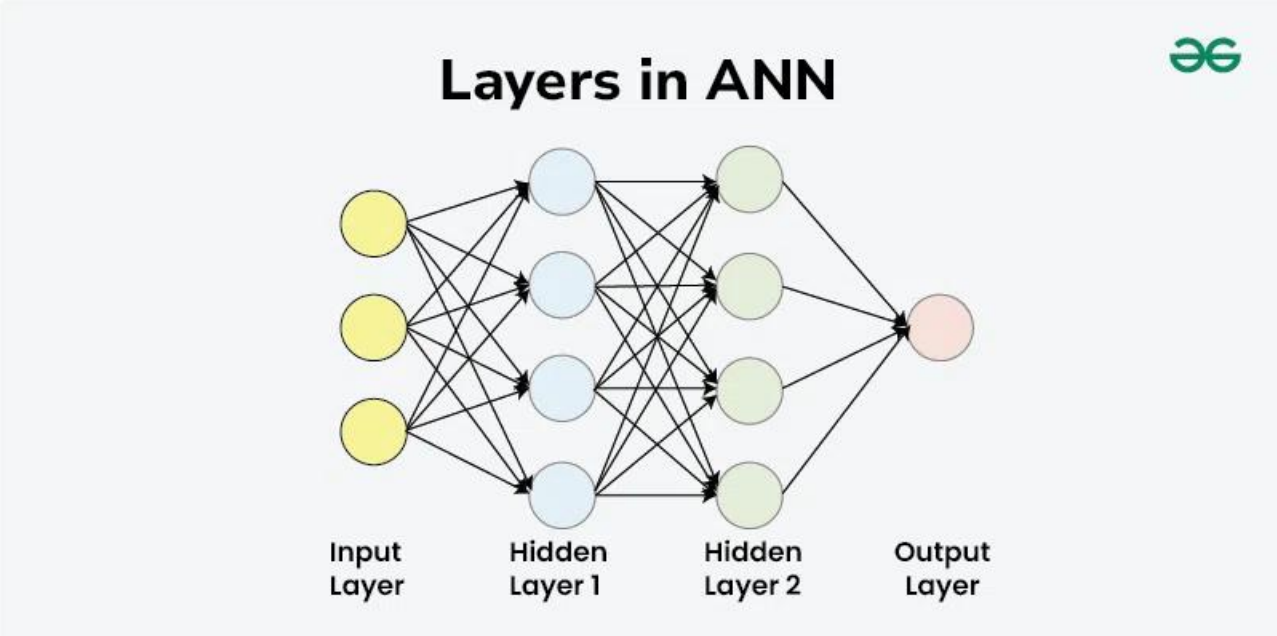


Figure 5 - Layers in ANN (GeeksforGeeks, 2025).

Some of the **most utilised deep learning approaches are** (Vasilev, 2019):

- **Convolutional Neural Networks (CNNs)** – primarily employed with image data. They assess spatial patterns by usage of filters and pooling operations.
- **Recurrent Neural Networks (RNNs)** – designed to work with sequential information.
- **Transformers** – created to work as natural language processing (NLP) algorithms, however research prove they are quite useful in image segmentation and multi-modal learning.
- **Generative Adversarial Networks (GANs)** – utilised to manufacture realistic artificial visual/audible output.

1.1.5. How Does Deep Learning Compare to Machine Learning?

Both DL and traditional ML architectures revolve around the common goal of learning from data. However, they differ significantly in the way they approach the tasks, scalability, performance and adaptability to similar or new tasks. ML models often rely on shallow architectures and depend heavily on manual feature designs, which can create bottlenecks in chores relating to unordered input such as pictures or audible content. On the other side, DL algorithms are able to learn from both low and high-level features directly from the raw information. Nonetheless, this is linked to the requirement of vast amounts of input to be provided. If this condition is met, the latter achieve a superior performance to ML models. DL networks are also more flexible in modelling relationships that are not linear and handling data that is of higher dimensions. However, this necessitates for the higher computing power they need when complete the task. Despite these challenges, DL has showcased deliverables across multiple domains that commonly outperforms ML techniques (Vasilev, 2019; Goodfellow *et al.*, 2016; Martin, 2019).

1.2. Problem Statement

Theoretically, a multi-class DL classifier could serve as an automated triage tool, highlighting high-risk scans for a quicker review, preventive diagnosis of GBDs in early stages, and appropriate patient plan forwarding. Practically, progress is throttled by two gaps (Obaid *et al.*, 2023):

1. **Data rarity and disparity** – publicly available collections of US images were historically limited to less than 2 000 shots, which comprised of selected pathologies. Thus, data was insufficient to appropriately train and benchmark AI architectures for the task of multi-class classification.

2. **Scarce external validity and interpretability** – developed models commonly fail to report patient-level cross validation and showcase little knowledge about whether the specific architecture adheres to the clinically relevant anatomy.

The recent release of the Gallbladder Diseases Dataset consisting of 10 692 high-resolution US pictures, that encompass nine distinct issues with the organ, sourced from four different hospitals – partially narrows the data gap but has not yet been systematically exploited or evaluated. No published work has combined this dataset with modern class-balancing losses and light weight DL algorithms that can be used on current clinically available ARM and computationally restricted devices. Thus, leaving unanswered questions about generalisability, clinical trust, cost-effectiveness, and possible adoption of the new advances (Turki *et al.*, 2024).

1.3. Aim and Objectives of the Research

1.3.1. Aim

The main goal of the current research is to investigate, develop and evaluate different light weight deep learning models capable of accurately detecting and classifying gallbladder diseases via ultrasound imaging data, thus providing a trustworthy approach for medical professionals in achieving rapid, accurate, low cost, and early diagnosis of GBDs. The emphasis here is on lightweight models that would be easily distributed over ARM and computationally restricted devices in hospitals, hence enabling real-time diagnostic assistance in clinical settings.

1.3.2. Objectives

The objectives of the current research are:

1. **Literature Review and Theoretical Framework**:
    - To review existing DL techniques in a critical manner that relate to gallbladder disease detection and classification.
    - To assess the limitations of the identified solutions in industry and academia, specifically in terms of resource complexity and requirements.
2. **Data Acquisition and Preprocessing:**
    - To obtain and explore the dataset from the related repository.
    - To analyse the class distribution of data and evaluate different techniques to address imbalances.
    - To perform data preprocessing, including normalisation, augmentation, and segmentation, and prepare the dataset for the training of the chosen models.
3. **Selection, Development, and Implementation of DL Models:**

- To investigate various lightweight architectures to outline the most suitable ones for the aforementioned medical imaging tasks.
- To highlight the most effective way to optimise the models for minimal computational requirements alongside the provision of high diagnostical accuracy.

4. **Performance Evaluation and Model Selection:**
    - Identify and apply performance metrics that correspond to the medical imaging tasks such as accuracy, sensitivity, precision, recall, and F1-score of developed models using validation techniques.
    - Conduct comparative analysis to highlight the best-performing lightweight model in terms of diagnostic accuracy and computational efficiency.
5. **Experimentation Settings:**
    - Assess the selected model's performance.
6. **Discussion and Recommendations:**
    - Critically discuss findings, highlight trade-offs between model complexity, accuracy, and resource constraints.
    - Provide recommendations for practical implementation and future work based on study outcomes.

1.4. Research Questions

The formulated research question related to the stated aims and objectives are:

1. **Literature and Theoretical Insights:**
    - What are the existing DL methodologies that have successfully been developed in detecting and classifying GBDs?
    - What are their limitations?
    - What is their resource effectiveness?
2. **Data Preparation:**
    - What is the data distribution in the nine-class GBDs dataset?
    - What are the appropriate techniques available to address possible class imbalances?
    - How does the preprocessing of the data influence the performance and generalisability of lightweight DL architectures?
    - Which pre-processing techniques are a good fit for the tasks?
3. **Model Architecture and Optimisation:**
    - Which lightweight DL models provide optimal balance between computational efficiency and diagnostic accuracy when applied to GBD detection and classification?
    - What optimisation techniques improve computational efficiency of DL architectures without compromising on diagnostic accuracy?
4. **Performance Evaluation:**
    - Which are the performance metrics that are most suited for assessing the performance of the models?

- Which are the best validation techniques for the models?
- What are the computational cost-benefit trade-offs between accuracy and resource efficiency among the architectures?

5. **Experiments:**
    - What experiments can be performed to analyse the chosen models' performance practically?
    - Which architecture outperforms the others?
6. **Discussions, Recommendations and Future Work:**
    - What knowledge insights can be drawn from the research?
    - What best practices and recommendations can be formulated based on this research to guide future implementations of lightweight AI models in medical diagnostics of GBDs?
    - How does the developed model compare to current research?
    - Can lightweight DL models deliver comparable results to larger networks in the field?

1.5. Research Contribution to Knowledge

The research presented in this work aims to make significant contributions to the current body of knowledge in both medical diagnostics and computationally efficient artificial intelligence. Thus, the contributions can be outlined as follows:

1. Emphasises on the gap identified in the computational complexity involved in the development of deep learning models that are utilised for medical imaging analysis, and more specifically in detecting and classifying gallbladder diseases. By focusing on lightweight architectures, the current work strives to deliver insights into maintaining high diagnostic accuracy without the employment of expensive hardware resources.
2. The study evaluates and applies knowledge regarding model optimisation approaches suitable for resource-limited medical devices. Hence, providing valuable guidance for adopting DL solutions in ARM-based or similar infrastructures. In such a manner, real time diagnosis of the latter diseases can be uplifted and its time frames reduced. Also, allowing for the current medical devices in use to be upgraded.
3. Such advancements can improve the patient's waiting times and reduce the medical employee's working load. The expertise level of inexperienced staff members can also be improved via such AI approaches, reducing their diagnostical learning curve.
4. The work also contributes to multiple stakeholders such as healthcare professionals and providers, patients, medical device manufacturers, and the wider AI research community.

1.6. Scope and Limitations of this Study

The scope of this study involves the development, evaluation, and optimisations of lightweight DL models for detecting and classifying GBDs from US imaging data. The focus remains on computational efficiency in order for the easier implementation of the solution to the current medical devices present in clinical settings or via its utilisation on ARM devices.

Nonetheless, the work has **several limitations**. Firstly, the data utilised is **the only available dataset for multiple classes of GBDs**. This can potentially limit its accuracy or be insufficient to achieve state-of-the-art performance. Following, there are **time constraints** that may impact the development of a solution that achieves such result. Thirdly, there are **limited computational resources** available to train, evaluate and experiment on the model performance. Lastly, the deployment and evaluation phases of this study are limited to **simulated experiments and environments rather than clinical trials**, hence a real-world testing would not be possible.

1.7. Thesis Organisation

The current research is organised in the following manner:

1. **Chapter Two** delivers a comprehensive literature review to investigate the existing DL techniques in a critical manner that relate to gallbladder disease detection and classification, assess their limitations of in terms of resource complexity and requirements.
2. **Chapter Three** completes the following tasks:
    - Highlights the appropriate techniques to address possible class imbalances in the data; outlines pre-processing approaches suitable for the lightweight

models; identifies lightweight DL models with optimal balance between computational efficiency and diagnostic accuracy.

- Locates appropriate optimisation techniques and training strategies.
- Identifies performance metrics for assessing the performance of the models.
- Highlights suitable validation approaches.
- Identifies appropriate experiment plan to justify the architecture's performance.
- Outlines the plan for this project's methodology, which research philosophy and design are applied and what experiments are conducted. Practical considerations and theoretical implications are also being discussed.

3. **Chapter Four** – introduces the training process that the models undertake, what their deliverables are, and which is the best performing model from the selection.
4. **Chapter Five** – discusses the experiments conducted on the best performing model to validate its results, and analyses them.
5. **Chapter Six** – provides a discussion of what knowledge was obtained by the current research.
6. **Chapter Seven** – concludes the work by highlighting all the acquired conclusions and offering further chances of improving the work.

2. LITERATURE REVIEW

The current chapter reviews and assesses the previous academic and industrial attempts that have been made to detect and classify multiple GB pathologies, and the related binary classification approaches of a single disease of the latter organ. Through the years, there have been a number of attempts to address the aforementioned problem, however most of them revolve around classifying a single pathology into benign or malignant. There are only two known attempts trying to tackle multi-class classification up to date. Thus, the current review is structured in a thematical (topical) manner and each work's relevance is graded in accordance with its relevance to the problem definition of this research. The chapter also presents a clear outline of the Methodology that was used for the Literature review, what were the inclusion and exclusion criteria, and draws insights and conclusions regarding the current body of knowledge on the topic.

2.1. Methodology of the Literature Review

2.1.1. Search Strategy of Literature

The latter is portrayed in **Figure 6,** where a **PRISMA diagram** has been drawn. Google Scholar and the Discover@Bolton Library databases were thoroughly used to uncover related work to detection and multi-class classification of GBDs. The **keywords** that were utilised in different combinations for the search are:

- **Deep Learning**
- **Machine Learning**
- **Artificial Intelligence**
- **Gallbladder Disease**

- **Ultrasound**

The reviewed research papers spanned from January 2015 to July 2025, and where the above keywords were found present in their titles, abstracts or conclusions, the reviewed work was assessed for relevance and included or excluded from the Literature Review.

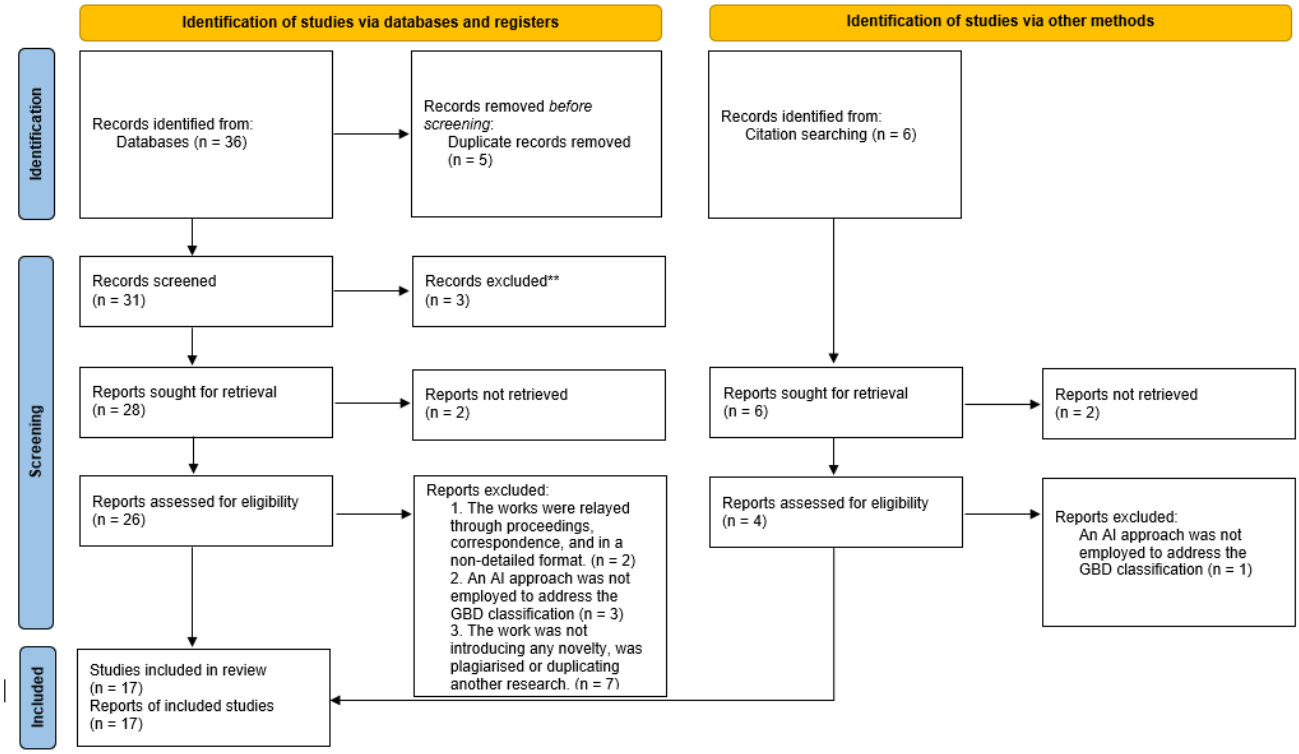


Figure 6 - PRISMA Diagram for Literature Review Research Methodology Process.

2.1.2. Inclusion Criteria

The thresholds to incorporate a research paper in this study are:

- **An application of AI (ML or DL) approach for the purpose of detection and diagnosis of GBDs through US imagery.**

- **The outcomes of the research included and justified evaluation metrics of the AI application** (Accuracy, Precision, F1 Recall, AUC, Sensitivity, Specificity).

2.1.3. Exclusion Criteria

The thresholds to avoid incorporation are:

- **The works were relayed through proceedings, correspondence, and in a non-detailed format.**
- **An AI approach was not employed to address the GBD classification.**
- **The work was not introducing any novelty, was plagiarised or duplicating another research.**

2.2. Related Academic Research

Before the literature is thoroughly reviewed in the below paragraphs, it is important to briefly discuss **two application fields** that are mentioned in the latter (Wang *et al.*, 2025):

- **Computer Aided Diagnostic (CAD)** – an AI application approach that aids medical experts in confirming the patient's condition. Such systems can assess both medical screening pictures, test data and patient's complaints to deliver the supposed diagnose through AI classification and/or propose a suitable treatment plan.

- **Radiomics** – focuses on analysing medical images by acquiring detailed information from them. It works by automatically collecting a large number of features from specific areas of the shot, known as regions of interest (ROIs). The latter are then thoroughly assessed through statistical and data mining approaches to highlight the most relevant information. This method provides a more objective and data-driven way to uplift diagnosing, classifying and grading diseases.

Since the detection and classification of multiple GBDs is an underdeveloped topic, the **existing academic research** revolves mainly around binary classification of different GBD pathologies. Thus, the aforementioned are **grouped** in the following subsections and in **Table 1**, where a concise summary of the reviewed papers is presented.

Table 1 - Literature Review Summary. Modified from (Wang et al., 2025)

Theme	Research Work	AI Technology Used	Relatedness Rating	Reported Patient Population	Reported Results	Merits	Demerits
Multi-class Classification	(Obaid et al., 2023)	DL- MobileNet	5	1782	Accuracy of 98.35	Tackles nine gallbladder diseases at once on a large, balanced ultrasound dataset collected across multiple institutions, with best performance from a mobile-friendly backbone (MobileNet). The enhancement/segmentation steps are designed to mitigate ultrasound noise and improve ROI quality before classification.	No parameter counts or compute benchmarks; No external/multicentre validation;
	(Bozdag et al., 2025)	DL - CBIR - GoogLeNet + InceptionV3 + NASNet-Large	5	1782	Overall Average Precision of 0.94	Avoids the training cost of large CNNs; ran on CPU with standard tools. Handles nine clinically relevant classes on a large dataset. Methods and per-class AP are reported with P-R curves, making behavior by pathology transparent.	Authors explicitly note the need to test on other centers/datasets. No sensitivity/specificity/AUC against a clinical ground truth or radiologist comparison. No parameter counts or runtime per image are provided;

GB Polyps	(Jeong et al., 2020)	DL - Inception V3 CNN	4	535	0.92 AUC	Improves diagnostic consistency across radiologists of varying experience Reduces false positives and potentially avoids unnecessary cholecystectomies Demonstrated high specificity and AUC, especially valuable in ambiguous cases with overlapping size ranges.	Single-center and retrospective study design, Lack of multicenter and prospective validation, Did not address detection of malignant polyps specifically due to limited malignant cases, The model was not compared against more recent architectures (e.g., DenseNet, NASNet).
	(Kim et al., 2021)	DL - Custom CNN	4	501	87% Accuracy, 0.88 Specificity, 0.90 AUC	Developed a custom lightweight CNN, avoiding overfitting on a moderately sized dataset. Achieved high accuracy and AUC, demonstrating feasibility of DL in gallbladder polyp classification. Used actual clinical US images, increasing real-world relevance. May aid in reducing unnecessary surgeries and improve risk stratification for cholecystectomy.	Single-center study with retrospective design. No external validation or multi-institutional data included. Model trained only on static images, not real-time video or longitudinal data. Model explainability was not implemented, limiting clinical interpretability. Differences in scanner models or imaging protocols may affect generalisation.
	(Choi et al., 2023)	DL - Custom Architecture	4	263	0.944 AUC, Accuracy 0.858, Sensitivity 0.856, Specificity 0.861	Better AUC than prior US studies; adds value when combined with conventional risk factors. First study (here) to show CAD improves diagnostic accuracy across several physicians, particularly aiding the inexperienced group.	Single-center data only - performance may vary across sites, thus needs large-scale validation. Ultrasound variability and operator dependence make standardisation hard.
	(Li et al., 2023)	ML - Bayesian Network	4	1,296	AUC 75.13%, Accuracy 80.47%	Practical, interpretable tool using routinely available pre-operative ultrasound + age; produces probability-based guidance rather than a single size cut-off. First BN model for this task; includes short-diameter information (often omitted in prior nomograms). Provides an online calculator and a score table for quick use. Outperforms their nomogram built on the same variables on the test set (BN AUC 0.751 vs nomogram AUC 0.721).	No external validation; findings need confirmation on larger, independent datasets. Only five variables included; authors note the sample size is modest for ML and that adding more clinical/imaging features could improve performance. Retrospective design; all cases are post-cholecystectomy, which may introduce selection bias toward operated patients. (Study cohort description.)
	(Yuan et al., 2023)	Radiomics - Dual-mode US	4	100	AUC 0.850 ± 0.090, Accuracy 0.828 ± 0.097, Sensitivity 0.892 ± 0.144, Specificity 0.803 ± 0.149, YI 0.695 ± 0.157	Dual-modality radiomics improves discrimination over either B-mode or SMI alone; High sensitivity and good calibration; Potential to reduce unnecessary cholecystectomies and missed neoplastic polyps.	Pilot, small sample, single-centre; Results preliminary; Requires multicentre, prospective validation; Authors plan to explore deep learning with larger cohorts.
	(Chen et al., 2020)	CAD (AdaBoost on segmented images)	4	224	Accuracy 87.54%, Sensitivity 86.52%, Specificity 89.40%	Fast and simple - millisecond-level inference with an interpretable, low-dimensional pipeline. Effective segmentation improves downstream classification; explicit demonstration that preprocessing matters. Clinically relevant comparison against novice vs. expert sonologists; CAD can assist less-experienced readers.	Data & generalisability - Single-centre HRUS dataset; requires high-quality HRUS and skilled acquisition—limits portability. Sample size is small, motivating more data and potentially a deep network when feasible. Feature scope - Authors suggest adding more clinical data/salient features to improve performance.
	(Wang et al., 2023)	US Radiomics - XG Boost	4	640	Discriminating neoplastic from non-neoplastic gallbladder lesions: AUC 0.822–0.853a; Discriminating carcinomas from benign gallbladder lesions: AUC 0.904–0.979	Better diagnostic performance than conventional US for both tasks and outperforms CEUS for carcinoma discrimination in this dataset. Clinically impactful - Potential to reduce unnecessary cholecystectomies versus the 10 mm rule. Generalisability steps - Prospective, multi-centre, external testing across scanners	Sample size relatively limited despite multi-centre design. Spectrum restrictions - Excluded lesions <6 mm and masses inseparable from adjacent liver (no delineable ROI); GB wall-thickening–only diseases not included. Not yet compared to real-world guideline pathways in a controlled clinical workflow; authors call for randomized clinical trials.
	(Li et al., 2023)	ML - Bayesian Network	4	759	Accuracy 82.35%	Accurate and practical pre-op risk tool for ≥10 mm polyps, aiming to refine surgical indications beyond a size-only rule. Inclusion of CSA (derived from long + short diameters) as a novel variable improved discrimination; BN outperforms current guidelines on AUC and calibration.	Retrospective design; sample modesty; only four input variables—authors suggest adding serologic biomarkers and radiomics/CT/MRI features in larger, preferably prospective, studies. Guidance threshold suggested (≥50% probability, thus consider prophylactic cholecystectomy) but no real-world pathway trial yet.
	Yuan et al., 2023	Ultrasound radiomic analysis through SVM	4	96	Accuracy 0.875, Sensitivity 0.885, Specificity 0.857, AUC 0.898	Radiomics captured pathology-consistent textural and shape differences (cholesterol polyps more heterogeneous/clustered highlights; adenomas larger/less regular), yielding higher diagnostic accuracy than visual assessment alone and potentially informing preoperative decision-making.	Small, single-centre dataset; No external validation; Only 2D grayscale US; Authors propose larger cohorts, multi-modal US, and deep learning in future work, and suggest integrating automated radiomics with radiologist visual descriptors.

GBC	(Basu et al., 2023)	DL - vision transformer	4	218	Accuracy 0.921±0.062, Specificity 0.961±0.049, Sensitivity 0.923±0.062, AUC 0.971±0.028 (mean ± SD)	Explainability consistent with GB-RADS and ability to surface previously unknown features (e.g., feature id 1607) for potential clinical study; Super-human accuracy vs experts on US images; Efficient inference; Can serve as a second reader to reduce inter-observer variability.	Relatively small dataset) Single-center B-mode still-image data, No external validation; Some highly predictive features lack immediate clinical semantics and require further clinical studies to characterise.
	(Basu et al., 2022)	DL - GBCNet	4	218	Accuracy 91.0%, specificity 95.0%, sensitivity 97.6%	Outperforms expert radiologists on held-out test images; Curriculum reduces texture bias and boosts specificity; Pipeline localises relevant regions better than vanilla CNNs	Single-centre, static B-mode stills , and no external validation limit generalisability; Performance depends on reliable ROI localisation.
	(Xue et al., 2021)	DL - Optimised SegNet	4	300	Increased the IoU by 7.3%, the precision by 8.2%, and the recall rate by 11.1%	The optimised SegNet offers faster, more accurate segmentation on US, which can help downstream diagnostic workflows. CEUS appears to improve detection versus conventional US in this series. The P16 findings add biological context for staging and differentiation.	Single-centre, retrospective, and limited sample size (especially for cancer cases); Authors recommend larger follow-up cohorts. There's no external validation of the segmentation model, and diagnostic performance is framed as "coincidence rate," which is not a full reader-study or ROC analysis.
	(Gupta et al., 2024)	DL - CNN Classifier	4	565	Sensitivity 92.3%, Specificity 74.4%, AUC 0.887	Clinically competitive performance, matching experienced readers and maintaining high sensitivity in challenging scenarios; Potential to assist less-experienced radiologists. Training with a blur-to-sharp curriculum addresses ultrasound texture bias, which is a known failure mode.	Single-centre study without external validation; Authors call for multicentre work to fully establish generalisability. Specificity lagged sensitivity in some subgroups (e.g., mural thickening) versus radiologists, implying potential false-positive trade-offs.
Gallstones	(Lian et al., 2017)	ML	4	Not reported	GB EVA is 86% GB Stones EVA is 80%	Practical, fully automatic segmentation for both GB and stones; Competitive accuracy with short runtimes, suggesting potential to assist diagnosis in routine ultrasound workflows.	Single-centre data and no external validation reported.
Biliary Artesia	(Zhou et al., 2021)	DL - EDLM	4	For training - 1141 For External Validation - 298	Internal Validation Data: Sensitivity - 93.3% Specificity - 85.2 % External Validation Data: Sensitivity - 93.1 % Specificity - 93.9 % AUC of 0.956	Multicenter design with external testing; Consistent performance; Improves clinician sensitivity when combined; Resilience to scanner settings; Workable smartphone/video use-cases; Initial interpretability via CAM	Manual bounding boxes add burden; Occasional attention to non-gallbladder regions; Rarity of BA limits dataset size; Needs fully automatic detection/segmentation and exploration of 3D ultrasound in future work.

2.2.1. GB Polyps

The latter occur in cases of people in good health with a frequency between three and seven percent. The extracted samples are rated as either harmless formations or pathogenic ones. Furthermore, the above are differentiated as adenomatous and ones with non-progressive behaviour. The last ones exhibit in majority of cases (ninety five percent) and rarely create medical complications. The adenomatous samples are considered to carry a cancer threat for the examined patient (Wang *et al.*, 2025; Xu *et al.*, 2017; Lin *et al.*, 2008; Wiles *et al.*, 2014; Inui *et al.*, 2011; Pavlidis *et al.*, 2023; Xu and Hu, 2017; Choi *et al.*, 2016; Dilek *et al.*, 2019; Lam *et al.*, 2021).

The organisational recommendations of medical professionals report that lesions spanning above 10 mm are supposed to be treated operationally. The assessment of such lesions is then happening through an invasive procedure known as Cholecystectomy. The results of which determines whether the formation is malignant and needs to be operated on or is harmless (Wang *et al.*, 2025).

A number of research works employed AI techniques through ML and DL algorithms to distinguish between patient samples exhibiting harmless or adenomatous presence of GB polyps.

A **recent research paper** presents an AI solution that differentiates between neoplastic and non-neoplastic GB polyps through US imagery. The utilised approach is **DL-based** and is developed as a **decision support system**. It employed the **Inception V3** CNN architecture through **transfer learning**. Techniques like **data augmentation and early stopping** were applied to prevent the model from overfitting. The **US data was fused with textual data** like polyp size, patient age, and multiplicity by a late fusion strategy. Although the **trainable**

parameters were **not explicitly discussed**, the architecture typically has around 23 Millon parameters, thus leaning towards a **moderately computationally expensive** application. The study was conducted on **535 patients**, from which 357 were diagnosed with non-neoplastic polyps and 179 with neoplastic polyps. The approach exhibits results of **0.92 AUC, Sensitivity - 74.3%, Specificity of 92.1%, and Accuracy of 85.7%**. The work also reports that when the application was utilised to support human reviewers its AUC grew noticeably. The system outperformed radiologists with less experienced. A **test set** from the previously pre-processed data was employed to perform a **validation experiment**. Also, the work reports that **three human reviewers** with different knowledge level were invited to participate and **review the solution**. Merits that are shared by the authors are the uplifting of diagnostic consistency when the experience levels of the professionals are different, a reduction of false positives and potentially avoiding the use of cholecystectomies, the exhibition of high AUC and specificity in cases with overlapping lesion sizes. Some of the reported **demerits** are that the study can be classified as one which centres around a **single pathology and is retrospective**. The limitation of **not detecting malignant polyps because of lack of such data**. Also, the architecture was **not benchmarked** against some well-known architectures like DenseNet (Jeong *et al.*, 2020).

Another research addressed the same classification problem as the above through a **DL model** on the same type of data. The authors aimed to develop a tool that could enhance diagnostic accuracy and assist clinicians in separating harmless samples from ones that require attention. For the aforementioned purpose, the work presents **a custom CNN** architecture **built from scratch** with 6 convolutional layers, 4 max-pooling ones, 2 fully connected layers and a SoftMax classification one. The model is relatively shallow and straightforward when compared to modern big CNNs. The number of **available parameters** to train is **not specifically**

mentioned, but taking into consideration its design, the design of the approach can be considered as **relatively light**. The study was conducted on **501 patients** where 208 were experiencing harmful polyps and 293 harmless ones. The total amount of **samples from US** used was **3,052**. The authors report the achieved results are **Accuracy of 87.6 %, Specificity of 88% and 0.908 AUC**. These were achieved through a **cross-fold validation of ten folds**. The dataset was split in such a manner for **some data to be separated for a test experiment that was used for a final evaluation**. **No external or multicentre validation** was conducted. Some of the **reported merits** are that **a custom lightweight CNN** is developed that achieved high accuracy and AUC through the utilisation of actual clinical US images and has the potential to reduce unnecessary surgeries. **Demerits** that are reported are that the study is with a **single centre and a retrospective design** with **no external validation** or data coming from multiple hospitals, the **model was trained only on static images** rather than real-time video data, the architecture was not utilised for clinical testing, and the differences in sample shots were under different imaging protocols (Kim *et al.*, 2021).

A **third work** on the topic introduces a **custom multi-head CNN** consisting of **six pipelines** that work on **different input information** (outline, inline, polyp, wall and polyp image patches, age, and sex). The architecture utilises three depth-wise convolutional layers with dropout, then a channel-wise one. The extracted information from the latter is then flattened and concatenated prior its delivery to a fully connected layer. To enhance the model's interpretability the authors employed a **Grad-CAM** framework. **Aggregation** is also used, because each patient's data contains a myriad of shots and a **top-k voting** by image confidence (excluding near -0.5 scores) is applied. The training was conducted with **Focal loss** to mitigate the class imbalance and **AdamW** was utilised as an **optimiser** with **batch size of 8**, within **50 iterations**. The developers did **not report** the number of **available parameters** for the

architecture. The **dataset** that was utilised for the model training and validation spans across **3,118 samples from 263 patients**, where 114 have a confirmed neoplastic pathology and 149 have benign formations. The reported results include **AUC 0.944, accuracy of 86% and specificity of 0.861** where the architecture was **fine-tuned and tested on the test set**. The baseline of the model, where it was not pre-trained showed a low AUC of 0.483. To justify and validate the findings the authors presented a Grad-CAM framework experiment where the physicians achieved accuracy of 0.634 without CAD, which raised to 0.785 with CAD, where the major uplift was scored among professionals with less experience. Authors claim that their study introduces better AUC results than previous binary classification attempts and it is the first research that proves CAD betters the diagnostic accuracy of physicians. On the other side they highlight that some **limitations** are present like the data is derived from a **single medical facility**, which can impact the results, if data from different machines/hospitals is introduced (Choi *et al.*, 2023).

The **next research paper** introduced a **ML architecture** based on a **Bayesian Network** by employing a **Tree-Augmented Naïve Bayes (TAN)** hierarchy. The utilised **data** was gathered **from variables that were taken before surgical intervention from US and hospital data**. The textual information involved include age, count of polyps, and their dimensions. The number of **trainable parameters** of the model were **not provided**. Since this is an ML approach, it can be classified as **lightweight and computationally inexpensive**. The study design involved **data gathered from various medical institutions** in China (11 hospitals), where the information of 1,296 cholecystectomy patients was involved. Compared to the previously discussed work the results are significantly lower with **AUC of 75% on the test set, Accuracy of 80.47%, Sensitivity of 64.58%, and Specificity of 84.13%**. The authors explain that the tool is **easier to interpret** when decision-making is in question due to the utilised AI technology, it also delivers a

probability-based guidance (based on the ML model employed), and that it is the first BN architecture used for the specific task. However, they also outline the fact that the study **lacks external validation**; **just five variables are taken into consideration** and that additional ones can uplift the model's accuracy; and the fact that **all patients represented in the information are after cholecystectomy**, an approach that **can lead to biased results in individuals that have already undergone surgery** (Li *et al.*, 2023).

Another research utilised **bimodal US screens** (B-mode and Superb Microvascular Imaging) through a moulded radiomics. Firstly, the attempt **separates the ROI** via the unhampered expertise of **two examiners** from which the radiomics features are acquired and more specifically the GLMC and binary textures alongside the morphology from the B-mode. Then, the **desired attributes** are picked out through **ICC** (Interclass Correlation Coefficient) filtering and **LASSO** (least absolute shrinkage and selection operator) selection. Then the binary separation into classes is performed through an **SVM model** (Support Vector Machine). **Five-fold cross validation** is also employed. The authors d**id not provide any data regarding the number of parameters**, but considering the applied architecture, the approach can be classified as a **lightweight** application. The research is **retrospective with a single centre** where samples from just **100 participants** were incorporated. The age of the latter is reported to be between 21 and 58, and the samples exemplified 71 troublesome cases against 29 without complications. The authors report reasonably high results with **AUC of 0.850, Accuracy 82%, Sensitivity of 0.892, and Specificity of 0.803**. Some of the **merits** mentioned by the authors are that dual-modality radiomics uplifts the segregation of both B-mode or SMI modalities taken separately; **the reported sensitivity is high**, and it is calibrating well; and that the approach is a **good way of avoiding cholecystectomies**. On the other side, they note that

the **sample size is very limited**, and they are acquired from just a single institution, which can impact generalisability (Yuan *et al.*, 2023).

The **next approach** that delved in the classification of GB polyps delivered a **CAD** method through the **analysis of HRUS images**. For the purpose, a classical **ML** algorithm is employed as the classifier. Due to the differentiation in the sample quality in the utilised dataset, the authors employed a **separation of the ROI** from the rest of the picture to improve the accuracy of the model. Also, a **shape prior wavelet frame-based segmentation** is applied to outline the GB organ, subsequently the sample's **contrast is being normalised**, followed by data **augmentation techniques** like centring and resizing. A **Gaussian PCA kernel** was also utilised (retaining 10 components) in conjunction with a calculation of the polyp diameter. **AdaBoost** is used as a classifier of the two trees with depth of seventy. Thus, the approach can be classified as a **lightweight** one. However, the number of **trainable parameters is not reported**. The **dataset** was derived from a **single medical facility** and comprises of the information of **224 individuals**. The deliverables of the project were assessed through **leave-one-out cross validation** (LOOCV) and **a blinded reader study**, where two experienced sonographers and two novice ones were involved. Through the LOOCV, the CAD approach achieved an **Accuracy of 87.54%, Sensitivity of 86.52%, Specificity of 89.40%, and AUC of 86.21%**. The reported **pros** of the attempt are that it is **computationally efficient, simple, interpretable, showcases that segmenting the ROI betters the classification output, and can reduce the learning curve for novice sonographers**. However, the authors also state that the research **lacks a multi-centre background, the data frame is quite small, which restrains its generalisability** (Chen *et al.*, 2020).

A **consequent work** undertook a **radiomics pipeline** approach. It acquires customised features from **grayscale B-mode images**, employs **feature stability and selection through ICC, t-tests, correlation filtering and LASSO**. This work **benchmarked eleven ML algorithms** for the same classification task and outlined that **XGBoost as the best performer**. The authors also created a logistic-regression comparators from conventional and CEUS images. This was done in order to create a variation of the architecture to include clinical predictors such as age, sex, and more to the radiomics features. The paper **does not report parameter** counts for the XGBoost ensemble, but due to the model used it can be concluded that it is a **lightweight** one when compared to DL solutions. The patient cohort comprised of **640 patients**, and the information was collected across **four hospitals**. Additionally, two independent tests sets were introduced from external sources. The results show that for **neoplastic vs. non-neoplastic classification** the model achieved **0.822 – 0.853 AUC through the validation and external test sets**. When discriminating **carcinomas from benign gallbladder lesions** the results were **AUC 0.904 – 0.979**. When the authors added the clinical predictors to the radiomics results this did not better the outputs significantly. Some of the mentioned **merits** are that the work can **uplift the diagnostic accuracy of professionals**, **reduce unnecessary cholecystectomy**, and improve the **cost-effectiveness** of the whole treatment process. On the other hand, **demerits** are that the utilised **information size is still modest**, and **models were not tested in real life scenarios**, thus clinical trials are needed (Wang *et al.*, 2023).

The **next research** attempt employed a **TAN** model from **multicentre retrospective data** where the number of individuals included totalled to 759. The data was collected from eleven medical institutions. The authors **did not report parameter count** of the model they developed, but due to its ML origins it is considered relatively **computationally lighter** than DL architectures. On a **fixed internal split**, the BN **delivered AUCs of 0.811** on the **training**

subset and **0.877 on the test one**, with **Accuracy of 82%, Specificity of 86% and sensitivity of 55 to 63%**. The study's **strengths** include a **multicentre data** approach and a simple, **interpretable model**. **Limitations** are **the limited sample size**, and **restricted feature set**. The authors outline the need for larger studies that would include biomarkers and radiomics to further better the generalisability and performance (Li *et al.*, 2023).

The **last** well-known published **research** paper that delves into the world of binary classification of GB polyps employed a pipeline consisting of **lightweight radiomics and an SVM** model to segregate true adenomatous formations from pseudo cholesterol ones. The US **samples** were taken **before surgical intervention**, and the **ROI were manually highlighted** by specialists. In such a manner the authors managed to acquire thorough first-order, binary textured, and morphological features, which they diminished through a **statistical filter consisting of two phases**. This resulted in **features** comprising of **eight keys**, which were fed to the SVM with a **5-fold CV on a 60/40 split**. To **verify the results** the authors performed an **internal experiment** through a test set which delivered **AUC 0.898, Accuracy of 99%, and Specificity of 0.857**. Hence, outperforming deliverables from only morphology and spatial subsets. The study argues that these quantitative features mirror pathological differences and may prove useful in the decision-making process before interventions. However, the work is **limited by its small and one centre derived design**. Also, **no external validation** is reported, and the emphasis is placed on 2D grayscale samples. The authors mention that a larger, multimodal, and DL study is required, that would be further validated by medical trials (Yuan *et al.*, 2023).

2.2.2. GB Carcinomas

Gallbladder Cancer (GBC) is an aggressive pathology and the most frequent cancer of the biliary tract. Its occurrence varies widely across regions and countries, with the highest rates reported in native communities in South America, Northern India and East Asia. Due to the fact that the pathology usually develops without external or internal signs it is often detected at an advanced phase. Thus, the average survival rate is about six months and less than 5% of the patients live to the 5-year mark. Hence, spotting the disease on time is of utmost importance to improve the patient's chances. As with the other GB pathologies, carcinomas are also diagnosed through US imagery, however the identification of GBC at an early stage through the latter remains difficult. Thus, the involvement of AI techniques into the aforementioned task is gathering pace (Wang *et al.*, 2025).

One of the well-known works suggests an **explainable transformer model with two stages** to detect GBC through US samples. A **global branch**, that is lying on the basis of **ResNet-50**, **segregates the ROI** and a **local branch that is inspired by BagNet-33** learns a **bag of visual features**. Then a **transformer with 4 layers fuses the global and local cues** to order the samples into classes, namely into normal; benign or malignant. The training process employs **ImageNet initialisation**, a **phased fine-tuning**, and a **Stochastic Gradient Descent** (SGD) with **cross entropy**. This is performed though **60 epochs** with batches of **16**. Explanations are then delivered through mapping the most activated local features. The authors **did not report** the availability of **trainable parameters**. The study is developed over a dataset consisting **of 1,255 images** from 218 individuals from a **single hospital**. The showcased results are **Accuracy 0.921, Specificity 0.961, Sensitivity 0.923, AUC 0.971**. The authors report that some of the **pros** of the work are that the **inference is very efficient and the accuracy is high**, thus it can serve as a second examiner to diminish doubts in diagnosis. However, they also note that the

data frame they employed was relatively **small** and coming from a **single medical environment**. Also, the **results have not been externally validated** (Basu *et al.*, 2023).

Another research from Basu and other scientists that employed the same dataset introduced the **GBCNet**. It is a model that is consisting of two phases based on **DL** techniques, which initially **spots the GB ROI** and then decides whether the formation is malignant or benign through a head that is multi-scaled and pools in second order. Texture bias is then assessed through a custom-made inspection. The delivered results are **Accuracy 91.0%, Specificity 95.0%, and Sensitivity 97.6%**. This was achieved through an **internal validation experiment** using a test subset. **Parameter** counts and runtime were **not reported** but through the architecture of the approach it can be concluded it is **computationally moderate**. **No external or multicentre design were employed**; thus this is a lack of the study. However, the delivered outcomes are rivalling expert sonographers (Basu *et al.*, 2022).

The **next study** examines a **tuned SegNet** for segmenting GB US and compares it with a medical outlook at CEUS and P16 expressions. From a technical perspective, the team adds **pyramid pooling**, so the network keeps more spatial detail without expanding the memory or training time. On the data the authors employed, the segregation metrics were improved by seven to eleven percent over the standard SegNet. CEUS delivered better results when compared with regular US samples for GBC detection. Low P16 staining showed that it had a connection with the advancing phase of the pathology and poorer differentiation, which exemplified as useful biological signals. The paper **did not report parameter counts**, but the architecture aims to be efficient and **not heavyweight**. **Limitations** are clear – data deriving from a **single hospital**, **no external validation and limited malignant samples**. Thus, the

outputs of the model need further confirmation in a multi-site and clinical studies (Xue *et al.*, 2021).

The **last study** focused on data from North India and the architecture that was employed comprises of a **multiscale and second-order pooling CNN**, which was **fine-tuned** on the gathered data. The data frame of US **did not have any ROI marking** and used **a blur-to-sharp training curriculum** to tackle bias in texture. It involved the samples from **233 individuals**, which were utilised for training and a **separate cohort from 273 patients**, and their data was used for testing the model. Hence, the achieved outputs reached **Sensitivity of 92.3%, Specificity of 74.4%, and AUC of 0.887**. Such results can be considered to rival experts in the field. No parameter counts are reported, but the shared GPU that was used for training can lead to the conclusion that the model is **moderate** as a load. Hence, a tool that could be considered practical and with high sensitivity is delivered, which can enhance the ability of even non-experts to diagnose the condition. However, there is still the need of external and multicentre validation before routine deployment (Gupta *et al.*, 2024).

2.2.3. Gallbladder Stones

The pathology is one of the frequently occurring GBD variations. Roughly, one in ten people will have them. The likelihood of the issue appearing is higher with progressing age and affects mostly women. In the majority of cases they are the simplest to spot through US diagnostics. However, on some appearances when they sit in an awkward position inside the organ or the individual is changing posture or when they go together with other GB abnormalities the diagnosis can become trickier (Wang *et al.*, 2025).

One **research** offers a glimpse into the classification of whether the patient's samples contain gallstones or not. It presents a classical (**ML**) pipeline that autonomously segregates both the GB and gallstones on US by employing **five steps**. Namely, these are (Lian *et al.*, 2017):

- **Anisotropic diffusion plus modified Otsu** to tame tackle speckle and lift contrast
- **Global Morphology Filtering** to output a clean GB mask
- **Parameter-adaptive pulse-coupled neural network** to outline bright formations that may be stones.
- **Modified region growing** to tackle too much segmentation and ignore marks left by medical experts.
- **Smoothing** to clarify boundaries.

The **data** was collected from a **single hospital**, and the authors report robust overlap and localisation – for the GB **EVA is 86%** and for stones **EVA of nearly 80%.** The authors underline that their approach's deliverables are competitive with other technologies and claim it could assist clinicians by accelerating the diagnostic process. However, **the evaluation was internal only** (Lian *et al.*, 2017).

2.2.4. Biliary Artesia (BA)

Biliary atresia (BA) is not a very frequent disorder of the GB. It is particularly seen in the bile ducts in infants in one of twenty thousand individuals worldwide. However, the occurrence is more frequent in East Asia. It is also the leading indication for paediatric liver transplantation, which is why spotting it early matters. When BA is diagnosed promptly and treated, outcomes are generally better, more specifically in cases where operative interventions are performed at a young age (Wang *et al.*, 2025).

US is the first test when BA is suspected. Abnormal gallbladder morphology on ultrasound is a particularly useful sign, with sensitivity and specificity commonly revolving around 90%. Although not every occurrence in children showcases classical disease marks, thus additional US features may be needed to avoid missed cases. Due to the fact that BA is uncommon and early clues can be subtle, misdiagnosis or delayed one are still a frequent scenario, hence some young patients lose the window for optimal treatment (Wang *et al.*, 2025).

An attempt in the field was made by training an **ensembled deep learning model (EDLM)** comprising of **five Se-ResNet CNNs**. Each network was **trained of a various fold** and **the output of each was averaged**. A **five-fold scheme** was applied in conjunction with **transfer learning, augmentation, class-weighting, and dropout** to segregate BA cases from the US samples. The team assembled **3,705 US** shots from 330 BA positive and 811 negative people across **five hospitals** for training purposes. The **set for external validations** was collected from **six more medical institutions** and comprised of **841 images**. The **sensitivity and specificity** for each individual outputted by EDLM **on the internal validation information** were **93.3% and 85.2%**. The same metrics reported **on the external validation** data frame were **93.1% and 93.9%, also an AUC of 0.956**. Thus, surpassing professionals' expertise deliverables. The **number of trainable parameters were not specifically mentioned**, but it can be considered, due to the technology used that it is a **moderate to heavy application**. Some of the **pros** of this research are that it is a **multicentre one** with available **external validation** and that it proves that when combined with human expertise it **improves the diagnostical sensitivity**. Also, the study reveals that **the model can be utilised on a smartphone or on live video data**. However, some of its **cons** are that the addition of **manual bounding boxes** in the processing adds up processing time, on occasions **the model pays attention to regions that do not**

include the GB organ, and the infrequency of the condition significantly limits the available data (Zhou *et al.*, 2021).

2.2.5. Classification on Nine Gallbladder Pathologies.

There are **only two known** and reputable works up to date that tackle the problem of classifying multiple GB pathologies.

The **first one** of them, introduces a large and mostly balanced US dataset consisting of **10,692 image** samples spanning **across nine GB conditions**. The authors' pipeline starts with **denoising and enhancing the data through non-local means and a bilateral filter, followed by a ROI segregation through a DNN to increase the attention of the classification architecture, then four models are benchmarked** (VGG-16, Inception-V3, ResNet-152, and MobileNet). The **preprocessing** steps taken are simple **rescaling and augmentations techniques within an 80 to 20 train-to-test ratio**. **MobilNet** exemplified the **highest scores** with an **overall accuracy of 98.35%**, which signifies that a compact model architecture can tackle the problem effectively and deliver high outputs. Nonetheless, the paper **did not report parameter counts** or runtime. Another **con** is that the **validation** is only performed **internally**, thus it is hard to conclude whether robustness is present for real-world scenarios. Another thing to note is that the authors do report that they applied **grouping by patient ID to prevent leakage** (Obaid *et al.*, 2023).

Another approach to the multi-class problem treats the diagnostic process as **case retrieval rather than direct classification.** Meaning, they submit an US query, and the system pulls back the most visually similar prior cases using features that are composed together. This is done through **GoogleNet, Inception-V3, and NASNet-Large**. The models are used as **pre-trained** and **are not specifically trained on the dataset**, which is the same one as the previous work.

With other words, the aim here is to provide the top matches to the clinician from the available library and quickly compare appearances. Hence, here there is no reliance on probabilistic output, but more of a guidance towards the professional. The authors report an **overall average precision of 0.94**. The whole approach was implemented on a normal Windows 10 PC, which supports the idea that this research can be considered lightweight to moderate computationally (Bozdag *et al.*, 2025).

The paper does not report specificity, AUC, and sensitivity or a reader comparison. Thus, it is unclear whether and how the system influences real clinical decision making. There is also **no external validation** (Bozdag *et al.*, 2025).

2.3. Literature Review Conclusions

Through the above sections it is evident that US remains the most practical diagnosis modality for GB disease. However, both classical radiomics and modern deep learning add clinically useful implications. For polyps, the compact radiomics pipelines on B-mode achieve strong discrimination and offer easy to interpret results that outline the pathology. In the scenario of GBC, the reviewed works showcased that expert-level accuracies are achievable through AI techniques, although specificity is slightly falling behind other metrics and wider external validation is needed for most of the works. In biliary atresia studies it is exemplified that multicentre experiments, external tests, and even smartphone or video implementations are possible, while taking into consideration the need for manual region selection and a following full automation of the pipeline. The multi-class classification approaches are promising but are not externally validated, thus the results on internal testing are high, but **more attention is needed to data on patient level, and their generalisability requires further research**.

Hence, these conclusions support the statement that thoughtful design is having an upper hand on the large number of trainable parameters that big DL models have, and a **light model can be adapted to the classification task**. The **choice of evaluation metrics also plays a vital role** in tackling the **class imbalance** in the limited datasets available. Another important point drawn from the Literature Review is that **enforcing patient-level splits and folding cross validation may improve the model's stability and accuracy as well as providing fairgrounds for transparent results**.

3. METHODOLOGY

3.1. Chapter Overview

This chapter explains how the study is designed and executed to evaluate Lightweight DL models for classification of nine GBD classification from US imagery. Two goals are balanced – diagnostic quality and usability on legacy or mobile devices to uplift medical staff's diagnostic abilities. Hence, the current chapter outlines how the research methodology is informed by researching different options to answer some of the research questions; presents the research stance in point **3.2**. **Informing the Methodology**, then it proceeds to **3.3. Research Philosophy** that explains the philosophical stances taken in the current research, and finally in point **3.4.** and onwards discusses the exact procedure undertaken by this research to address the aims and objectives and what are the potential theoretical and practical implications.

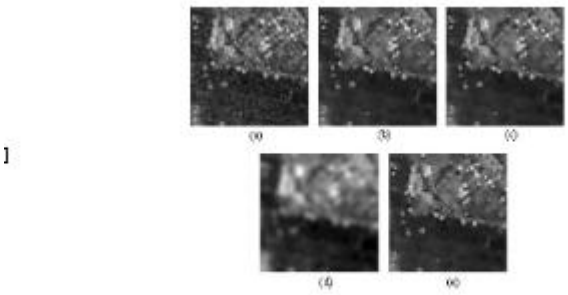
3.2. Informing the Methodology

3.2.1. Why Preprocessing Matters for Lightweight DL? What is the Impact on Performance and Generalisability? What are the Suitable Approaches to Tackle the Issues?

US as a diagnostical is fast and accessible in most modern-day hospital environments. However, its output is quite often resulting in messy samples due to the speckle noise, variable gain, probe angle and patient positioning when the imaging is performed. Thus, leading to occasional subtle boundaries and textures resulting in limited GB visibility. Lightweight models have less capacity to smooth out this variability. They are most useful when fed with clean and standardised inputs. In practice, careful preprocessing improves these algorithms' abilities to segregate and distinguish between classes and aids them in generalising across different scanners and medical sites (Takahashi *et al.*, 2024; Moinuddin *et al.*, 2022; Yu and Acton, 2002).

Intensity standardisation proves to be a suitable technique to tackle the aforementioned issues whereby normalising the intensity levels across the image samples diminishes local and caused by the imaging protocol shifts in colour. Evidence from medical imaging shows that harmonisation and intensity normalisation can tangibly uplift the performance and robustness of algorithms trained on the data, especially in cases where data was gathered from multiple medical institutions and different imaging machines (Reinhold *et al.*, 2019; Cobo *et al.*, 2023).

Another issue with US imaging modalities is increasing **noise** that can hide borders. Thus, **denoising** can be applied with attention to speckle formations, which can improve boundary visibility while keeping the important textures of the imaging sample (**Figure 7**). The technique is also known as **speckle-reducing anisotropic diffusion (SRAD)**. However, it is important to note that such interventions should be approached with caution, because there is the danger of smoothing the image too much and loosing of diagnostically useful patterns (Yu and Acton, 2002; Moinuddin *et al.*, 2022).



(a) Original 125 × 125 SIR C SAR L-band HH image. (b)–(e) Filtered images by the En-Lee, En-Frost, AD-Homomorph, and SRAD filters.

Figure 7 - SRAD Application Example (Yu and Acton, 2002).

Another approach that is suitable for the project is the **standardisation of the size and geometry** of the input data. The application of resizing, padding and cropping, so the GB region remains centred and scaled across samples is useful during model training and testing,

because the algorithm can spot similar spatial layouts more efficiently (Shorten and Khoshgoftaar, 2019).

In the scenario of medical imaging, **simple augmentations** are also useful if applied with modesty. Hence, small rotations of five to ten degrees, horizontal flips, mild scaling, and gentle brightness or contrast interventions are proven and well-supported in academia to uplift generalisability. However, if the above are utilised aggressively this can distort the sensitive modality data and cause architecture confusion (Shorten and Khoshgoftaar, 2019; Cubuk *et al.*, 2018).

There is also the **problem of blurriness**. Images that contain high blur can confuse the model during training and reduce its performance. A simple and widely used check is the **variance of Laplacian** (**Figure 8**), which is effectively a sharpness score. Through it samples which are not clear enough can be highlighted and excluded if they fall below a certain threshold to avoid the aforementioned issue (Pertuz *et al.*, 2013).

$$\phi_{i,j} = \sum_{(i,j)\in\Omega(x,y)} (\Delta I(i,j) - \overline{\Delta I})^2,$$

where $\overline{\Delta I}$ is the mean value of the image Laplacian within $\Omega(x,y)$.

Figure 8 - Variance of Laplacian (Pertuz *et al.*, 2013).

3.2.2. Which Lightweight Models Offer Best Balance Between Computational Needs and Diagnostic Accuracy?

Since the goal of the current work is to assess and develop an approach that will be suitable for application on existing medical infrastructure there are several families of architectures that prove suitable for the task and are researched and summarised below.

GhostNet is a small and fast CNN that avoids performing heavy calculations. It substitutes the multiple and computationally expensive convolutions with cheaper “ghost” feature maps (**Figure 9**) created by uncomplicated linear ops. Meaning, it first creates a small set of key feature maps and then generates the rest by quick and inexpensive transformations of the latter. The blocks of the architecture are ordered as “Ghost bottlenecks” (**Figure 10**), which are similar to MobileNetV2, but are trimmed to improve work efficiency. Hence, the model achieves similar results with less computations. The algorithm delivers **5.2M parameters** and takes **224x224 pixel RGB inputs**. It is **pretrained on ImageNet** and achieved 75.7% top-1 accuracy. It proves as a good fit for the current project as it provides the lowest latency among CNNs at a set accuracy. Thus, on typical hospital infrastructure or small edge devices, it will perform quickly. It can also be shrunk and made to work faster by using INT8 quantisation, which is storing the numbers in 8-bit format. However, that approach may cost the accuracy to go down a bit, but it can be recovered if the model is retrained with quantisation turned on (Han *et al.*, 2019; Luo *et al.*, 2024).

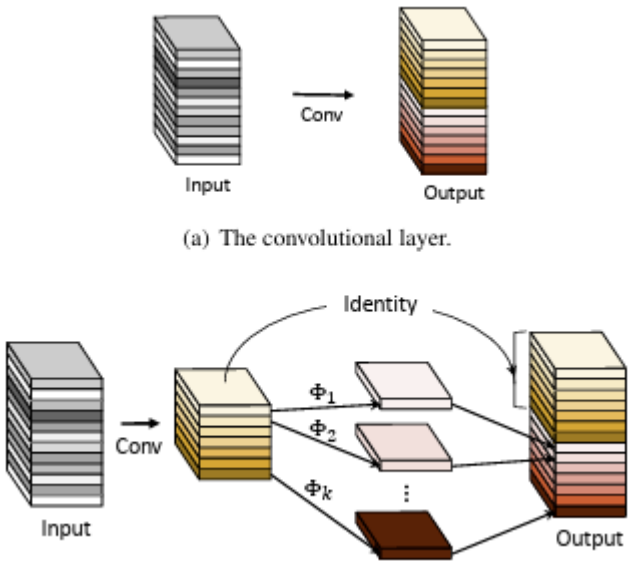


Figure 9 - The Ghost Module (Han *et al.*, 2019).

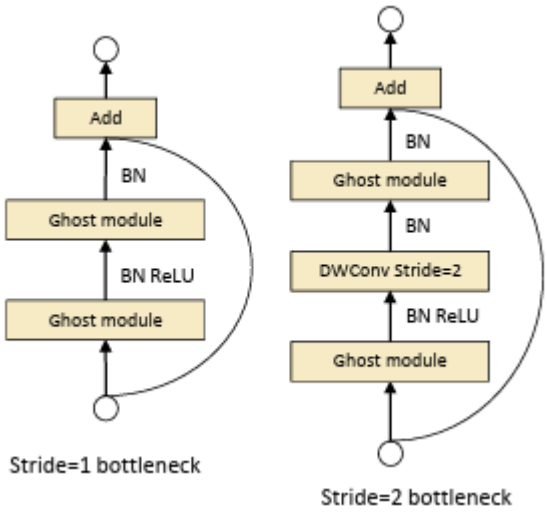


Figure 10 - Ghost Bottleneck (Han *et al.*, 2019).

Another suitable architecture for the purpose is **TinyViT (Figure 11).** As the name suggests it is a small vision transformer model trained with a fast-pretraining distillation scheme from

large teacher models. In simpler terms instead of learning everything from the beginning it utilises a larger model as a teacher during pretraining. In such a manner, it looks at the whole image simultaneously (the global context) and employs fewer weights, thus making it light and quick. It has several variations consisting of **5, 11, and 21 million parameters**. It is also capable of working well with higher spatial planes when the accuracy is important. Due to the fact that it looks at the image from a global perspective it can prove useful when multiple GB pathologies are present in different parts of the gallbladder. The input requirements of the algorithm are **224x224 pixels RGB samples**. It delivered 84.8 % top-1 accuracy on the cleaned **ImageNet-21k** dataset (Wu *et al.*, 2022).

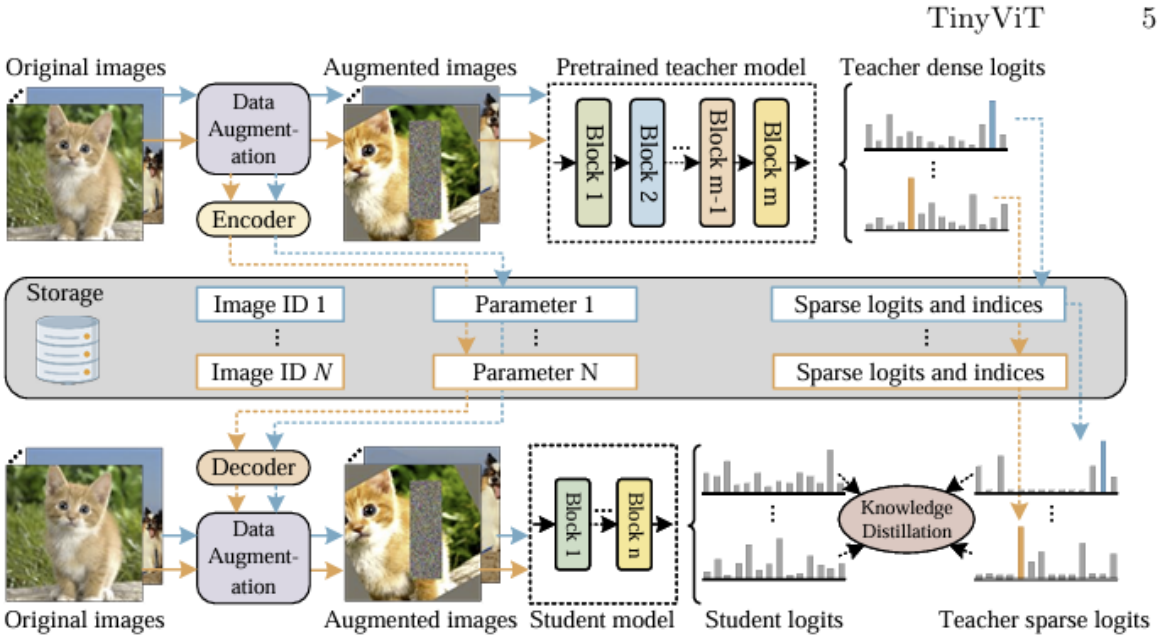


Figure 11 - TinyViT Architecture Workflow (Wu *et al.*, 2022).

ShuffleNetV2 (Figure 12) is a CNN that is specifically developed to target speedy outputs. The architecture maintains a good balance between channels, does not put a great load on the memory, and sticks to simple and fast operations. It is based on a “split, shuffle and mix” strategy where it separates the feature channels, jumbles them to blend the information, and then continues. This is what makes it really cheap and effective, especially on ARM devices and mobiles. It delivers **2.3 million parameters** and is very good on latency. It is also trained on **224x224 RGB input**. It is pretrained on the **ImageNet-1k** dataset and achieved top-1 accuracy of 74.9% (Ma *et al.*, 2018).

Building blocks of ShuffleNet v1 [35] and this work. (a): the basic ShuffleNet unit; (b) the ShuffleNet unit for spatial down sampling (2×); (c) our basic unit; (d) our unit for spatial down sampling (2×). **DWConv**: depthwise convolution. **GConv**: group convolution.

Figure 12 - ShuffleNet Architecture Workflow Ma *et al.*, 2018.

Another suitable architecture is the **MobileNetV3 (Figure 13).** It is a mobile friendly CNN that is specifically developed so it can be performing quickly on mobile and computationally restricted devices and provide good accuracy. It merges automated architecture search with tweaks like h-swish activation and lighter end layers to cut down on latency without salvaging accuracy. It is available in two versions – Large and Small. Where the **large** brings **5.4 million** parameters and the **small 2.9 million**. The results they deliver are Large – 75.2% top-1 and Small – 67.4% top-1. Both of them benchmarked on the **ImageNet-1k** dataset. Compared to the previous version the current model presents 20% lower latency and 3.2% more accuracy. The input it accepts is **224x224 pixels RBG** shots (Howard *et al.*, 2019).

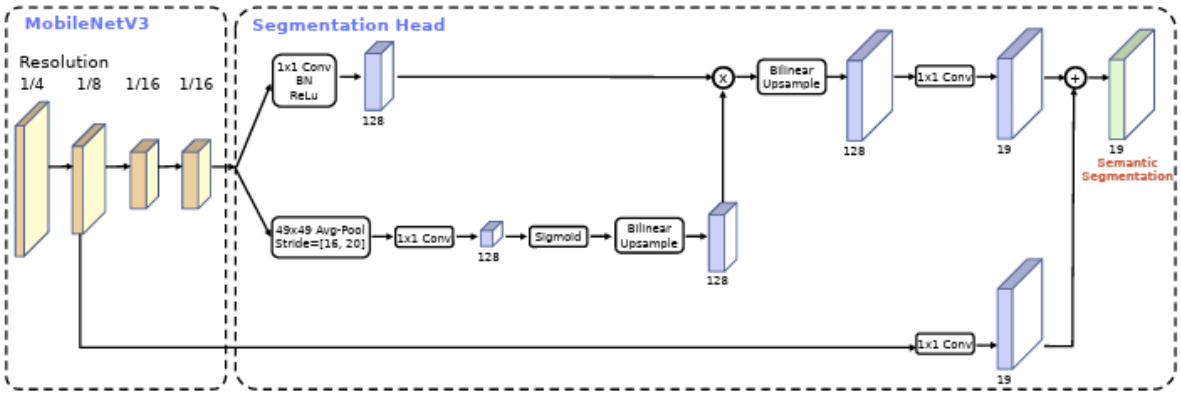


Figure 13 - MobileNetV3 Architecture Design (Howard *et al.*, 2019).

3.2.3. Which Are the Benchmark Models Used in Medical Imaging Diagnostics?

Across multiple diagnostic classification papers and standard benchmarks, **the most common baselines and benchmarking** backbones are (Rajpurkar *et al.*, 2017; Kundu *et al.*, 2021; Elshennawy and Ibrahim, 2020; Yang *et al.*, 2023):

- **ResNet-50** – it is the classical backbone for models to be compared against, employed in large public studies and in more recent RSNA/MedMNIST benchmarks. It is widely adopted as standard benchmarks for diagnostic performance.
- **DenseNet-121 –** again it is a classical algorithm used as a standard benchmark in medical image analysis.

3.2.4. Suitable Training Strategy and Optimisation Techniques with Focus on Preserving Diagnostical Accuracy.

1. **Suitable Training Approach.**

The first training technique that is a suitable fit for the project is **Transfer Learning**. It is a training method that allows a model to be trained on new data without the need of starting from the beginning. Meaning the network is pre-trained on another similar dataset and the user just finetunes it on the new information. The technique is useful because the first blocks are able to acquire knowledge regarding the reusable patterns like edges and textures and then the new layers are trained on the new information with less data and compute to reach good output levels. Thus, the so called “freezing” and “unfreezing” of layers, should be performed with caution as it can destabilise the architecture. Thus, since the project has time and resource limitations that have been mentioned earlier in this report, and is aiming at developing a lightweight approach, its suitability is justified (Pan and Yang, 2010).

2. **Optimiser and Learning Rate.**

Two other important features in the training procedure that affect the output are the **optimiser** and the **learning rate**. Where an **optimiser** is an algorithm that controls the architecture’s weights to uplift its prediction capabilities. Common choices are **stochastic gradient descend (SGD)** with momentum and **Adam/AdamW**. The main difference between them is the way in which they combine recent gradients to make the training faster and more stable. Where **SGD** employs a global learning rate and a velocity term that irons out updates by extracting and moving forward some of the previous steps and information. Its benefit is the simplicity; however it needs a good learning rate schedule. **Adam / AdamW** utilise per-

parameter adaptive learning rates. The latter are derived from the averages of the mean and variance of recent gradients, so the training can be quicker and more stable on noisy and sparse problems. AdamW betters Adam by decoupling weight decay from the gradient step, thus aiding generalisation (Goodfellow *et al.*, 2016; Kingma and Ba, 2014; Loshchilov and Hutter, 2016, 2017).

On the other hand, the **learning rate and schedule** is the size of each step for every update of the weights. Meaning, if it is too large the training can blow up, or if it is too little it can stall. Hence, the schedule is simply a blueprint for how the step size can variate over time. Planning good schedules usually lead to smoother training and better final accuracy (Goodfellow *et al.*, 2016; Smith, 2015).

A scheduling form is **Cosine Schedule with short warmup.** The idea here is that the training starts with a lesser learning rate for a few epochs, which is called the “warmup” phase. This is done to provide stability for early weight updates. Following, is the gradual decrease or increase of the learning rate on a smooth cosine curve until the training ends. Thus, warmup prevents the model from getting “shocked” and in the meantime the cosine decay pushes bigger steps earlier, which is helping the architecture to learn quicker, followed by the smaller steps in later training stages, where the finetuning comes into play. That is why it is a suitable solution for the present scenario of lightweight DL applications. Such learning rate adjustments are decided on a dip of improvement in the validation metric during training (Goodfellow *et al.*, 2016; Goyal *et al.*, 2017; Loshchilov and Hutter, 2016).

3. Loss metrics and their impact on tackling class imbalances.

Batch-Balanced Focal Loss (BBFL) merges two concepts that help with imbalanced data which are – Focal Loss that helps focus when tough samples are used, and batch balancing where each small sized batch of samples aids in balanced manner across classes. Which means that classes with less information present in them do not get overwhelmed by larger ones, and the network pays equal attention to mistakes made. There is scientific evidence that the approach uplifts single-class and multi classification in medical imaging. Thus, its suitability for the current work (Singh *et al.*, 2023).

Another solution that is a good fit is **Focal Loss.** As explained in the previous paragraph, it aims at the hard cases and re-calculates the loss in a manner that samples that are easy to discriminate are having less importance, and the hard ones have more weight (Lin *et al.*, 2017).

4. Other Suitable Techniques.

Early Stopping is another approach that is a good fit for the current work where the validation metric is being observed and if it stops getting better for a few epochs, which is called the “patience”, the training is halted, and the best model checkpoint is saved. This avoids overfitting, saves computational costs, and is especially handy where small and efficient architectures are applied (Prechelt, 1998; Goodfellow *et al.*, 2016).

Validation for Leakage (Leakage-safe Validation) separates the information at the patient level so images from one and the same individual never appear either in the training or in the validation fold. In such a manner data leakage is not occurring, meaning that duplicate appearance of the same samples in both folds is evaded. Hence, the outputs are not

artificially inflated, which provides with a realistic performance estimation (Kaufman *et al.*, 2012; Tejani *et al.*, 2024).

Bounded Hyper-parameters such as learning rate, weight decay, and loss settings manage the learning process of the network. Thus, instead of looking endlessly for the proper configuration, a small grid with pre-declared ones that are common amongst models can be employed. This is what the technique actually is. Thus, it diminishes the dangers of over-fitting (C. and L.C., 2010; Bergstra *et al.*, 2012).

5-Fold Cross Validation is an approach that segregates the samples into five equal parts called “folds” at the patient level. The architecture is then being taught on four of them and tested on the remaining one. This procedure re-occurs five times, where each time a different mixture of samples is utilised for testing. When this is done, the results are averaged and in such a way a more reliable estimate of performance is derived when compared to a simple training and validation split. Also, when limited information is available this method proves handy. Thus, it is a good choice for the current research as the utilised dataset is not large, hence each case will contribute to both training and testing and improve output efficiency (Arlot and Celisse, 2009; Brownlee, 2020).

3.2.5. Which Metrics Best Assess Performance in Medical Imaging?

Since the main problem of this work is multi-class classification of GBD and some conditions are rarer than others, slight or average imbalance in the available data samples is possible. Thus, the utilised metrics should be carefully selected to assess each class fairly, portray clinical risk (meaning missed positives against false alarms), and report uncertainty across the

5-fold cross validation. Hence, the following metrics emerged through research of different medical imaging research papers to be the best fit:

1. **Accuracy** – this is the overall percentage of correct identifications of disease types, in the current scenario. It is easy to read and comprehend but can be misleading with imbalance. This is so because the architecture can show good performance on the surface, but actually it can be segregating common classes better and make a lot of errors on the rare ones (Powers and Ailab, 2020).
2. **Sensitivity (Recall) and Specificity (per class)** (Powers and Ailab, 2020)**:**
    - Sensitivity or Recall is the representation of all true cases of the class "k", and how many was the algorithm able to catch effectively (true positive rate).
    - Specificity is the count of non-k occurrences that were correctly identified and ruled out. This metrics is very clinically useful and utilised as it depicts the logic "catch disease vs. avoid false alarms". In such a manner it supports the highlighting of classes that the model is failing to distinguish correctly.
3. **Precision** is a positive predictive value. Meaning, it checks when the model claims that it recognised a class as class "k", is true or not. This metric is especially useful when false positives carry costs (Powers and Ailab, 2020).
4. **F1-score** is the harmonic mean of precision and recall (Powers and Ailab, 2020):
    - The **Macro-F1** variation is an average F1 across all classes, thus rare classes matter as much as common ones. In such a manner imbalance is tackled.
    - **Per-class F1** is employed when classes that are being weakly recognised are in need to be highlighted.

5. **ROC-AUC** summarises ranking quality across thresholds. For example, whether the model grades a true “k” higher than a “non-k” sample. However, it is good to note that if the imbalance is heavy this metrics can be regarded as inaccurate (Hand and Till, 2001; Saito and Rehmsmeier, 2015).
6. **PR-AUC (Precision-Recall Area, macro-averaged) is** more sensitive to the way the architecture is performing on rare classes and to false positives in the latter. Thus, it can prove more informative than ROC-AUC in cases where imbalance is present (Saito and Rehmsmeier, 2015).

3.2.6. What Experiments Can be Performed to Analyse the Performance of the Best Performing Model?

There are various different methods that the best performer can be tested and validated through. Below are summarised and briefly overviewed some of the most common practices in Medical Image Screening research. However, it is good to point out that in order for such research to be widely recognised a multi-centre validation is required. This is something that is lacking in most of the reviewed works in the Literature Review Section. Due to time constraints and resource limitations, no such validation is going to be performed in the current work, also.

7. **Grad-CAM ++** is an approach that employs a heatmap over the sample that outlines where the network “looked at” to give its foretelling. The framework utilises the model’s gradients (which is how much each feature map affects the chosen class) so it can colour important areas. The method is called “++”, because it tops up the original approach by managing multiple pixels better. It is used in Medical Imaging research as it underlines

tiny findings that lead to one diagnosis. In the GBD scenario the attention should be on the Gallbladder organ and its parts, instead of borders or clinical text. Thus, the heatmap will allow a visual confirmation whether the algorithm's attention is really on the anatomical structure (Chattopadhyay *et al.*, 2018).

8. **Error analysis through Confusion Matrices and Per-class PR Curves –** such a matrix is a table consisting of rows that contain the true class and columns with the predicted one. Each cell counts how often the network made that mapping. This aids to highlight where the model is systematically making mistakes and provides a track for improvements and fixes. For example, if it is in the current work's scenario between the classes of polyps against GB stones (Singh *et al.*, 2021; Powers and Ailab, 2020).

   On the other hand, PR curves exemplify the imbalanced classes or ones that are appearing rarely in the data. Thus, if PR-AUC per class is calculated it would be handy to adjust a minimum recall target and analyse the outputs (Saito and Rehmsmeier, 2015).

9. **Simple testing on a held-out test set** is one of the classical experiments to test a classification algorithm. It is performed through and evaluation on a test set that was never used for training or tuning (Brownlee, 2019; Powers and Ailab, 2020).

3.3. Research Philosophy

This project assumes a **Positivist, Quantitative, and Experimental** stance.

When **Ontology** is concerned the Gallbladder pathologies are real and exist in the objective reality, thus the sample labels reflect that reality.

From **Epistemological** perspective the study assumes an **Objectivist** approach as the knowledge is acquired through observable data and statistical evidence like metrics and tests.

When **Axiology** is concerned the research aims for **neutral, transparent and reproducible** results, while taking into consideration clinical risk and ethical utilisation of secondary data that is anonymous.

The **Methodological stance** of this work is **deductive** and **experimental**.

3.4. Research Methodology

The whole Methodology of this research is portrayed in **Figure 15**.

3.4.1. The Dataset

The current research is using the **UIdataGB: Multi-Class ultrasound images dataset for gallbladder disease detection**. It is available to download and preview from this link - https://data.mendeley.com/datasets/r6h24d2d3y/1 . The collection consists of **10,692 US image samples from 1,782 participants,** which are segregated into **nine clinically meaningful categories of GBD pathologies** (Turki *et al.*, 2024).

The data was acquired over four years in Baghdad, Iraq across **multiple clinical sites** (four), through different modern US scanners. The information was collected by medical employees and labelled by sonographers. Consequently, it was checked for quality assurance by senior experts and where disagreements were present for a specific diagnosis, a consensus was reached to resolve it (Turki *et al.*, 2024).

The dataset groups the samples into nine categories which reflect common and serious GB conditions (**Figure 2**). Overall, the dataset is large when compared to other available GBD datasets and mostly balanced by class (approximately between 1,000 to 1,600 images per class). For instance, Carcinoma cases comprise of 1,590 samples from 265 individuals and GB wall thickening consists of 990 US shots derived from 165 patients (**Table 2**). Gender distribution is slightly imbalanced (**Table 3**), with 6,246 female samples from women with an average age of 47 years, and 4,446 male shots from participants with an average age of 53 (Turki *et al.*, 2024).

Table 2 - Overview of Details in the Dataset by Pathology Variation (Turki *et al.*, 2024).

Disease type	Number of Images	Number of patients	Female	Male
Gallstones	1326	221	137	84
Intraabdominal and Retroperitoneum problems	1170	195	110	85
Cholecystitis	1146	191	102	89
Membranous and gangrenous cholecystitis	1224	204	109	95
Perforation	1062	177	95	82
Polyps and cholesterol crystals	1020	170	99	71
Adenomyomatosis	1164	194	108	86
Carcinoma	1590	265	155	110
Various causes of GB wall thickening	990	165	92	73

Table 3 - Overview of Gender Segregation of Information in the Dataset (Turki et al., 2024).

Sex	Number of images	Average age
Female	6246	47 years
Male	4446	53 years

The initial sample dimensions were 450x600 pixels with aspect ratio of 3:4 in a 24-bit format. Nonetheless, the authors re-dimensioned them to **900x1200 pixels** for the purpose of uniformity. They also administered normalisation techniques like zero-mean and unit-variance alongside denoising through a median filter for frames that had noise. Some data augmentation like small rotations, flips, translations, and brightness modifications were also employed. The labels were encoded for ML/DL utilisation (Turki et al., 2024).

From an ethical perspective, **the consent of the participants was obtained** and their **personal information removed** from the samples (Turki *et al.*, 2024).

Some of the drawbacks of the data frame that were highlighted by the authors are occasional low contrast between the assessed bodily part and the background and scanning artefacts that appear in the imagery and diminish the effective segregation and diagnosis (Turki *et al.*, 2024).

3.4.2. Hardware Resource Setup for the Training, Testing and Experimentation on the Algorithms.

The following details are shared for reproducibility purposes, due to the fact that the incorrect setup setting may affect speeds and results in case of an attempt of a reproduction of the current work.

The selected lightweight DL algorithms to perform the classification task of GBD is as follow:

- **Platform** – Google Colab.
- **Programming Language** – Python v. 3.11.13
- **GPU** - Nvidia A100 Tensor Core GPU (40 GB)

3.4.3. Preprocessing and Quality Control

As mentioned earlier US data is frequently noisy and inconsistent due to variations in the scanners used and their settings. Thus, a simple preprocessing pipeline (**Figure 14**) and some quality controls to highlight blurry frames are outlined below:

1. **Input format and standardisation:**
    - Grayscale converted to RGB (one to three channels) – Since all of the reviewed algorithms accept input in the RGB channel format and the US dataset consists of samples that are grayscale, they need to be converted.
    - Normalisation of intensities – to tackle the contrast and brightness shift of each machine utilised for the collection of the dataset, a z-score is calculated for each US shot.
    - Resolution and Aspect Ratio – images are padded to a "letterbox", which is a square image version. Then all samples will be resized to 320x320 pixels as this is the default input size the models accept.

2. **Denoising to address Speckle occurrence:**
    - A light Median Filter of 3x3 pixels to address the noise and speckle but not lose textures that are important for diagnosis (Yu and Acton, 2002).
3. **Augmentation** to enlarge the training distribution without affecting the organ visibility (Shorten and Khoshgoftaar, 2019):
    - Horizontal flips and small rotations of plus/minus five to ten degrees.
    - Light Scaling of plus/minus 10%.

- Light brightness and contrast jitter of plus/minus 10 to 15%.

4. **Quality Control (QC):**

    - Detection of Blur through a computation of the variance of Laplacian for each sample. Poor sharpness frames are excluded or corrected (Pertuz *et al.*, 2013).

It is important to note that to ensure the comparison and evaluation of the chosen model is fair the normalisation is applied to each sample, and the augmentations is the same for all of them. Also, the validation and test images have no augmentations applied.

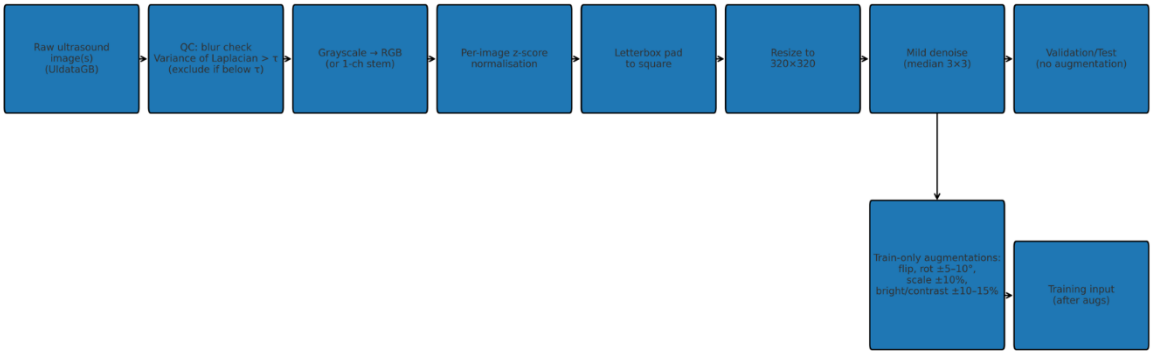


Figure 14 - Preprocessing Pipeline.

3.4.4. Data Splits and Leakage Controls.

Since the current work is employing a **5-Fold Cross Validation** technique to prevent data leakage, in the beginning the information is **split at patient level** into a **development set – 85%** of the data, and a **test set of 15%**. The latter are stratified by class, so rare pathologies appear in both subsets. The test set stays separate and "sealed" until the very end, where evaluation experiments of the best performing algorithm are performed, and for which it is employed.

The **Five-Fold Cross Validation on the development subset** data of participants is divided into 5 folds which are stratified by class. On each iteration the training is performed on four folds and evaluated on the remaining one. In such a way, each fold acts at least once as a validation one. Thus, all samples from one patient live in one partition. The randomness of seeds is set, so the exact splits can be reproduced (Arlot and Celisse, 2009; Kohavi, 1995).

This ensures that data is not leaking across patients, and they do not appear in more than one partition. Any step that learns parameters is fit on the training patients of that fold only and then used on the inner validation and the outer fold (Arlot and Celisse, 2009; Kohavi, 1995).

3.4.5. Selected Models.

From the reviewed models in the Literature Review section the choice fell on **TinyViT** with 11 million parameters and **GhostNet - 1.0** with 5.2 million parameters. Both of them are compared against a widely utilised reference model for classification in Medical Image Analysis, which is **ResNet-50**. This provides the current study with two different architecture technologies which are - a modern transformer with global context (TinyViT) and an efficient CNN that is friendly to quantisation (GhostNet). Both of these are compared against a strong and familiar baseline to justify results (Han *et al.*, 2019; Wu *et al.*, 2022; He *et al.*, 2016).

The choice of the two main models is derived from the fact that the aim is to compare lightweight models, thus one (TinyViT) has more parameters and is capable of delivering a better accuracy and learn features more efficiently, and the other one (GhostNet) has less parameters, but is focused on response times and latency, hence can deliver faster outputs. The idea is to compare the cost and trade-offs between the two and see which one is the better performer and better fit for the classification purpose, keeping in mind that the ability to update legacy medical equipment or mobile devices is one of the drivers of this project. Also, both algorithms use a 3-channel RGB input with 320x320 pixel dimensions as their default input (Han *et al.*, 2019; Wu *et al.*, 2022).

3.4.6. Training Strategy.

Since the main aim is to train the two architectures for classification of the nine GBD pathologies where they are accurate, stable, and deployable, first some finetuning from the pre-learned ImageNet weights is taking place, where an optimiser and schedule are employed to converge the models smoothly and handle class imbalances. The default hyperparameters for the training can be revied in **Table 4**.

Since early layers already know generic visual patterns, they only need to be tweaked to adjust to the US image setting. Thus, a two-stage training process is taking place where (Yosinski *et al.*, 2014; Kornblith *et al.*, 2018):

1. **Phase 1 – Frozen backbone**:
    - All backbone layers are frozen and only the new classifier layer is trained for 3-5 epochs.
    - The idea here is to train the head without tampering the good features in the low level.

2. **Phase 2** – Phased unfreezing with discriminative Learning Rates:
    - Unfreeze layers from last to first over 2-3 steps.
    - Utilise smaller learning rates for early layers and a larger one for late layers as they need better adaptation.
    - This runs for 25 - 40 epochs, inclusive of the first phase and early stopping implemented.

The **Optimiser** of choice is **AdamW** with decoupled weight decay, where the learning rate is equal to **2e-4** for the **first phase** and **1e-4** after the unfreezing. The **wight decay** is set to **1e-4.**

The chosen **Learning Schedule** is **Cosine decay** with a short warmup of 5 epochs in order to start training gently and finish with finer steps. The chosen **batch size is 32** samples per batch, and mixed precision is also employed to save memory and training time (Loshchilov and Hutter, 2016, 2017; Goyal *et al.*, 2017; Narang *et al.*, 2017).

As for the **loss function**, **BBFL** is applied as the primary since the dataset is slightly unbalanced. It is chosen as it aims at the hard examples and balances the contribution from each class within a given batch. The initial **starting point** for it is **1.5 – 2.0**. In such a way the class weights are calculated for every batch and the less frequent classes are not suppressed. A mix-up lightweight regulariser of 0.2 - 0.4 is also utilised to diminish the chance of overfitting and disregarding some anatomical features (Singh *et al.*, 2023; Zhang *et al.*, 2017).

As mentioned a bit earlier, **Early Stopping** is also employed based on the macro-F1 on the inner validation split. The **Patience** is set to **8 epochs** without improvement, where it stops and returns to the model variation with the best learned weights. Thus, checkpoints of the models are also saved for this purpose (Prechelt, 1998).

Table 4 - Default Training Hyperparameters Table.

Item	Default
Optimiser	AdamW (lr 2e-4 head, after 1e-4 full), weight_decay 1e-4
LR schedule	Cosine decay, 5-epoch warmup
Epochs	25–40 total (incl. 3–5 for linear probe)
Batch size	32 (AMP on)
Loss	BBFL (γ 1.5–2.0), with mixup α 0.2–0.4
Early stopping	patience 5 on macro-F1, restore best

3.4.7. Metrics and Model Selection Rule.

Since there is a small imbalance in the data and this is a classification of multiple pathologies, the chosen metrics must assess each class fairly by properly reflecting clinical risk like missed positives versus false alarms and exemplify if the probabilities can be trusted.

The **Primary Metric** is **Macro-F1**, because it combines precision (how often a positive prediction is correct) and recall/sensitivity (how many true positives are observed). The meaning of "macro" is that the F1 per class is computed and then averaged, so rare classes have the same weight as the ones that appear more often. Thus, it is chosen as the primary score because it balances evenly across all nine diagnoses (Powers and Ailab, 2020).

Other metrics that are reported are **precision, recall, F1 and specificity** and they are **addressed towards each class**. This is so, in order to highlight which classes need work (Powers and Ailab, 2020).

Balanced Accuracy is another metrics that is reported as an average of sensitivity and specificity per class, which are consequently macro-averaged. This is so, because this calculation is more informative than raw accuracy when classes are unbalanced (Brodersen *et al.*, 2010).

PR-AUC (macro) is the last one, which is area under the precision-recall curve. It is more informative that usual ROC-AUC when positives are rare. PR-AUC is emphasised for minority classes (Saito and Rehmsmeier, 2015).

The **Model Selection Rule** for the best performing model is as follows – **Macro-F1** which is calculated as a median across folds is the primary selection criteria. Where a higher one is better. If the results would be almost the same, meaning a difference of +/- 0.5 points, the higher value is preferred.

3.4.8. Experimental Setup.

Three experiments are run to prove the best model's explainability, diagnose errors, and provide a final and unbiased estimate of performance on unseen patients. These are:

1. **Grad-CAM++** - to prove where the model is focusing its attention in the image plane and explainability.
2. **Error analysis through confusion matrices and per-class PR curves**.
3. **An evaluation on a held-out test set**.

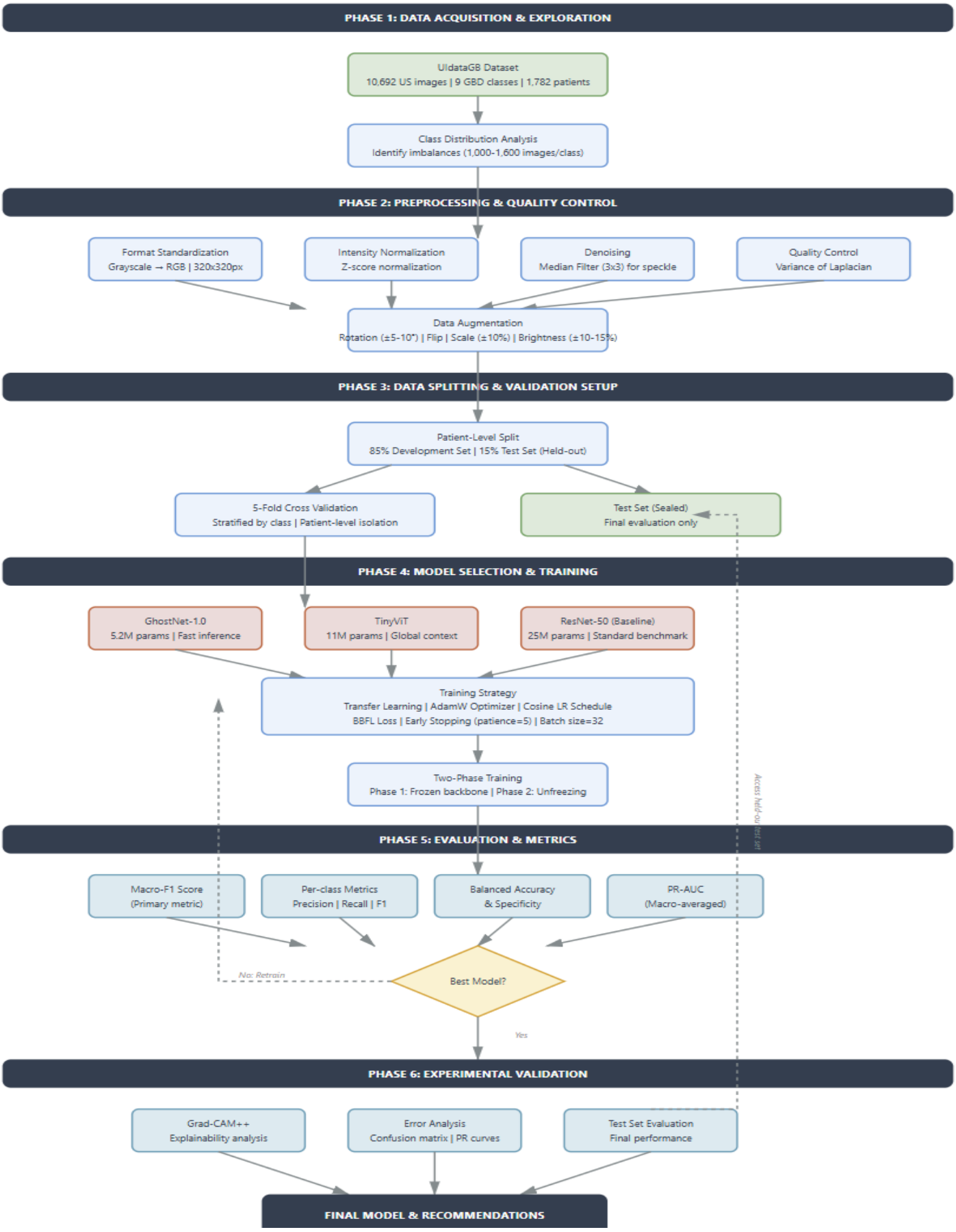


Figure 15 - Methodology Flow Chart.

3.5. Practical Considerations

1. **Computing Resources and Time** – the whole training and evaluation processes are run on a Google Colab notebook, with early stopping to most efficiently limit runtime and utilise resources. This is done to use the time and hardware resources effectively.
2. **Data handling, governance and ethics** – The open-source dataset is with removed identities and used a secondary data source.
3. **Deployment constraints** – due to limited time and resources the research is not able to deploy the best performing model on ARM or legacy medical devices.
4. **External validation** – for the same reasons as stated above, the forming of a multi-centre external validation teams to further experiment and validate the solution are not possible. This is another limitation to the study.

Some **Risk Management** techniques are applied to tackle some of the above limitations such as **BBFL and stratified folds** to handle **imbalances**; and **Cross Validation** at patient level combined with **early stopping** and **strict fold isolation** to tackle **Overfitting and Data Leakage**.

3.6. Theoretical implications

1. **Capacity versus Efficiency in US imagery** – results will map how far lightweight backbones can compete with heavier DL architectures in the classification task of multiple US samples of GB pathologies by clarifying their outputs.
2. **Value of limited preprocessing on US data** – the research will produce evidence on image normalisation on every image, mild speckle denoising and standardisation will exemplify if small and stable pipelines improve generalisation and diagnostic efficiency without losing the latter detail.

3. **Learning Strategy that is oriented towards imbalances and fair class metrics** – it is assumed that BBFL is going to improve the macro-F1 and the recall of smaller classes. Thus, uplifting the use of macro-averaged metrics as a default in clinical classification.
4. **Methodological implications** – the presented workflow consisting of patient-level 5-fold CV and leakage controls can become a standard for similar US problems.

4. IMPLEMENTATION, RESULTS AND ANALYSIS

4.1. Chapter Overview

The code for the current research, development and evaluation of models can be found through accessing this link: https://colab.research.google.com/drive/1gDyn_WXGn3repUKswgfRuNbnHzNo2Alf?usp=sharing .

The current chapter presents the precise implementation, training and development of the algorithms. It also delivers the changes made to the methodology due to problems that occurred along the way and the ways in which attempts were made, so they can be tackled.

4.2. Data Preparation

4.2.1. Exploratory Data Analysis (EDA)

Although, in the Methodology section, there is a relayed EDA that is derived from the release paper of the dataset, a brief EDA was performed to confirm the information (Turki *et al.*, 2024).

Class Balance and patient coverage (Figures 16 & 17). The dataset contains **10,692 US** samples across nine GBD pathological classes. The count of images for each class varies from 990 for “Various causes of gallbladder wall thickening” to 1,590 for "Carcinoma”. Hence, a moderate class imbalance is present. The patient “id_coverage” that was extracted is 100%,

which means that each sample can be linked to a patient record, it is important for the splitting and to prevent data leakage. This is important, because as mentioned earlier, **even mild imbalances can lead to learning bias and shift the networks attention towards more frequent labels and suppressing ones that are rarer, thus inflating overall accuracy** (Japkowicz and Stephen, 2002; Saito and Rehmsmeier, 2015; He and Garcia, 2009).

Total images discovered: 10,692

Class counts:

class	count
8Carcinoma	1590
1Gallstones	1326
4Membranous and gangrenous cholecystitis	1224
2Abdomen and retroperitoneum	1170
7Adenomyomatosis	1164
3cholecystitis	1146
5Perforation	1062
6Polyps and cholesterol crystals	1020
9Various causes of gallbladder wall thickening	990

dtype: int64
[patient_id] coverage: 100.0% (10692 / 10692)

Figure 16 - Class Balance and Patient Coverage.

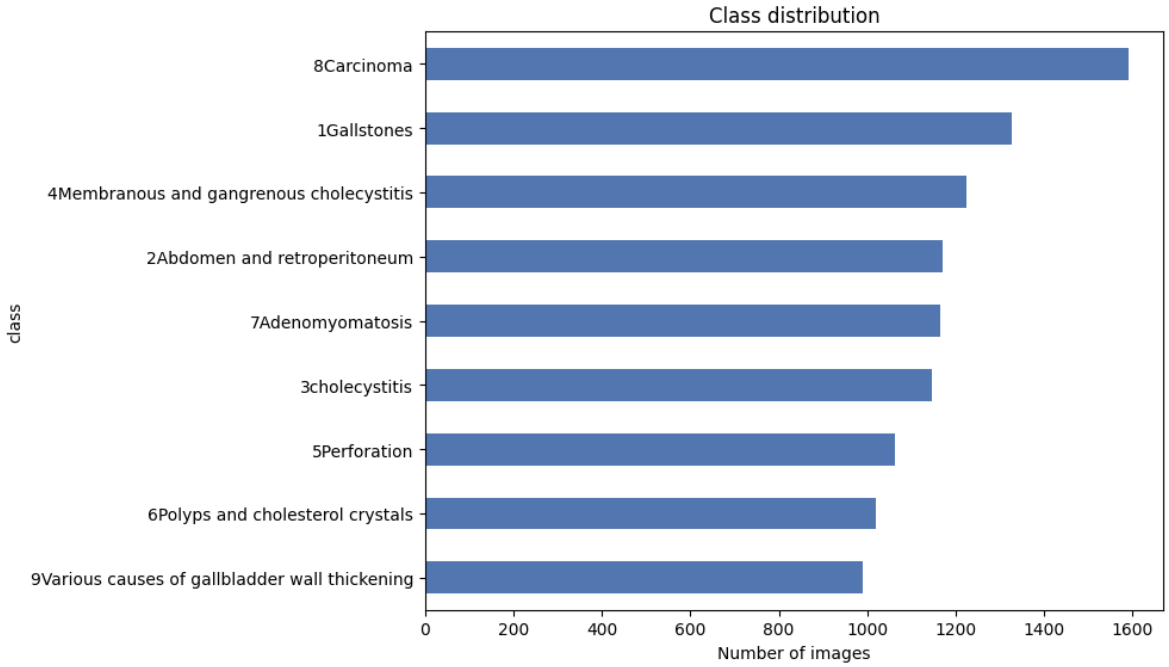


Figure 17 - Sample Distribution Across Classes.

The **File index snapshot** represented in **Figure 18**, depicts the first ten rows of the "Gallstone" class. The samples are consistent with **width of 1,200 pixels and height of 900 pixels**. The **aspect ratio** is also consistent as reported by the authors and is **4:3** across the images. The "patient_id" is the same for this batch, which is "a1", pointing to multiple samples taken from the same patient in one examination. The "class_idx" maps human labels like "1Gallstones" to integers. The repeated id's of the participants hint that several frames come from the same person, which is one of the reasons for the decision in the Methodology section to undertake a patient-level split, so it can be ensured that the model is not "seeing" the same anatomy in both train and validation (Turki *et al.*, 2024; Roberts *et al.*, 2021).

	filepath	class	filename	patient_id	width	height	aspect	class_idx
0	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (1).jpg	a1	1200	900	1.333333	0
1	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (10).jpg	a1	1200	900	1.333333	0
2	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (11).jpg	a1	1200	900	1.333333	0
3	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (12).jpg	a1	1200	900	1.333333	0
4	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (13).jpg	a1	1200	900	1.333333	0
5	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (14).jpg	a1	1200	900	1.333333	0
6	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (15).jpg	a1	1200	900	1.333333	0
7	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (16).jpg	a1	1200	900	1.333333	0
8	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (17).jpg	a1	1200	900	1.333333	0
9	/content/drive/MyDrive/AI/Project_Summer_Modul...	1Gallstones	a1 (18).jpg	a1	1200	900	1.333333	0

Figure 18 - First Ten Rows of the "1Gallstones" Class.

Figures 19 and 20 portray the image geometry and channels that are present in the GB dataset. The data is primarily consisting **of RGB-encoded grayscale samples**. The histograms (**Figure 20**) for width, height and aspect ratio are very similar and consistent. They show that almost all frames sit at 1200x900 pixels with an aspect ratio of 4:3 (1.33). **Only a small number** of images deviate and go up to **2400 by 1800 pixels**. The performed channel assessment of 1,000 frames showed that there are three (RGB) channels available in the samples, meaning that the US pictures are saved as RGB files, although the content is visually grayscale.

	width	height	aspect
count	10692.000000	10692.000000	10692.000000
mean	1198.919753	901.447811	1.330825
std	25.610173	23.172051	0.038181
min	900.000000	876.000000	0.750000
5%	1200.000000	900.000000	1.333333
50%	1200.000000	900.000000	1.333333
95%	1200.000000	900.000000	1.333333
max	2400.000000	1800.000000	1.335616

Figure 19 - Shape and Geometry of the Data in the Dataset.

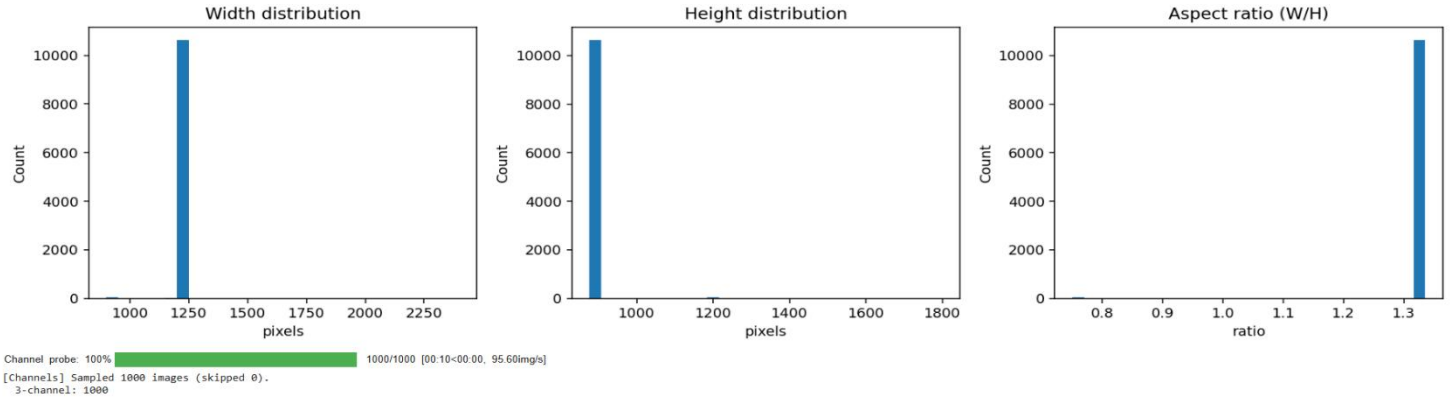


Figure 20 - Histograms of Dimensions of Images in the Dataset per 1,000 Random Samples.

Figures 21 and 22 depict the results of the performed **Variance of Laplacian scoring across all 10,692 available frames** at an image dimension of **320x320 pixels.** The higher values of the variance mean a sharper image with stronger edges and the lower ones hint of blurring present in the sample. The histogram (**Figure 22**) highlights a right skew with a small tail towards the lower variance. The textual output (**Figure 21)** suggests conservative levels at the 5-th. (104.6), 8-th. (122.8), and 10-th. (131.8) percentiles. Thus, a **provisional threshold was set at the 8-th. percentile of 122.8**, represented by the red dashed line. All available files were

included in the latter scoring. The test was performed to see the blur status of the data, as the US data is frequently plagued by motion blur or defocus, which can affect clinically important boundaries. Thus, training on such frames can cause negative consequences for the output of the algorithms (Pertuz *et al.*, 2013).

```
VoL scoring @320×320: 100% ████████████████████████████████ 10692/10692 [01:47<00:00, 99.
[VoL] Scored: 10692 images  |  Skipped: 0
Suggested VoL thresholds (dataset-wide; we will recompute per train fold):
  τ_5%  = 104.58
  τ_8%  = 122.76
  τ_10% = 131.84
[SET] Provisional blur threshold τ = 122.76 (8th percentile)
```

Figure 21 - Textual Output from Variance of Laplacian Scoring.

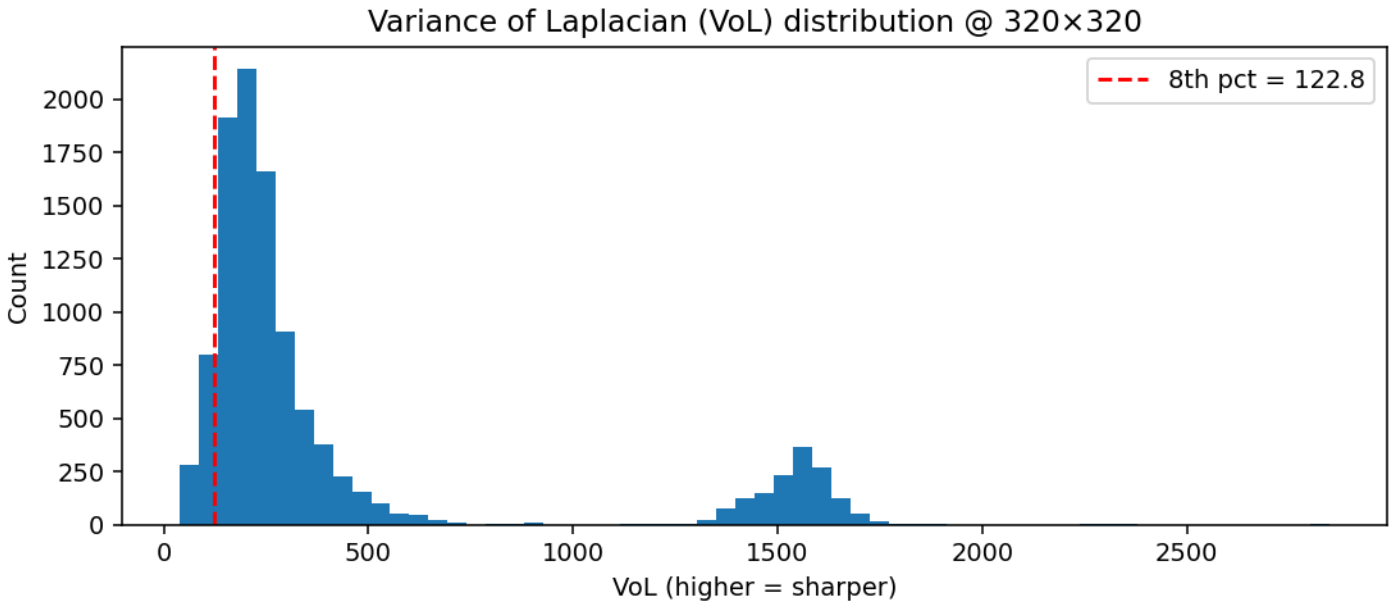


Figure 22 - Variance of Laplacian Distribution Histogram.

In **Figures 23 to 31** can be observed random plots of images from all nine pathology classes.

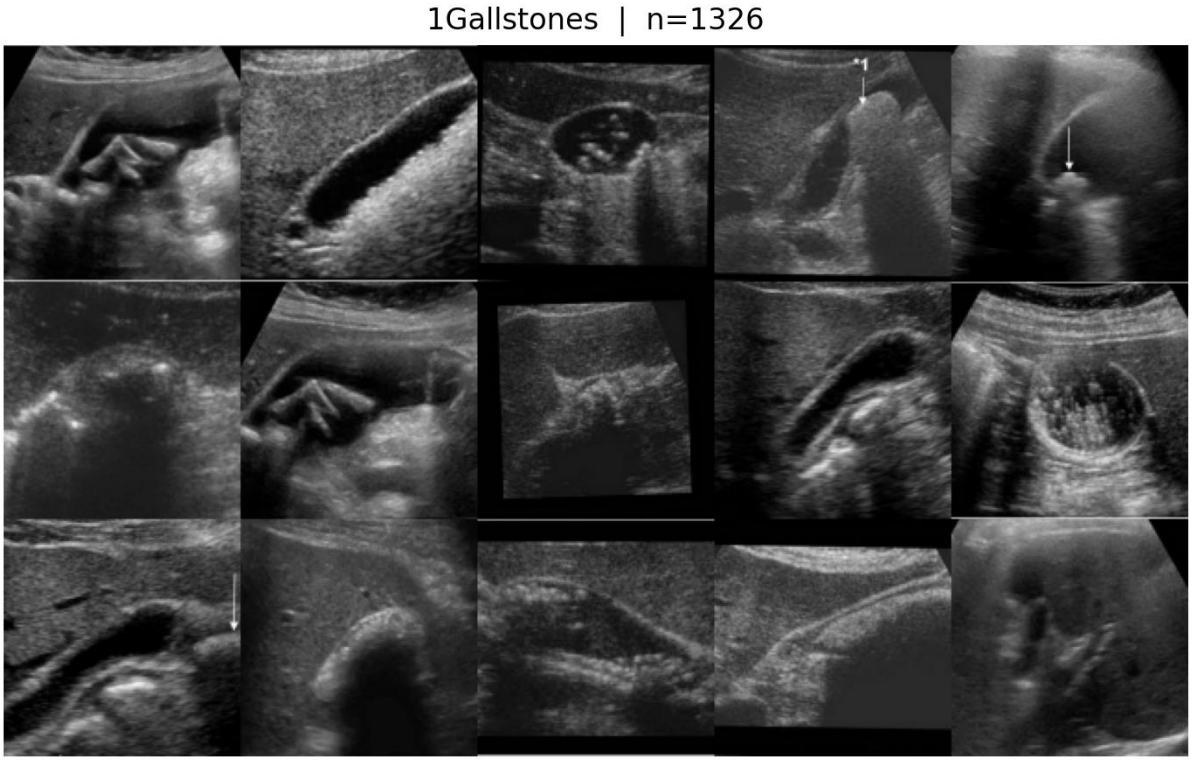


Figure 23 - Random "1Gallstones" Samples.

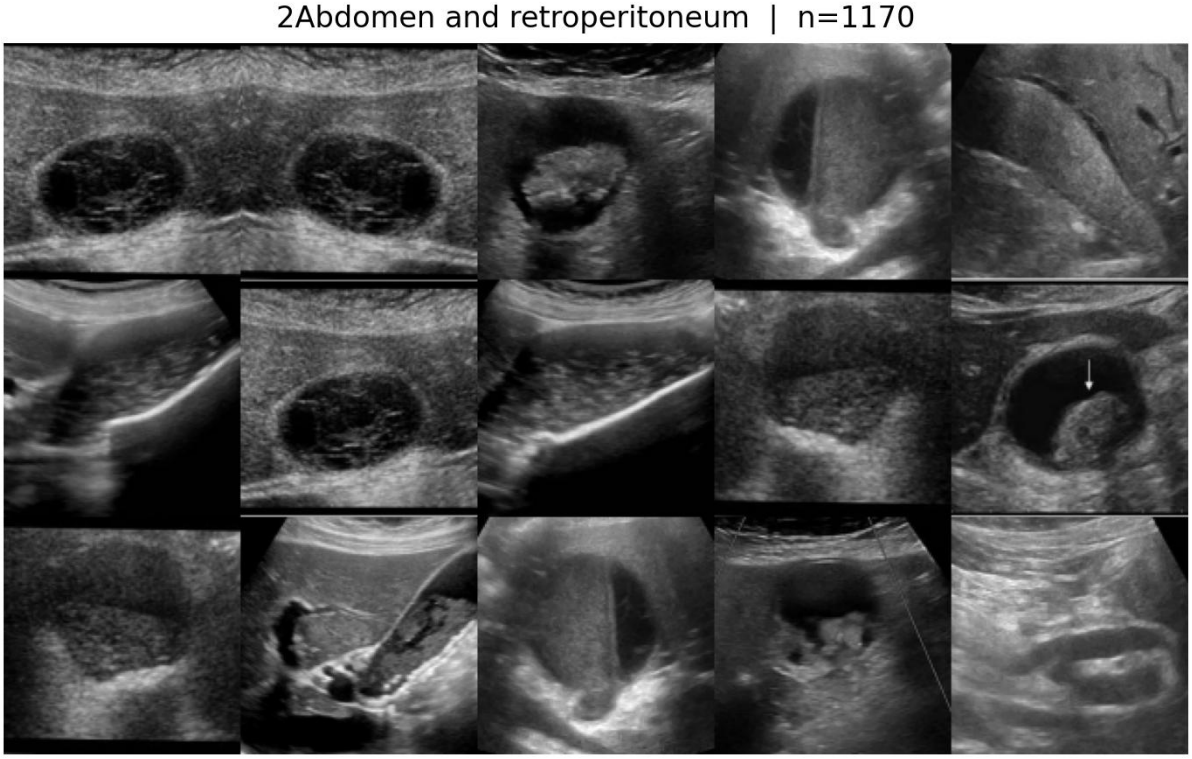


Figure 24 – Random “2Abdomen_and_retroperitoneum” Frames.

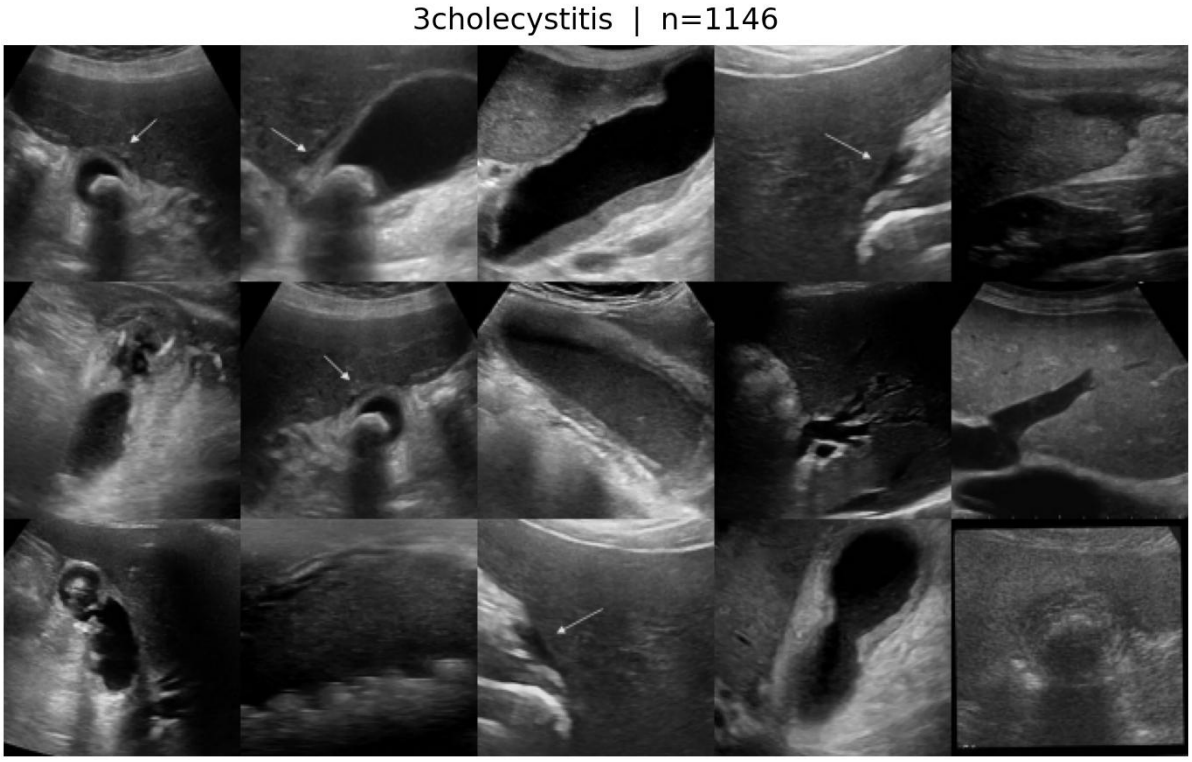


Figure 25 - Random "3cholecystitis" Frames.

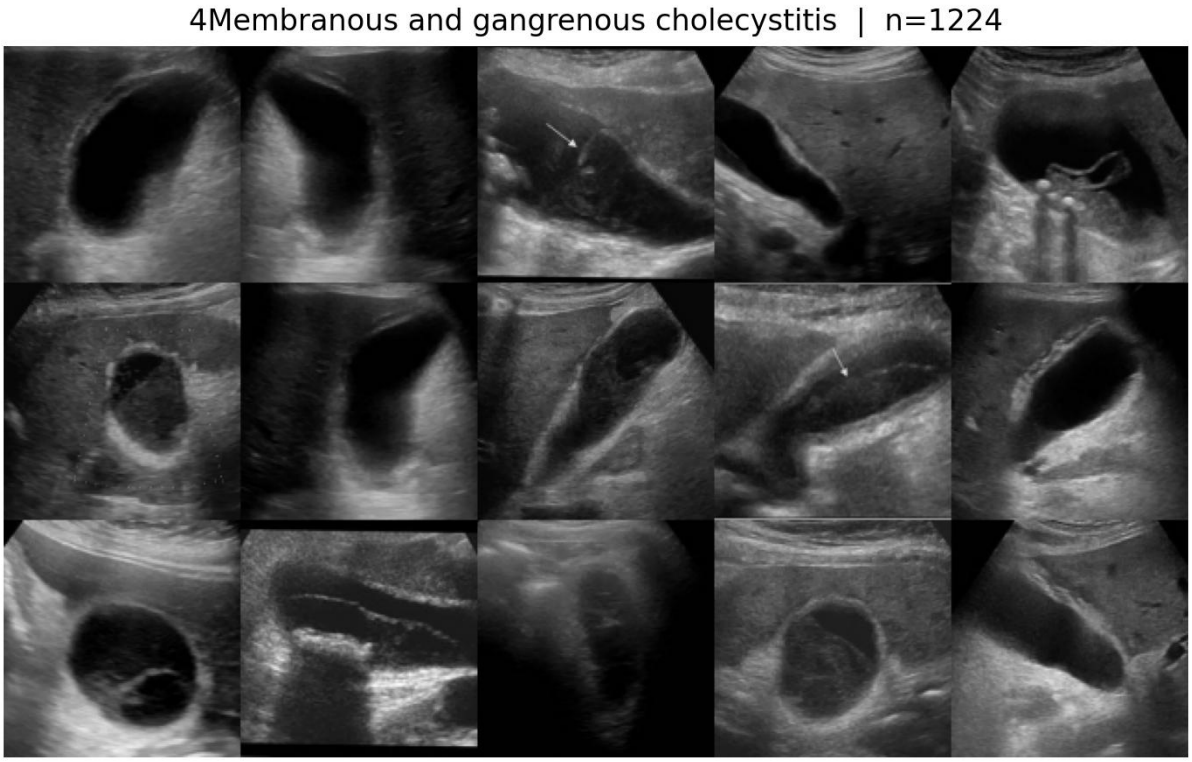


Figure 26 - Random "4Membranous_and_gangrenous_cholecystitis" Samples.

Figure 27 - Random "5Perforation" Samples.

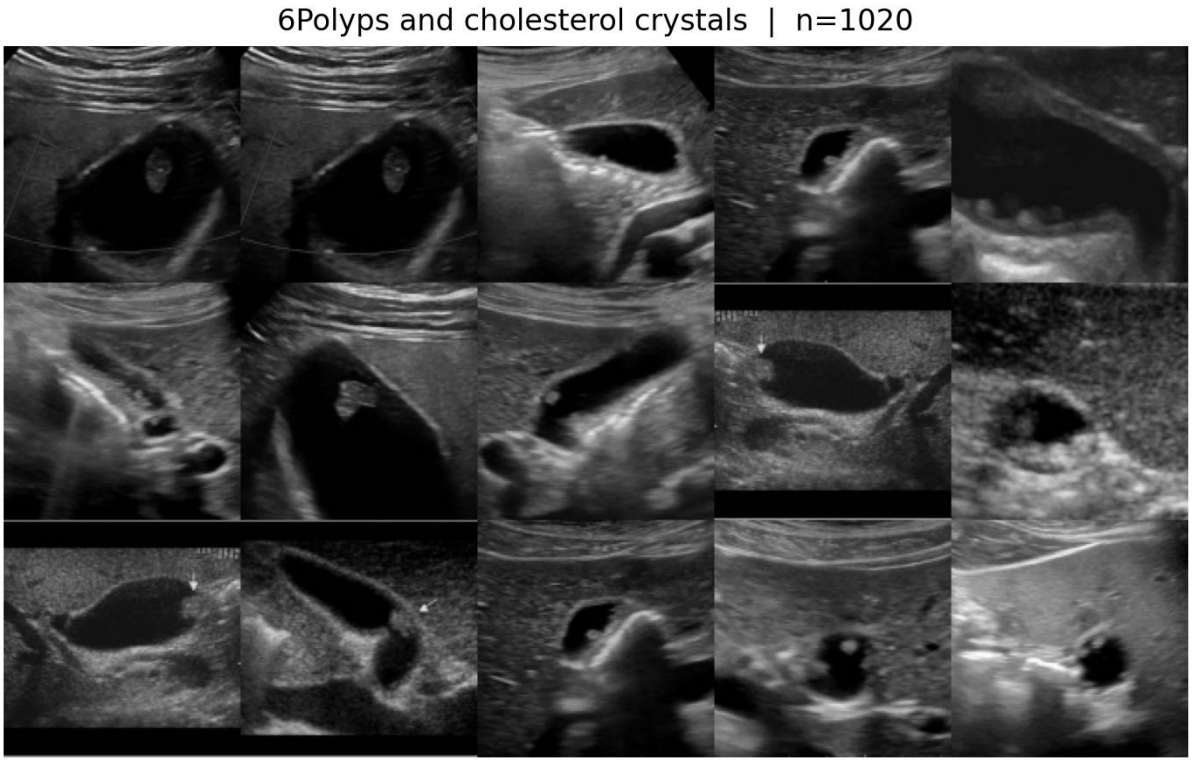


Figure 28 - Random "6Polyps_and_cholesterol_crystals" Frames.

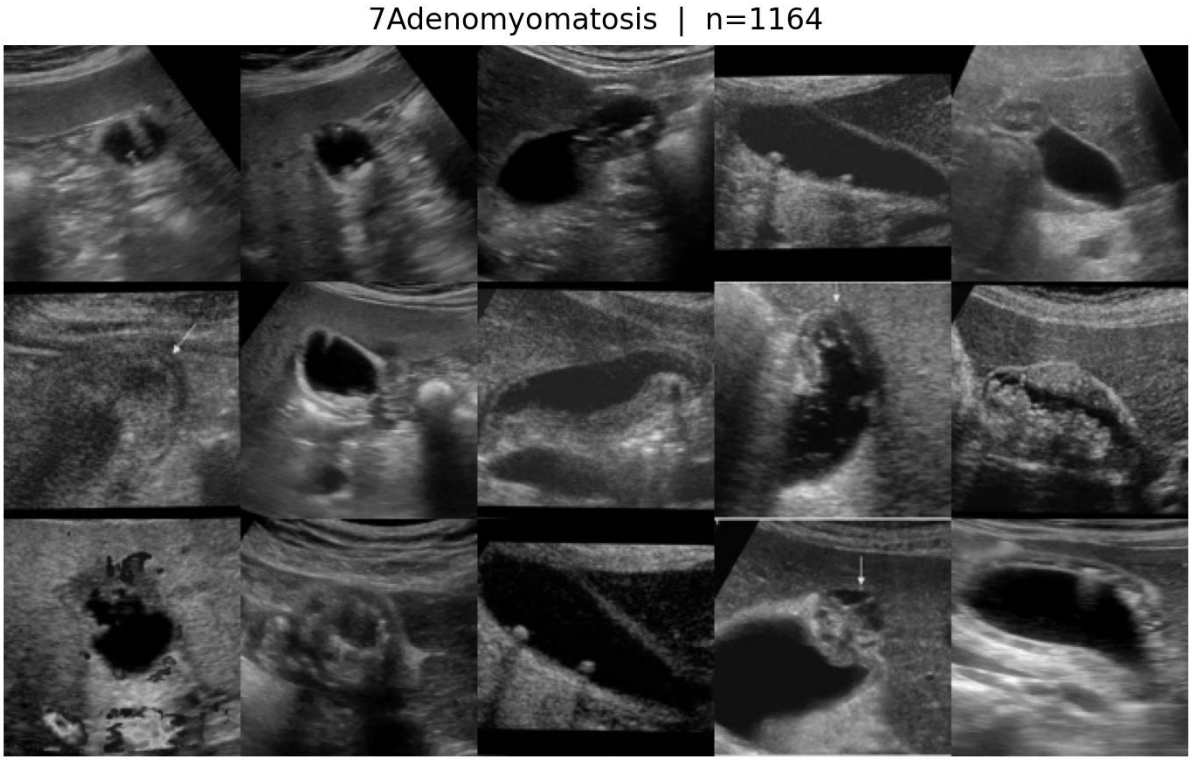


Figure 29 - Random "7Adenomyomatosis" Frames.

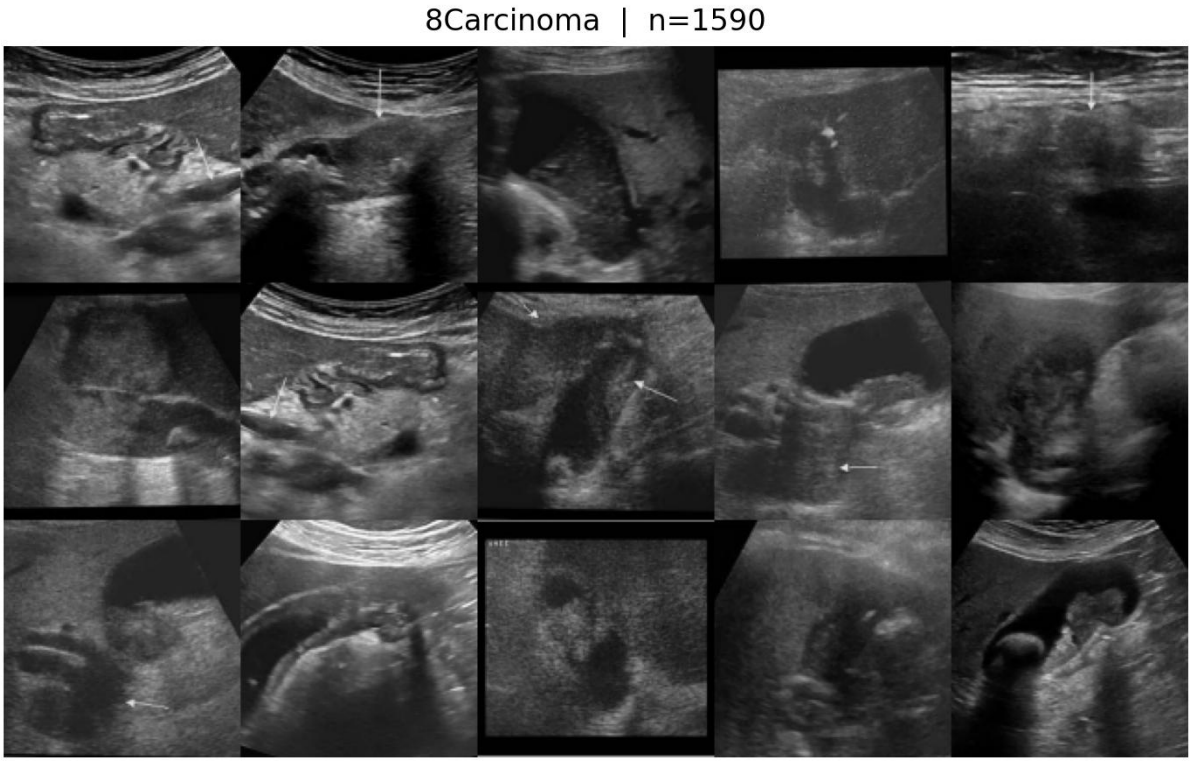


Figure 30 - Random "8Carcinoma" Samples.

Figure 31 - Random "9Various_causes_of_gallbladder_wall_thickening" Frames.

4.2.2. Data Preprocessing.

This implementation stage followed the steps mentioned in the Methodology section as planned. It standardised the US frames, removed obviously unhelpful images, and prepared tensors for the models. The whole pipeline is deterministic and re-runnable so that the numbers can be reproduced exactly.

Firstly, the **random seeds** were fixed for Python, NumPy and Pytorch to enforce deterministic CuDNN behaviour. It is known that such an approach slows down the training process but avoids the slight deviation in results with repetitive training trying to reproduce the original results of the algorithm (also known as ”run-to-run drift”). Afterwards, the images were **reduced to 320x320 pixels** to comply with the chosen algorithms' input requirements. The utilised **batch size was 32 and 2 to 4 workers per loader** were employed for a stable throughput on Colab and the Nvidia A100.

As stated in the Methodology, the raw images' aspect ratio was not stretched, but instead **each frame was letterboxed** to a square plane and the short side of it was padded. Thus, the anatomy of the organ is best preserved, and geometric artefacts are not lost (Shorten and Khoshgoftaar, 2019).

Since US contains speckle, like it was already mentioned earlier, overly aggressive filters can erase diagnostically useful texture. Therefore, a **3x3 median filter** was applied bringing in a very mild denoise that reduces isolated speckles without smoothing edges too much (Yu and Acton, 2002).

To help the models cope with everyday scan differences like how the probe is held or the machine settings, some very light **augmentations** were utilised:

- Horizontal flips of 0.5.
- Small shifts, scale and rotation of +/-10 degrees and +/- 10%.
- Some gentle brightness and contrast jitter of +/- 0.15.
- Then, each image was normalised through z-score.

The augmentations were planned as modest intentionally so subtle structures in the image plane are not distorted. The Z-score centred and normalised each frame so the models would see a comparable intensity range across different scanner machines and gain settings (Shorten and Khoshgoftaar, 2019).

Blur quality control techniques were also utilised, as planned. Frames that were **too blurry** and would cause label noise were **filtered out by scoring sharpness through the Variance of Laplacian (VoL) at 320x320 pixels**. Where the higher VoL highlighted sharper images and the lower one – blurrier ones. This was **only performed on the training split with 8-th. percentile** of the VoL scores and shots with values below that level were excluded. The **validation and testing images were left untouched**, so evaluation reflects real-world quality. Hence, in the development split 9,088 images were scored with a computed threshold of 122.8. Thus, **the QC step kept 8,361 frames and dropped 727 images (nearly 8%),** which matched the chosen percentile **(Figures 32 and 33)** (Pertuz *et al.*, 2013).

Thus, in such a manner the inputs are clean and standardised without losing on the variations in the pathologies that are clinically valuable. Also, the QC filter removed the worst 8% of frames while the augmentations expose the models to changes that improve generalisation. All thresholds are derived from the data and chosen inside the training fold to avoid leakage.

```
Dev=9,088 | Test=1,604
VoL: 100% 9088/9088 [01:46<00:00, 107.40img/s]
[VoL] Scored=9088 | Skipped=0
[QC] τ (p8) = 122.76
QC filter: 100% 9088/9088 [01:35<00:00, 103.24img/s]
[QC] Kept 8,361 | Dropped 727 (τ=122.76)
Train batch: torch.Size([32, 3, 320, 320]) torch.Size([32]) | dtype torch.float32
Classes in this batch: [0, 1, 2, 3, 4, 5, 6, 7, 8]
```

Figure 32 - Textual Output of the QC Filtering Step.

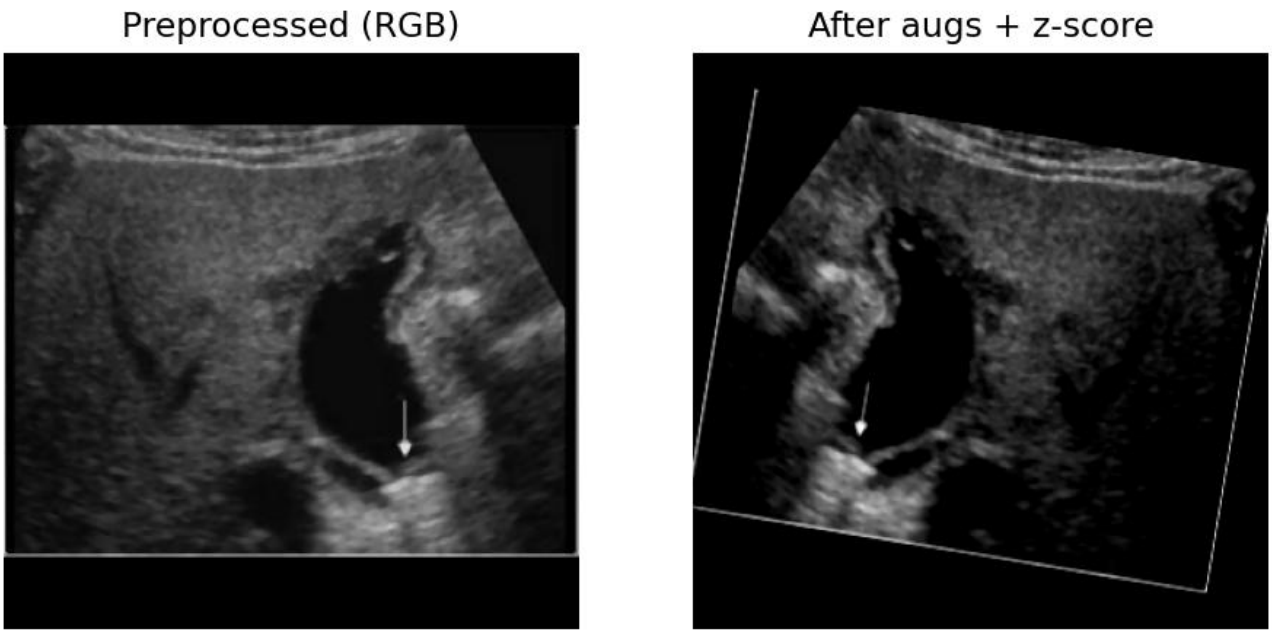


Figure 33 - Plot of Pre-processed (RGB) US Image and the Resulting Example After Augmentations and Z-scoring.

4.2.3. Data Splitting.

The idea behind planning this part in the Methodology was to evaluate the models fairly while **avoiding any patient leakage of data** – meaning the same patient has no chance of appearing

in both training and evaluation phases. The splitting procedure below enforces that separation and keeps the class balance.

Before splitting, the pre-processed images were written on a cached copy to Google Drive. Then, the **hold-out test set** for the experimentation phase of the project was separated through "**GroupShuffleSplit" with patient ID** as the grouping key to segregate 15% of the participants in the dataset as the final test set. Only the remaining 85% is used for training and validation as the development set. Then the generated .CSV file along the process is saved to the disk and never touched by pre-processing step. In such a manner is created a real deployment on unseen patient data and patient leakage is evaded (Roberts *et al.*, 2017; Varma and Simon, 2006).

The planned **Five-fold cross validation** was also applied on the development set with a **StratifiedGroupKFold** of k=5. Thus, when stratified by class each fold retains a similar class mix, which is important with mildly imbalanced data. A small meta.json file with the image size, QC percentile and counts was also written to provide a loader stub that reads the .csv files generated for the train.csv/val.csv from the cache. This is done to accommodate a resuming of the training session if Colab disconnects the session due to non-interaction with the platform.

```
[FOLD 1] τ=59.34 | train kept=6849 dropped=595 | val=1736
[FOLD 2] τ=60.27 | train kept=6769 dropped=589 | val=1822
[FOLD 3] τ=59.47 | train kept=6872 dropped=598 | val=1710
[FOLD 4] τ=58.46 | train kept=6624 dropped=576 | val=1980
[FOLD 5] τ=59.63 | train kept=6668 dropped=580 | val=1932
[OK] All folds saved under: /content/drive/MyDrive/AI/Project_Summer_Module/splits_5fold_v1
```

Figure 34 - Information of the Data Representations in the Folds.

4.3. Model Training.

The model training phase of this research work started with the plan outlined in the Methodology section. However, due to the unsatisfactory level of results achieved other training strategies were also tested to try to tweak and improve the models and the outputted deliverables. More than twelve different approaches to the training strategy were tested. Below are presented the most important and major ones.

4.3.1. Model Training with the Original Methodology Training Strategy and Approach.

During the original training and validation session and across all the folds, **the three models exemplified that the training macro-F1 reached nearly 1.0 after the backbone was unfrozen**, while the **validation one stalled around 0.20 to 0.33** and usually went down afterwards **(Figures 35, 36 and 37).** This is a showcase example of **overfitting**, where the network starts to memorise the training distribution but fails to generalise on unseen patterns. The effect is most obvious in GhostNet and ResNet. Things are more gradual in TinyViT which still ended up with a wide train and validation gap. The loss graphs are not presented because the values generated by the BBFL are computed small and when rounded they are close to zero (Goodfellow *et al.*, 2016).

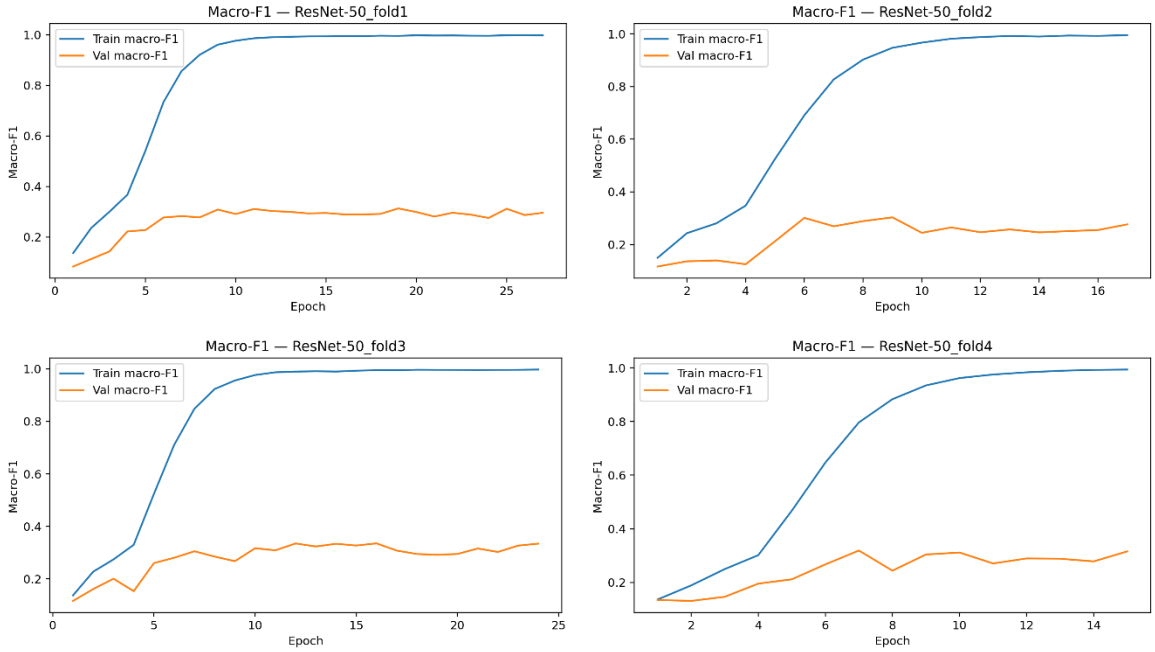


Figure 35 - ResNet-50 Training and Validation Macro-F1 Results with the Original Methodology Approach.

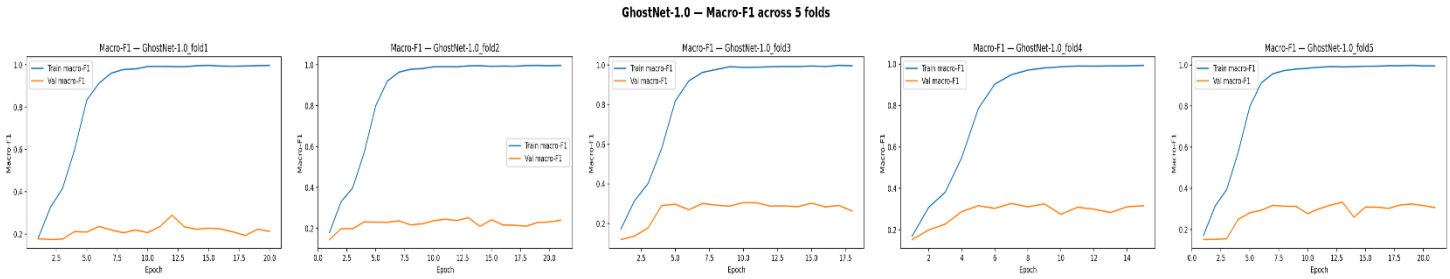


Figure 36 - GhostNet-1.0 Training and Validation Results with the Original Methodology Approach.

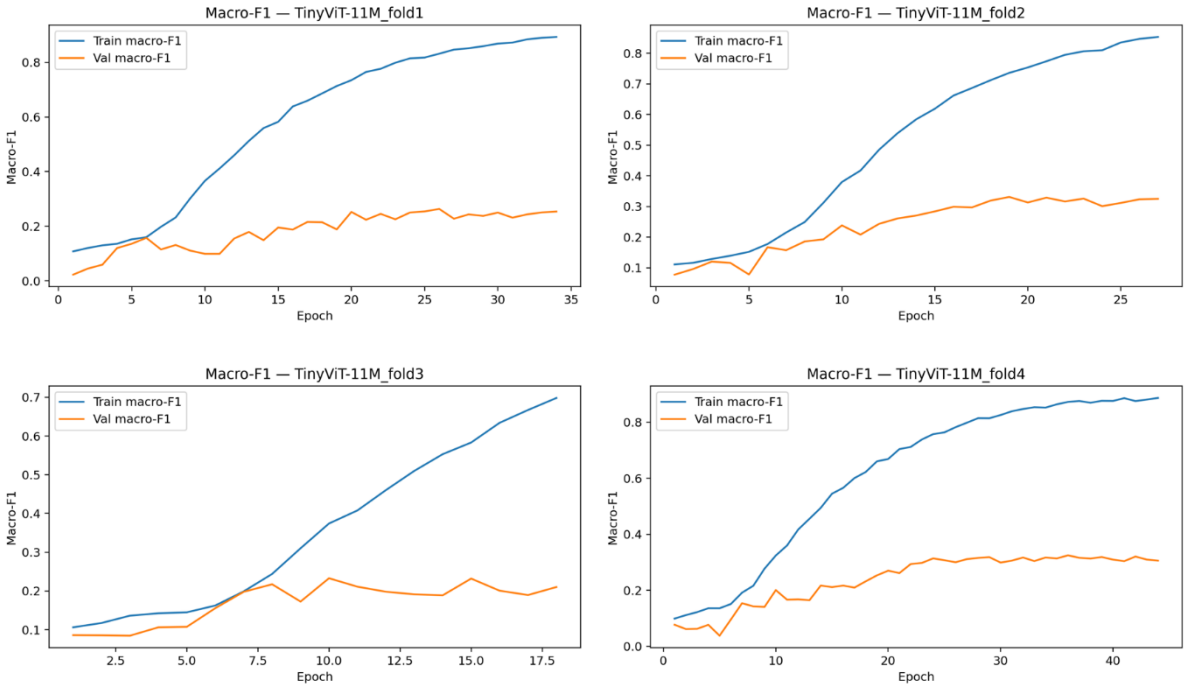


Figure 37 - TinyViT Training and Validation Results with the Original Methodology Approach.

Through an analysis of the results some additional research was conducted to highlight the possible culprit for the underperformance and possible ways to adjust the approach for better results. Thus, the following were highlighted:

- **Unfreezing of too many layers too early** – suspected because once a large part of the pretrained backbone was unlocked, the algorithm quickly started to memorise the data. Thus, potentially erasing the usefulness of the pretrained features (also known as “catastrophic forgetting”) (Kirkpatrick *et al.*, 2017).
- **Limited variability vs. strong capacity** – even though conservative augmentations were applied; US machinery still exhibits patient or style sameness. Thus, the networks learn false features instead of assessing the pathologies depicted. It is a common occurrence in medical imaging with modest datasets (Zech *et al.*, 2018).

- **Class Imbalance and decision boundary hardness** – Macro-F1 improves with a small pace because rarer classes remain hard to distinguish. The model achieves high training accuracy while it systematically misses the rare pathologies, which is penalised by the metrics.
- **A side effect of the QC filtering** – possibly the approach removes the softest frames from the training part of the folds and hence it narrows the distribution. This would make the validation relatively harder.

With **GhostNet-1.0**. (**Figure 36**) a very fast and intermittent convergence can be observed after the first unfreeze, followed by a major uplift in training F1 by epoch 6 to 8 in several folds. However, the validation F1 starts fluctuating early, then hits a plateau of around 0.30. This suggests that the head has learned linearly separable features on the seen distribution, but the deeper layers specialise too much when they are unfrozen.

In the case of **TinyViT** with 11 million parameters (**Figure 37**), there is a slower and steadier rise of training F1 while validation is playing up and down around 0.30 to 0.33 in the better folds, but the gap still remains. Despite the built-in regularisation, the model still overfits heavily.

ResNet-50 behaviour is similar to GhostNet's (**Figure 35**). There is an explosive training F1 once the backbone is unlocked, and a validation Macro-F1 which revolves around 0.28 to 0.34 on separate folds. Peaks in learning performance are short and are usually followed by a rapid decline, which again suggests overfitting.

Thus, **the verdict is consistent across architectures** – **they are overfitting and the validation macro-F1 stalls around 0.3 for all of them**. Hence, some follow up changes were made to the recipe to try and improve the overfitting and learning process.

In **Table 5** are presented the best results achieved by the algorithms during the original methodology training session.

Table 5 - Best Results Achieved by the Algorithms with the Original Methodology Approach.

Model	Best val macro-F1 (Approximately)	Fold	Epoch (Approximately)
GhostNet-1.0	0.33	5	12–14
TinyViT-11M	0.33	2	19–21
ResNet-50	0.34	3	23–24

4.3.2. Model Training with the First Changes in Training Strategy and Approach.

As mentioned above, there was a classic overfitting pattern across all architectures with the original methodological approach. To tackle this the following changes were made (Brownlee, 2019; Goodfellow *et al.*, 2016):

- **Stronger** but still light **augmentation** – a small CoarseDropout/Cutout (randomly masking a tiny patch) was introduced. This teaches the model to rely on distributed cues and be less sensitive to missing pixels or artefacts. This typically helps generalisation on small or imbalanced data.
- Keeping the **gradual unfreezing** and **discriminative learning rates –** Again, firstly the classification head was trained and then the blocks were a bit more gradually unlocked for learning while keeping the lower layers on small learning rates. This was done to address the “catastrophic forgetting”.
- **Mild regularisation** – a small dropout was added to the classifier layer and drop-path or stochastic depth where it was supported by the current backbone. This helps to reduce the adaptation of layers to the new domain.

In **Table 6** is summarised the changed Training strategy for this round.

Table 6 - First Changes in Training Strategy Table.

Component	Setting (this stage)	Notes / Rationale
Data augmentation	Baseline transforms + small CoarseDropout	Safe robustness to occlusions/artefacts.
Training schedule	Linear-probe (frozen backbone) to staged unfreeze (top to all)	Head settles first; backbone adapted gradually.
Learning rates	Discriminative LRs across groups	Higher LR for head, progressively smaller for lower layers.
Regularisation (head)	Dropout (light) - 0.2	Reduces over-confident head; combats overfit.
Regularisation (backbone)	Drop-path / stochastic depth (when supported)	Encourages path diversity; tames deep overfit.
Early stopping/checkpointing	Monitor validation Macro-F1; keep best; patience = 7 epochs	Picks the real best epoch .
Other training knobs	Optimiser/schedule unchanged from baseline	Changes isolate the effect of the strategy above.

After the changes were applied and the training process initiated and finished, **GhostNet-1.0** (**Figures 38 and 39**), exemplified that the training validation still climbed very quickly to a Macro-F1 of 1.0 by epochs 8 to 10, but the validation one was rising with a more steady pace and peaked at between 0.32 to 0.35 in different folds. The best result was achieved at fold 3 and epoch 34 where it hit 0.349 Macro-F1. The training loss dipped to almost 0 values, while the validation one was going upwards creating a big fork. Thus, **the changes that were introduced brought in slight improvements, but still not satisfactory results**.

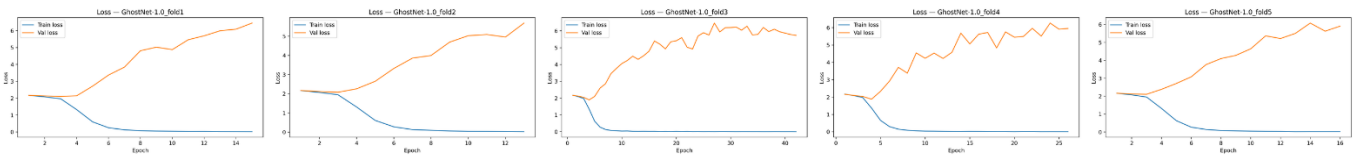


Figure 38 - Loss Curves of GhostNet with the First Training Changes.

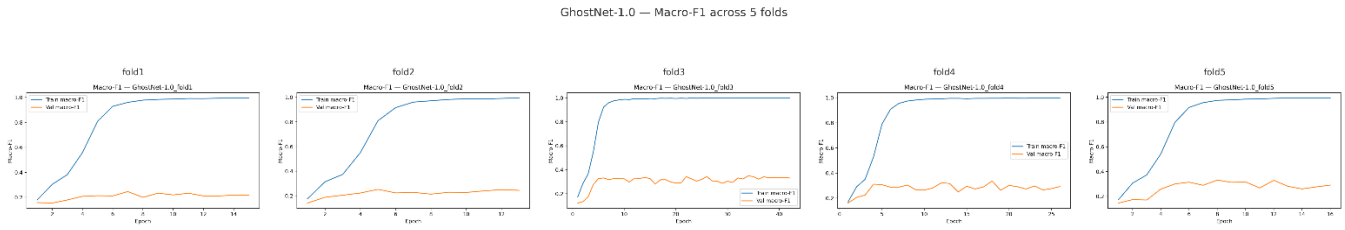


Figure 39 - Macro-F1 Curves of GhostNet with the First Training Changes.

On the other hand, the baseline – **ResNet-50** (**Figures 40 and 41**), reached results peaking in the 0.33 to 0.37 range across folds. The best fold was fold 2, where during epoch 21 the model hit 0.368 Macro-F1. The loss movements were similar to the previous model, but the plateau emerged a bit higher and held longer once the stage unfreeze was engaged. **Although here slight improvements are also observed, the results were still not nearly high enough**.

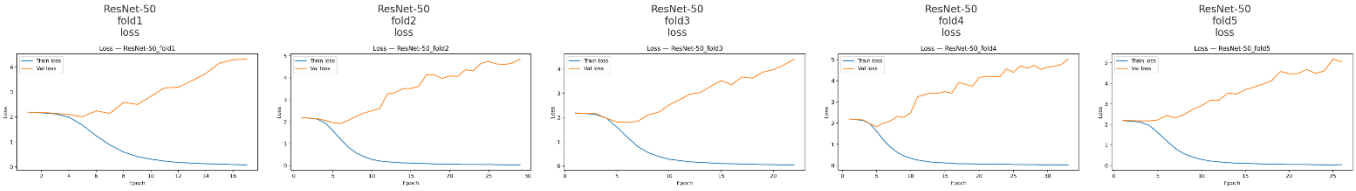


Figure 40 - Loss Curves of ResNet-50 with the First Training Changes.

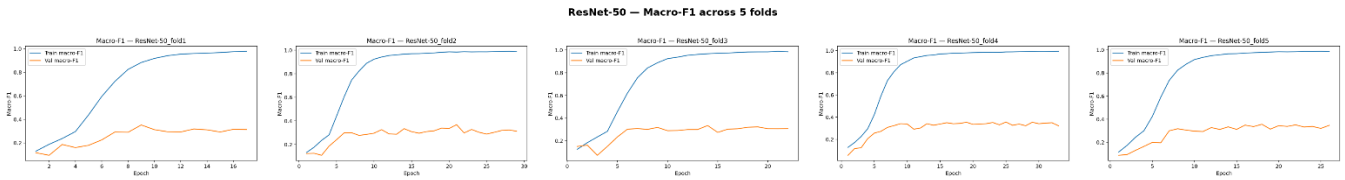


Figure 41 - Macro-F1 Curves of ResNet-50 with the First Training Changes.

Lastly, **TinyViT gained the most** from the new approach (**Figures 42 and 43**). The validation Macro-F1 started rising slowly early on, then kept improving later into the training after additional blocks were unfrozen. The best fold was fold 5 and epoch 24, where it hit the 0.369 mark. The training loss was gliding gown smoothly and the validation still climbed upwards, but slightly better than the previous approach. Thus, **the results were deemed unsatisfactory**.

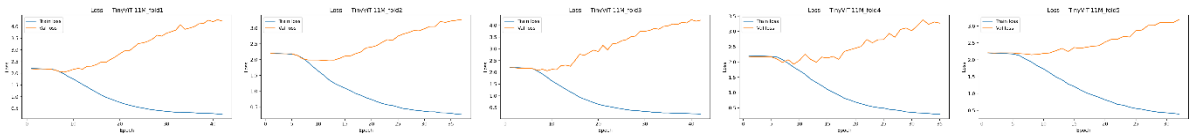


Figure 42 - Loss Curves of TinyViT with the First Training Changes.

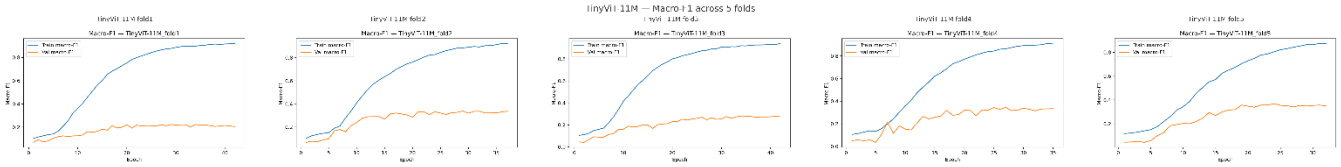


Figure 43 - Macro-F1 Curves of TinyViT with the First Training Changes.

Overall, **the new recipe delivered small but consistent gains in the validation results**. However, the results were still far away from state-of-the-art and (**Figure 44 and Table 7**) and **the suspicion that the fast unlocking of layers could still be the culprit** was still lingering in the air. Thus, more alternative approaches for a change were researched (Goodfellow *et al.*, 2016; Brownlee, 2019).

	model	fold	best_val_macro_f1	best_epoch
0	GhostNet-1.0	1	0.246358	7
1	TinyViT-11M	1	0.223066	34
2	ResNet-50	1	0.352394	9
3	GhostNet-1.0	2	0.252096	5
4	TinyViT-11M	2	0.338604	29
5	ResNet-50	2	0.368371	21
6	GhostNet-1.0	3	0.348992	34
7	TinyViT-11M	3	0.281155	34
8	ResNet-50	3	0.335023	14
9	GhostNet-1.0	4	0.336319	18
10	TinyViT-11M	4	0.344874	27
11	ResNet-50	4	0.356552	25
12	GhostNet-1.0	5	0.332036	8
13	TinyViT-11M	5	0.369353	24
14	ResNet-50	5	0.356210	18

Figure 44 - Best Macro-F1's Achieved through Different Folds and Epochs with First Training Strategy Changes.

Table 7 - Best Macro-F1's Achieved with the First Training Strategy Changes.

Model	Best Val Macro-F1	Fold	Epoch
GhostNet-1.0	0.349	3	34
ResNet-50	0.368	2	21
TinyViT-11M	0.369	5	24

4.3.3. Model Training with the Second Changes in Training Strategy and Approach.

Like it was mentioned earlier the first round of tweaks improved stability and validation loss slightly but still tended to creep up while the training curves were not even nearly depicting perfection. That would usually mean that the pretrained backbone is being unfrozen too fast (Goodfellow *et al.*, 2016; Brownlee, 2019).

Hence, some improvements (**Table 8**) were made as follows(Ioffe and Szegedy, 2015; Li *et al.*, 2018; Szegedy *et al.*, 2015; Loshchilov and Hutter, 2016; Zhang *et al.*, 2017):

- **Gradual and partial unlocking of the backbone** – firstly, unfreeze only the top 10%, then top 30%. This was implemented to limit the risk of forgetting the pre-learnt features and introduce less shock into the training process.
- The **learning rate for the backbone was reduced, and longer warmup was provided** – in such a manner the classifier layer adapts quickly while each deeper layer of the backbone changes with more caution. Hence, the fine-tuning stability should improve, when deeper and more general layers are less impacted.
- **L2-SP regularisation on the unfrozen backbone weights** – it adds a penalty if the weights that are undergoing tunning move far from their ImageNet initialisation. Thus, the model is restricted to pretrained norms while still adapting to the new domain.
- **MixUp plus staging** – the approach diminishes over confidence and picks the true best epoch.

Theoretically, this should only let the last part of the backbone move forward very slowly, and constrained by the L2-SP, the models should retain the powerful pretrained features intact

while the classifier and top layers learn the new domain. Thus, tackling the suspected failure points.

However, after the training was initiated and two rounds of folds were performed, the training process was not resulting in a major uplift in outputs. That is why the training procedure was interrupted to save on computational resources.

Table 8 - Configuration of Second Changes in Model Training Strategy.

Component	Setting we use in this stage	Rationale
Unfreezing schedule	Start top 10% then top 30% of layers	Avoid abrupt feature drift; preserve generic features in early blocks
Linear-probe phase	Train head only for 3-5 warm-up epochs (same as baseline)	Let the head "find its footing" before touching the backbone
Backbone LR	Very-low, Oriented towards layers LR decay factor of 0.7 per stage;	Deeper layers move least; head adapts fastest
LR schedule	Cosine decay with longer warm-up (10–15% of total epochs)	Smooth optimisation; reduces early instability
BatchNorm	Frozen in the backbone (stats + affine)	Prevent small-batch stat drift
Regularisation (backbone)	L2-SP	Tethers weights to ImageNet initialisation (Li et al., 2018)
Regularisation (criterion)	MixUp	Softer targets and better calibration (Zhang et al., 2018; Szegedy et al., 2016)
Other keeps	Discriminative LRs across groups; checkpoint best val Macro-F1	Matches earlier protocol; picks true best epoch

4.3.4. Model Training with the Last Changes in Training Strategy.

In the final attempt **(Table 9)** to improve results delivered by the algorithm and resolve the issues that were suspected to cause them, the following last changes were applied (Buda *et al.*, 2018; Xue *et al.*, 2011; Nia and Shih, 2024; Shorten and Khoshgoftaar, 2019):

- **Data Input was switched to grayscale** (single channel).
- **Added conditional image repairs for each image in the data** – CLAHE when contrast was low and Non-Local Means (NLM) denoising when noise with higher frequency was present.
- **Additional augmentation with Gaussian blur** – p = 0.3.
- **Disabled MixUp** entirely.

- Added **WeigthtedRandomSampler** on the training loader to address the imbalance in classes.
- **Kept early stopping and patient-level splits**.

These were used because (Buda *et al.*, 2018; Xue *et al.*, 2011; Nia and Shih, 2024; Shorten and Khoshgoftaar, 2019):

- Grayscale avoids colour artefacts and betters modality matching.
- CLAHE can rescue faint structures.
- NLM removes noise while preserving edges.
- Weighted sampling uplifts exposure to rare classes and frequently betters macro-averaged metrics in imbalanced information sets.

Table 9 - Final Change of Strategy Attempt Parameters.

Component	Setting (this phase)	Rationale
Input channels	1 (grayscale) with backbones instantiated as in_chans=1	Match modality; simplify intensity distribution
Dataset repairs	CLAHE when per-image contrast low; NLM denoising when high-pass energy high	Enhance faint structure; reduce noise while preserving texture (Zuiderveld, 1994; Buades et al., 2005)
Augmentation (extra)	Gaussian blur, p = 0.3	Robustness to motion/defocus blur (Shorten & Khoshgoftaar, 2019)
Class imbalance	WeightedRandomSampler (weight = 1 / class_freq, with replacement) on train only	Raise minority exposure; targets Macro-F1 (Buda et al., 2018)
MixUp	Disabled	Avoid label mixing for subtle, local cues (Zhang et al., 2018)
Optimiser / scheduler	Unchanged from previous phase (discriminative LRs, warm-up, cosine)	Keep optimisation constant to isolate effects
Early stopping & selection	Best validation Macro-F1, patient-level splits; checkpointing unchanged	Optimise directly for the report metric

For all three architectures and across all folds **(Figures 45, 46 and 47)**, the training loss decreased smoothly while the validation loss or slightly declined for a number of epochs. However, the overall loss metrics acquired were still very high and above 1.0.

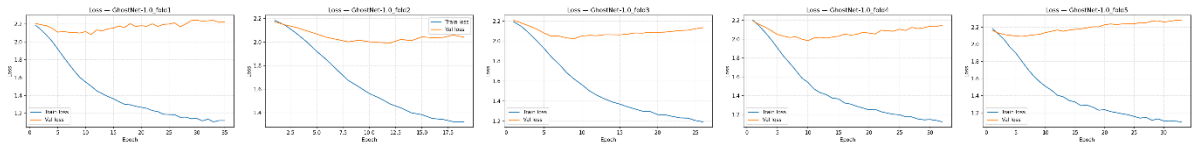


Figure 45 - GhostNet Validation Loss with the Last Training Strategy Attempt.

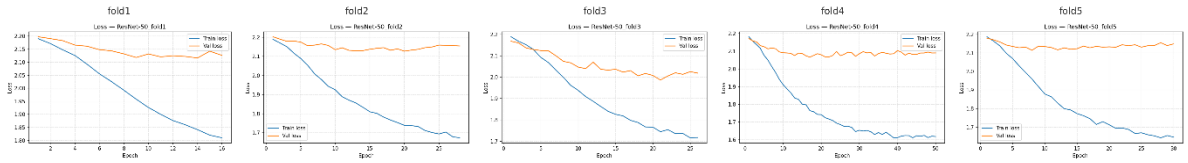


Figure 46 - ResNet-50 Validation Loss with the Last Training Strategy Attempt.

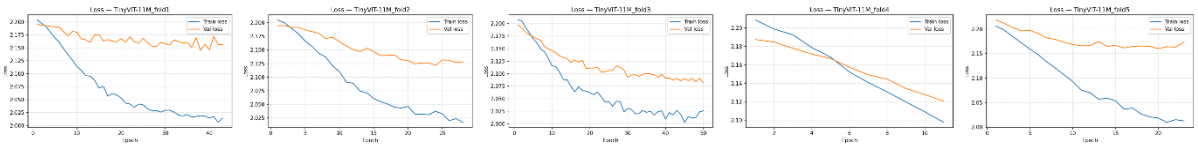


Figure 47 - TinyViT Validation Loss with the Last Training Strategy Attempt.

The Macro-F1 dropped when compared to the previous phase:

- GhostNet validation Macro-F1 plateaued around 0.20 to 0.23. **(Figure 48)**
- ResNet-50 was fluctuating around 0.16 to 0.18. **(Figure 49)**
- TinyViT was the most affected performing at the range of 0.10 to 0.14. **(Figure 50)**

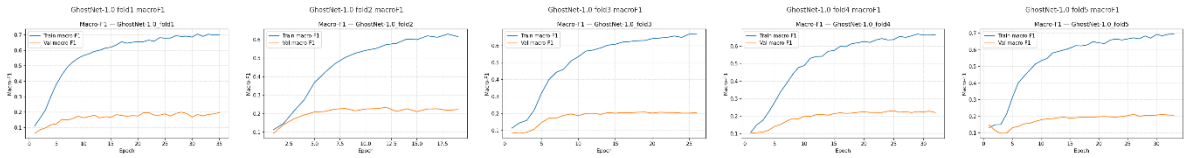


Figure 48 - GhostNet Validation Macro-F1 with the Last Training Strategy Attempt.

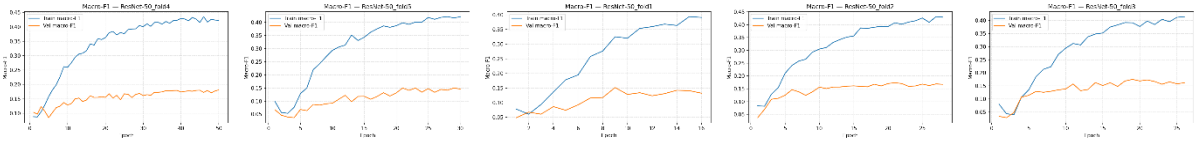


Figure 49 - ResNet-50 Validation Macro-F1 with the Last Training Strategy Attempt.

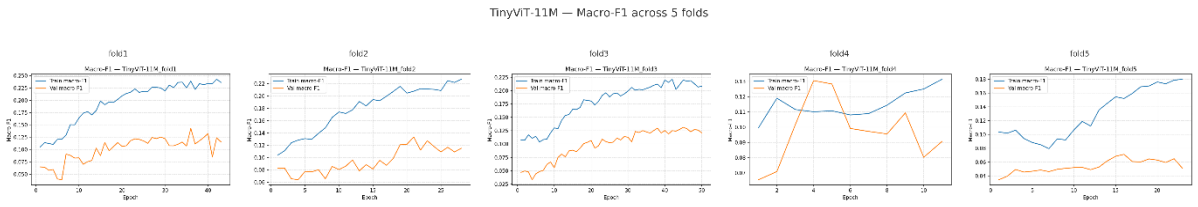


Figure 50 - TinyViT Validation Macro-F1 with the Last Training Strategy Attempt.

The more stable loss and the dip in Macro-F1 suggest the models became more calibrated but also more conservative. This is probably affected by the swap to 1 channel input and the disabling of MixUp.

GhostNet was the best of the three under this setting with an F1 of 0.23 **(Table 10)**.

Table 10 - Final Changes Best Algorithm Results

Model	Best Val Macro-F1	Fold	Epoch
GhostNet-1.0	0.234	2	12
ResNet-50	0.181	4	8
TinyViT-11M	0.143	1	36

5. EXPERIMENTS AND RESULTS

5.1. Chapter Overview

Although, unsatisfactory results were achieved during the training and validation of the models in the previous chapter, the methodologically planned experiments were still conducted with the best performing model from the original methodological approach – GhostNet-1.0 under the fifth cross validation fold. Thus, the current chapter explores the conducted experiments and the results derived from them.

5.2. Experiments

The conducted experiments are the ones planned in the original methodology, namely:

- **Confusion matrix and PR Curves.**
- **GradCAM++.**
- **A test experiment on the held-out test dataset.**

5.3. Results from the Tests

5.3.1. Results from the Evaluation Experiment on the Held-out Test Dataset

The evaluation test was performed on the **held-out test set** comprising of **1512 US samples** with size of **320x320 pixels**. As mentioned earlier, the samples were according to the patient-level requirement, and **no data had leaked towards it through the splitting process**. The

GhostNet model managed to achieve an averaged **macro-F1 of 0.272**. The metrics that were calculated for each class can be observed in **Table 11,** below.

Table 11 - Metrics Results from the Evaluation Experiment Performed on the Held-out Test Set.

[RESULT] Averaged Macro-F1 (TEST): 0.2724

	class	support	precision	recall	f1	specificity
0	1Gallstones	402	0.286458	0.136816	0.185185	0.876577
1	2Abdomen and retroperitoneum	24	0.000000	0.000000	0.000000	0.903898
2	3cholecystitis	168	0.157895	0.160714	0.159292	0.892857
3	4Membranous and gangrenous cholecystitis	132	0.500000	0.151515	0.232558	0.985507
4	5Perforation	222	0.205224	0.247748	0.224490	0.834884
5	6Polyps and cholesterol crystals	54	0.268519	0.537037	0.358025	0.945816
6	7Adenomyomatosis	114	0.219512	0.315789	0.258993	0.908441
7	8Carcinoma	228	0.420290	0.508772	0.460317	0.875389
8	9Various causes of gallbladder wall thickening	168	0.606667	0.541667	0.572327	0.956101

Hence, **class 9 – Various causes of gallbladder wall thickening** was highlighted as **the best performer** with an approximate F1 of 0.57, precision of 0.61, recall 0.54, and specificity of 0.96. This is the only class where a reasonable balance can be found between precision and recall. Thus, hinting that the model was able to learn consistent texture and boundary cues for wall thickening.

Following was **Class 8 – Carcinoma**, which achieved an F1 of 0.42, precision of 0.42 and recall of 0.51. As it can be seen, the recall was slightly higher than precision, which suggests that the model finds many of the positives but also generates a good number of false alarms.

The **third** best results were exhibited by **class 6 – Polyps and Cholesterol Crystals** with an F1 of 0.36, precision of 0.27 and recall of 0.54. Thus, the model flags polyps way too often. It finds many real cases, rectified in the higher recall, but also makes more false alarms, as shown with the lower precision. This is possibly linked to the appearance of polyps as small bright spots or minor wall bumps in data.

Classes 7 – Adenomyomatosis, 4 – Membranous or gangrenous cholecystitis, 5 – Perforation, 3 – Cholecystitis, and 1 – Gallstones showed F1s between 0.16 and 0.26. Frequently, these categories can be confused with each other or with findings related to wall thickening.

Class 2 – Abdomen and Retroperitoneum showed an F1 of 0. Which indicates that the model never actually predicted the class.

If the global behaviour of the model is concerned, it is observable that the specificity is high across all classes, while recall is mostly modest. Hence, it can be said that the algorithm rarely labels a sample as a specified class unless it is confident, thus keeping false positives low. This is reflected in the high specificity measures. However, it misses many true cases as shown with the low recall.

Potential causes to these results can be:

- The **patient-level splitting served its purpose** – kept close images and samples from the same patient leaking into subsets, thus providing a more honest result rather than inflating the numbers.
- **Class imbalance** – Class 7 collapsed to zero recall and the ones with medium data availability also show unstable precision and recall. Although, the metrics are balanced

the macro-F1 still penalises such dips and that can bring the overall averaged results down.

- **Classes are visually similar** – thus, without an approach applied to segment the sample and form a ROI, the algorithm can struggle to differentiate between similar features.

Hence, some **future improvements** can be made:

- If more data with smaller stones is acquired, the algorithm may improve recall by properly distinguishing between polyps and stones.
- ROI algorithm to localise and remove non-useful background.

5.3.2. Results from the Confusion Matrix and PR Curves Experiment.

The confusion matrix is normalised by row, thus each one sums to 1. Hence, each cell shows how many of all true class samples from a certain class were predicted as another class. The following curves are drawn per class together with the average precision (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015).

Thus, the delivered confusion matrix (**Figure 51**) shows that the model was strongest at class 9 – Various gallbladder wall thickening scoring nearly 0.54. Followed by class 6 – Polyps and cholesterol crystals with nearly the same result of approximately 0.54. And the third ranking class is 8 – Carcinoma with 0.51. Thus, these categories show relatively easy to spot global texture and changes of the organ pattern, that the network finds easier to learn.

However, the model seems to struggle with class 1 – Gallstones where exemplifies low recall of 0.14 with most of the predictions flowing to wall change classes. Overlap is likely to make

such features similar to the latter. Class 2 – Abdomen and retroperitoneum is also rarely recognised and frequently misjudged for other classes. The classes from 3 to 5 are also being confused with one another.

A good amount of the errors from various rows end up in columns eight and nine, which highlights the model is exhibiting bias towards the easier to spot pathologies and ones that are showing larger deviations in the organ structure. Thus, when it is unsure it leans towards these predictions (Saito & Rehmsmeier, 2015).

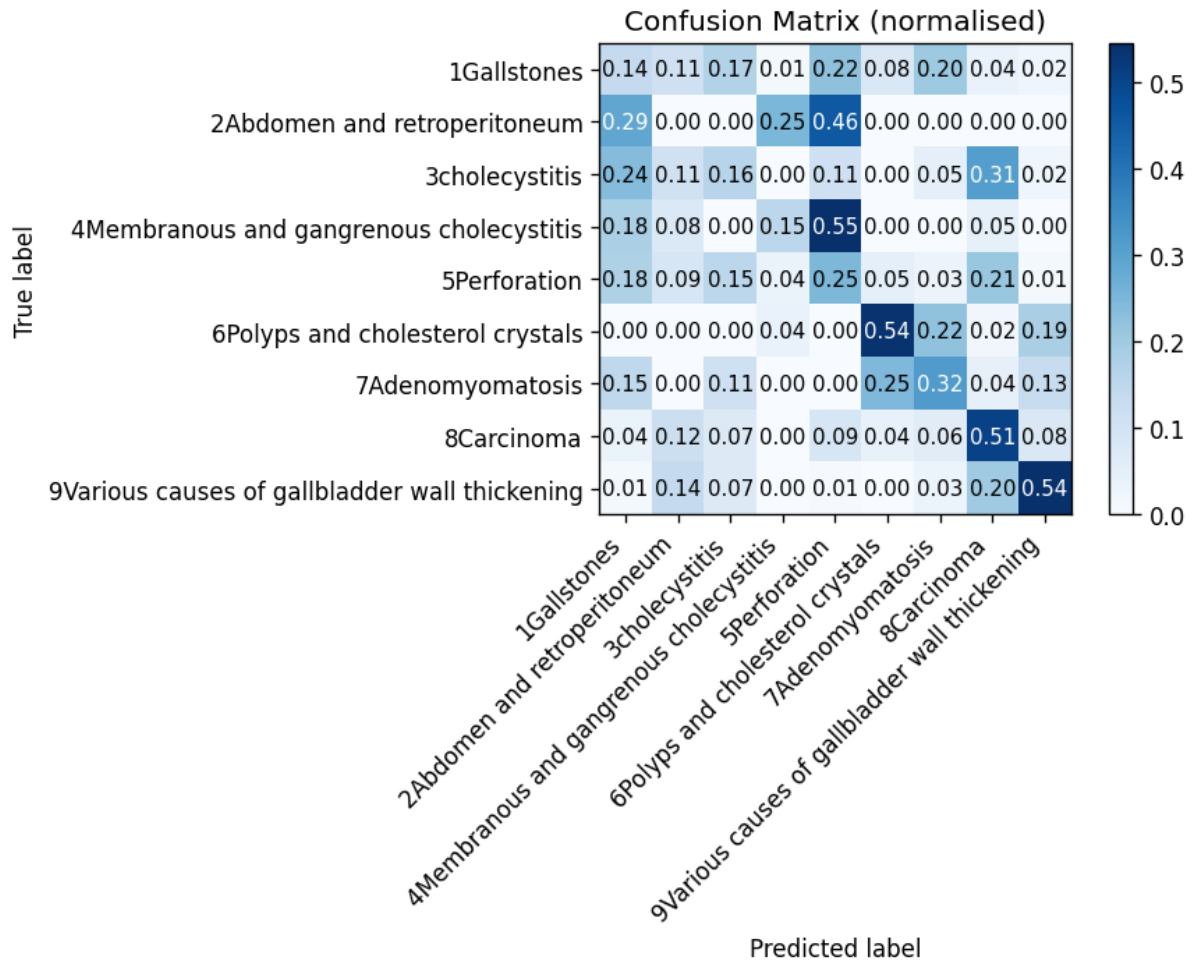


Figure 51 - Confusion Matrix.

The PR curves **(Figure 52)**, on the other hand, show that class 9 is delivering an average precision (AP) of 0.621 and class 8 – AP of 0.403. They are closely followed by class 4 with 0.381 and class 1 with 0.306. These curves depict moderate precision as recall increases, which indicates the model can retrieve many positives before precision collapses. In the middle are classes 5 with 0.204, 6 with 0.223 and 7 with 0.209. They draw shallow curves where precision lingers between 0.2 to 0.35 for a broader recall range. The weakest AP is shown by class 2 – 0.012. This would mean that the classifier did not manage to learn good features for the class.

This confirms the findings and possible solution presented in the previous subsection of this report.

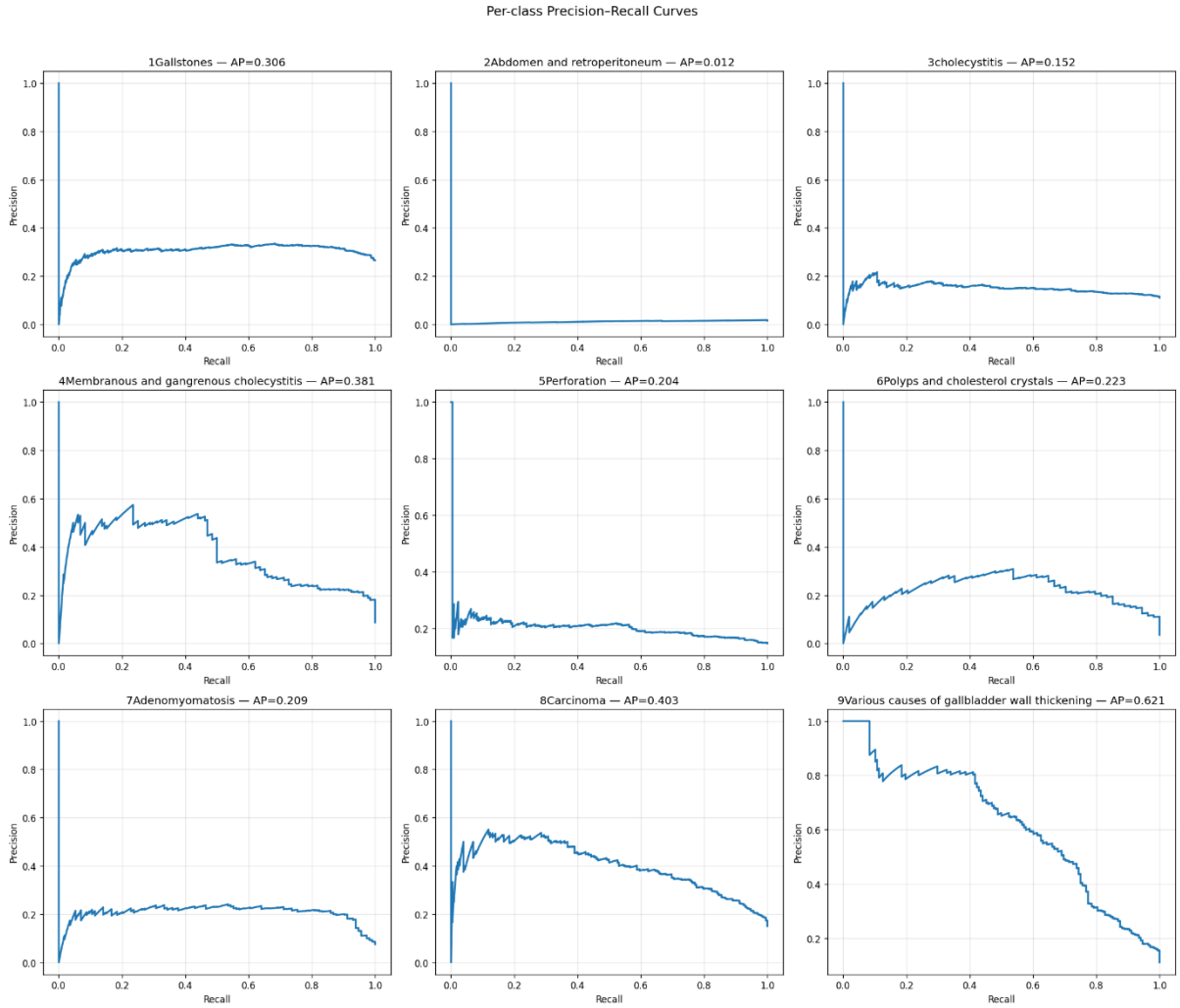


Figure 52 - PR Curves of the Best GhostNet Algorithm.

5.3.3. Results from the GradCAM++ Experiment

In order to understand where the GhostNet model is paying attention when it makes a prediction the Grad-CAM++ heatmaps were utilised. Thus, twelve test images were employed, from which six were correct and six incorrect. They were then overlaid on the ultrasound frames. The result is presented in **Figure 53.** The hot colours like yellow and green in the figure indicate regions with strong positive contribution to the predicted class. Where the cold ones appear they have little contribution to the prediction (Chattopadhyay et al., 2018).

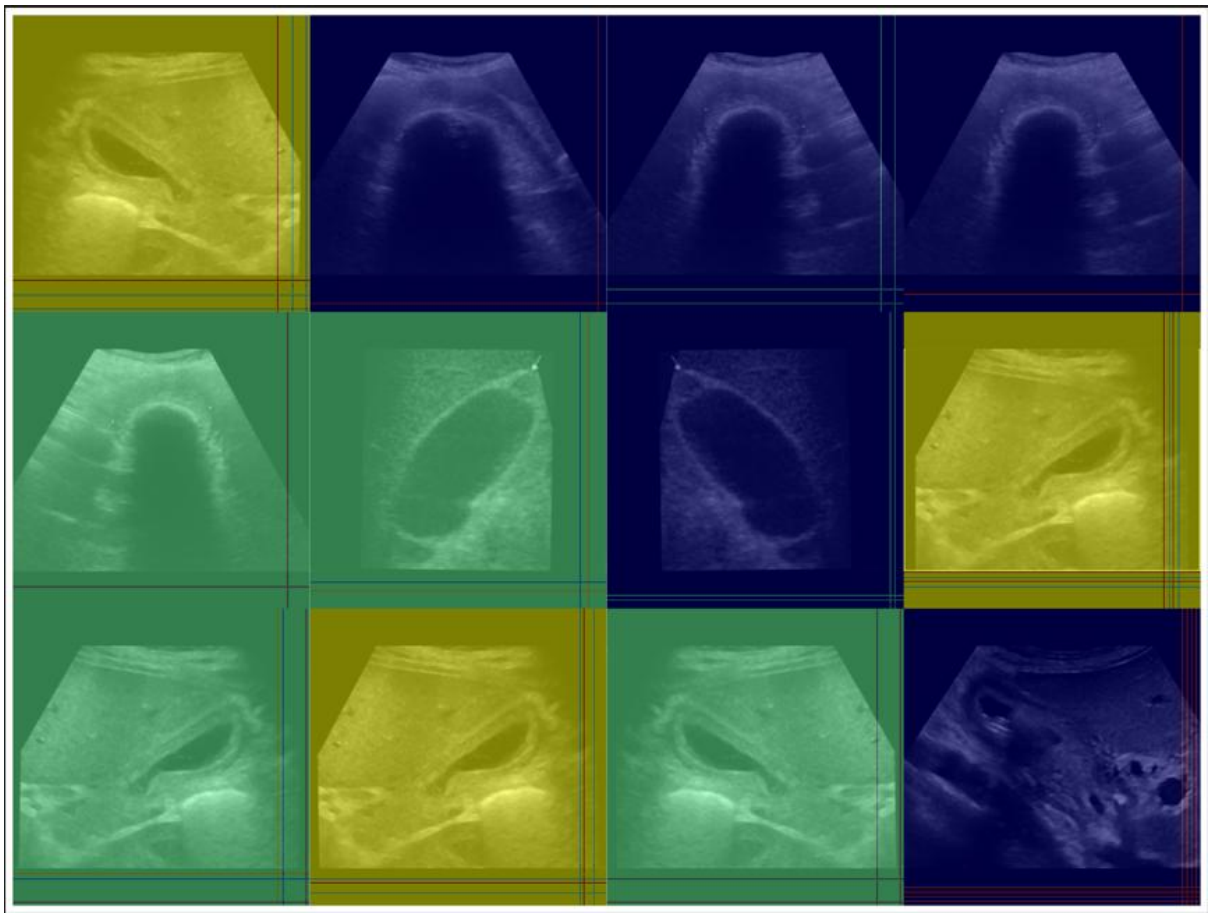


Figure 53- Grad-CAM++ Maps of GhostNet.

Hence, the findings can be summarised as follows:

- **The focus is falling on the gallbladder wall** which is a good behaviour. This is an important area as many target signs for pathologies reside in the region. This suggests that the model has learnt at least some of the pathological cues.
- **The heatmaps light up the dark shadows behind bright spots** – which is highlighting stones in the organ, however other bodily parts also cause US shadows. Thus, the model can mislabel images confusing them for stones (Rumack et.al., 2011).
- Some heat maps light up along the **edges of the sample** and not on the GB.
- On some cases that are wrong the **heatmap spreads inconsistently across the sample** US. Thus, suggesting that the model cannot locate the pathology and is guessing from the texture.

Thus, the focus on the wall area is consistent with what was seen in the higher AP for the classes “Carcinoma” and “Various causes of gallbladder wall thickening”. These precisely the categories that depend on wall pathology cues. Also, the concentration on shadows aligns with the PR curves that exemplified high recall but low precision. Meaning, the model makes a guess when it “spots” a combination of shadow and bright marks, which is usually a mislabelled decision.

6. DISCUSSIONS

6.1. Chapter Overview

This chapter ties together what we built, what actually happened in training and testing, and why. It also explains why the final numbers that were achieved are much lower than the state-of-the-art results reported in most of the papers included in the **2. Literature review**, and why that is actually a good sign for methodological quality rather than a failure of modelling.

6.2. Why Performance Stayed Low?

For all of the three models included in this research, namely - GhostNet-1.0, ResNet-50 and TinyViT, and across all methodological training strategies attempted, the latter showcased where the training loss fell steadily while validation loss plateaued and Macro-F1 on validation was not surpassing the 0.36 range. The Macro-F1 of the best performing model – GhostNet-1.0 trained withing the original methodological strategy, delivered Macro-F1 around 0.27. More precisely it can be said that the models learnt patterns, but they did not generalise well to data from new patients.

Behind these outcomes lay four practical reasons:

- **Generalisation on patient-level splits with no data leakage** is genuinely hard for the algorithms – as discussed in the 1.1. Background Study section, the US appearance varies between patients, scanners and precise US technology utilised. Thus, lightweight models have less capacity to tackle such variability. Hence, there is the need of very strong and consistent cues in order to respond well to new data. All the curves presented in the previous chapter that exemplified high training Macro-F1 and

flat validation Macro-F1, which is classic textbook example of overfitting (Goodfellow et al., 2016).

- **Overlapping between classes –** several classes can be outlined to share similar appearances in the data. Examples are polyps and wall irregularities. Others, on the other hand, are much less frequent in the data – “Abdomen and retroperitoneum”. The precision and recall plots from the previous chapter exemplify this clearly with some classes hitting modest AP (like “Various wall thickening”), while others show near zero results (Saito & Rehmsmeier, 2015).
- **Artefacts in the US samples –** The Grad-CAM++ experiment maps revealed that the model turns its attention to area borders and black padding in some errors (Section 5.3.3.). Hence, some of the network’s resources are spent on cues in the layout and background that are not important and cannot be transferred and generalise well on new data. This explains the gap between the training and validation loss that occurred.

6.3. What Were the Derived Methodological Conclusions?

After having attempted twelve different training setups and tweaking virtually almost every single aspect of the pre-processing and modelling strategy, the only reasonable conclusion was that the patient-level data splits and the disallowance of data leakage between them actually did their job. Deductively, this makes a lot of sense, because in such a manner no patient data from the same patient is allowed to flow between different splits. Meaning, the model is only trained on the data from a number of patients, validated on data of others and tested on third, which is also unseen. In that way it is ensured that the algorithm does not

simply “recognise” similar shots from the same patients that leaked between the splits, but rather it learns the pathological ques effectively. This is the proper and honest methodological approach that ensures there is not bias in the research.

Although, **the results** achieved in the current research are not state-of-the-art, they do follow the aforementioned policy and **are unbiased, fair, and realistic**. Of course, after this assumption was made it was easy to understand why a lot of the reviewed papers in **Chapter 2**. **Literature Review** did not apply strict patient-level data splitting and a data policy that disallows data leakage between splits. In such a manner, and combined with metrics that do not consider false alarms (e.g. Accuracy), the performance results are heavily inflated. To prove this point a Demo training was performed of the best performing model that was trained in the current research (GhostNet-1.0), with a simplified training strategy which totally excludes patient-level data splitting and purposely allows data leakage. The latter is relayed in the following subsection.

6.4. The DEMO GhostNet-1.0: Why “Easy Wins” Are a Red Flag?

To make this point concrete, as already mentioned, a separate DEMO training was conducted with the pre-trained GhostNet-1.0 that does not use patient-level splitting or cross-validation and performs a simple split on image level before the network is trained **(Tables 12 and 13)**.

The code for this DEMO experiment can be accessed through: https://colab.research.google.com/drive/1U0C_LyGEpcp49287NNCaH9JUNUlNkmt4?usp=sharing .

In such a manner, frames from the same patient can appear in both the training, validation and test sets. Because US studies usually contain many images that are nearly identical, the

network would effectively “recognise” patients rather than pathologies, and thus validation metrics will go up (Goodfellow *et al.*, 2016; Brownlee, 2019).

Such behaviour is well known. The leakage across subsets inflates performance and usually fails at deployment or real-world testing. Thus, the DEMO model is not a better architecture with better training design and deliverables, but a definitive attempt to prove why is strict patient-level pipeline necessary for credible and acknowledged results. The huge gap exemplified by the derived training metrics (**Figure 54**) and the original patient-level CV results is evidence that the lower scores are more realistic for real clinical use.

Table 12 - Pre-processing Approach for the DEMO Training of GhostNet-1.0.

Pre-processing Approach DEMO Training of GhostNet-1.0.		
Item	**Value**	**Notes**
Colour space	RGB (3-channel)	Each image is read in grayscale, letterboxed & resized, then converted to RGB for the backbone.
Input size	320 × 320 px	Consistent across train/val/test.
Aspect handling	Letterbox to square	Pads shorter side to keep anatomy proportions before resize.
Resize interpolation	Bilinear (cv2.INTER_LINEAR)	Standard for natural images/US frames.
Normalisation	ImageNet mean/std	A.Normalise(mean=IMAGENET_MEAN, std=IMAGENET_STD) (not per-image z-score).
Median denoise	No	No SRAD/median filtering in DEMO.
Train augs	H-flip (p=0.5); Shift/Scale/Rotate (scale ±0.10, rotate ±10, p=0.8, border=constant); Brightness/Contrast (±0.15, p=0.8)	Conservative, geometry-safe edits.
Val/Test augs	Normalisation only	No flips/rotations applied.

Table 13 - Training and Splitting Approach of DEMO GhostNet-1.0 Training.

Training and Splitting of DEMO GhostNet-1.0 Training.		
Item	**Value**	**Notes**
Backbone	GhostNet-1.0 (timm: ghostnet_100)	Pretrained on ImageNet; in_chans=3; drop_rate=0.20.
Classes	Derived from folder names	
Split strategy	Image-level stratified (no patient grouping)	70% train / 15% val / 15% test via StratifiedShuffleSplit.
Batch size / workers	64 / 4	
Optimiser	AdamW (2 param groups)	Head LR = 3e-4; Backbone LR = 1e-4; weight_decay=1e-4.
Loss	Cross-Entropy + label smoothing (0.05)	Stabilises early training.
LR schedule	Cosine with warm-up (3 epochs)	Implemented via LambdaLR. Total epochs 50.
AMP	Enabled	torch.amp.GradScaler; autocast during fwd pass.
Early stopping	Patience 7, monitor val macro-F1	Stops if no improvement.
Checkpoints	last.pt every epoch, best.pt on F1↑	Saved under /DEMO/checkpoints/; resume from last.pt if present.

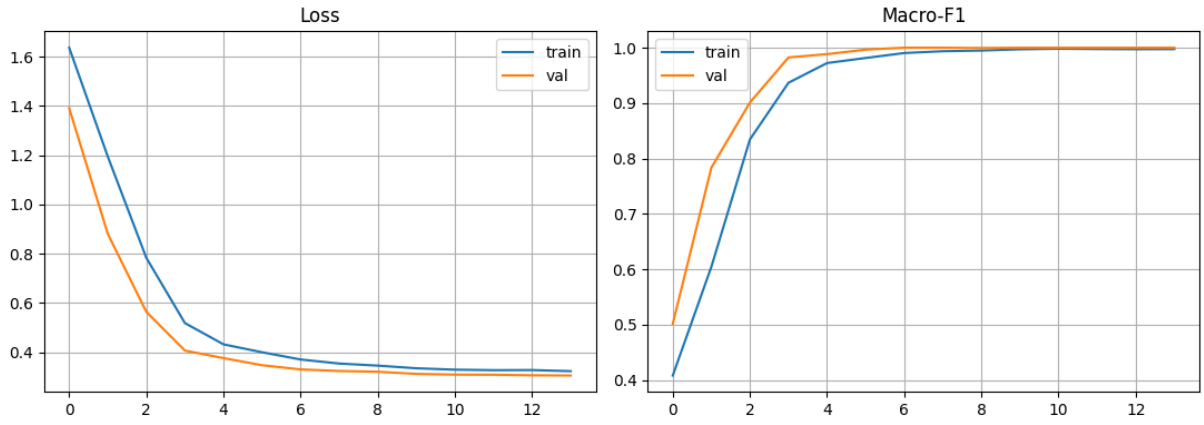


Figure 54 - Training Metrics Graphs from DEMO Training of GhostNet-1.0.

It is clearly visible from **Figure 54** that with this methodological approach the network loss goes to almost zero by epoch 10 and the training and validation Macro-F1 to 1.0 by epoch 7. Which proves the statement that the model is artificially inflating the results by simply recognising the frames of a specific patient.

6.5. How Our Results Compare to Published Works?

As discussed in Chapter 2 – Literature Review, there are two recent multi-class studies on the same dataset that report substantially higher figures:

- **Obaid et al. (2023)** apply denoising and segmentation that is utilising ROI. Then the authors benchmark several CNNs. MobileNet is reported to be the top performer with

accuracies around 98% on a single 80/20 split. Importantly, the paper states that the split was made at the patient level by grouping the images by patient ID to avoid leakage. However, the authors do not report cross-validation or any external validation, so performance is calculated over just one split.

- **Bozdag et al. (2025)** attempt diagnosis through content-based image retrieval and report a mean average precision around 0.94. The evaluation is conducted on the same dataset, but the paper does not describe a patient-level data separation or any external validation. Thus, the performance may be inflated in the presence of frames that can be duplicate or are very similar to each other due to data leakage.

In contrast, the pipeline the current research utilises enforces patient-level splitting for both cross-validation and the held-out test set, alongside early stopping and model selection on Macro-F1. The lower macro-F1 is therefore consistent with generalisation in real world situations that are controlled for leakage. Notably, even Obaid et al. highlight the importance of such grouping to avoid optimistic and biased results.

7. CONCLUSION AND FUTURE WORK

7.1. Conclusion

This research set out to build a lightweight DL solution to diagnose and classify nine GB disease pathologies trough US imagery. This was attempted through a careful pipeline, which enforced data leakage controls and three compact models (baseline – ResNet-50; GhostNet-1.0 and TinyViT as the main competitors). The **patient-level cross-validation** and held-out test split were imposed in conjunction with early stopping on macro-F1 values to assess the training behaviour of the models. The best performer – GhostNet-1.0 was audited through an evaluation on the test set, PR-curves, confusion matrix and Grad-CAM++ maps.

Across all training strategies attempted due to unsatisfactory results, the core picture was consistent – **training curves climbed fast while validation macro-F1 remained modest.** The latter revolved and peaked around 0.30 in the best folds for the original and first change recipes (around 0.35 for GhostNet 0.5 @ fold 3 epoch 34; ResNet-50 at app. 0.37 @ fold 2 epoch 21; TinyViT around 0.37 @ fold 5 epoch 24). These gains were short-lived and often followed by overfitting even though gradual unfreezing of the backbone and discriminative learning rates with mild augmentation and regularisation were applied.

The final and “conservative” recipe (grayscale input, conditional CLAHE with NLM repairs, Gaussian blur, WeightedRandomSampler) **stabilised losses but further decreased validation macro-F1** with GhostNet ranging between 0.20 to 0.23; ResNet-

50 around 0.16 to 0.18; TinyViT at approximately 0.10 to 0.14. Which suggests the classifier became more careful but did not generalise better on the data.

On the held-out test set, the outlined as the best performer GhostNet network hit macro-F1 of 0.272. Performance was uneven across classes with “Various causes of wall thickening” showing up as the most reliable (F1 around 0.57; precision – 0.61; recall 0.54), while “Carcinoma” (F1 at approximately 0.42) and “Polyps and cholesterol crystals” (F1 around 0.36) showed the familiar precision and recall tension pattern; one rare class – “Abdomen and retroperitoneum” collapsed to zero. Specificity was generally high and recall modest, which is consistent with a conservative decision boundary under class imbalance.

The low performance was underline because of three main factors:

- **The planned methodological strictness worked –** patient-level splitting provided a fair ground not allowing data leakage between sets, so the networks had to generalise to new patient data in different phases and not just recognise nearly identical frames. The separate DEMO (split on the image level) proved that and quickly drove training and validation macro-F1 towards 1.0.
- **Class imbalance and visual overlap of pathologies –** minority classes were hard to recover, and several categories share similar appearances under US. Which reduced macro-F1 even when common classes provided reasonable output.
- **Cues and artefacts that were not related to the pathologies –** Grad-CAM++ showed attention on area boundaries and shadows in some errors, which

highlighted a sensitivity towards the layout, background and artefacts rather than robust pathology capturing.

Thus, methodologically the **main contribution** is a **reproducible, justified by leakage control evaluation** with transparent diagnostics. The lower results are a **true and fair estimate of real-world generalisation** rather than a modelling failure. This research also answers the question w**hether a lightweight DL approach (up to 11M trainable parameters) can deliver state-of-the-art results,** and the answer can be deducted as – "**No**". At least not under a fair and conservative data leakage policy, without artificially inflating the results by allowing data to slip between splits.

7.2. Future Work

However, there are still techniques that were not utilised in this research due to time and resource limitations, and that are capable of uplifting the performance delivered by the algorithms. Such are:

- **The utilisation of ROI** – a model can be further added to the pipeline to define a region of interest and segregate the background from the actual GB organ, thus focusing the attention of the main model on the actual organ and the present pathology.
- **More multi-centre data is gathered –** the addition of more data that can be split on patient-level will significantly improve the learning abilities of the model.
- **The organisation of external validation experiments** with multiple experts from various hospital centres to confirm the output of the Network.

8. BIBLIOGRAPHY

Arlot, S. & Celisse, A. (2009) A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79. [Accessed: 15 August 2025].

Atlas University Hospital (2022) *Gallbladder Diseases and Symptoms*. Available at: https://atlasuniversitesihastanesi.com/en/gallbladder-diseases-and-symptoms/ [Accessed: 23 June 2025].

Awad, M. & Khanna, R. (2015) Efficient learning machines: Theories, concepts, and applications for engineers and system designers. In: Apress Media LLC *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. [Online]. Apress Media LLC. Available at: doi:10.1007/978-1-4302-5990-9 [Accessed: 10 July 2025].

Basu, S. et al. (2022) Surpassing the Human Accuracy: Detecting Gallbladder Cancer from USG Images with Curriculum Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June, 20854–20864. IEEE Computer Society. [Accessed: 10 August 2025].

Basu, S. et al. (2023) RadFormer: Transformers with global–local attention for interpretable and accurate Gallbladder Cancer detection. *Medical Image Analysis*, 83. Elsevier B.V. [Accessed: 10 August 2025].

Bergstra, J. et al. (2012) Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13, 281–305. JMLR.orgPUB6573. [Accessed: 15 August 2025].

Bozdag, A. et al. (2025) Detection of Gallbladder Disease Types Using a Feature Engineering-Based Developed CBIR System. *Diagnostics 2025, Vol. 15, Page 552*, 15(5), 552. Multidisciplinary Digital Publishing Institute. [Accessed: 27 May 2025].

Brodersen, K.H. et al. (2010) The balanced accuracy and its posterior distribution. *Proceedings - International Conference on Pattern Recognition*, 3121–3124. [Accessed: 18 August 2025].

Brownlee, J. (2019) *Better deep learning*. v1.81. [Online]. Available at: https://machinelearningmastery.com/better-deep-learning/ [Accessed: 23 November 2023].

Brownlee, J. (2020) How to Fix k-Fold Cross-Validation for Imbalanced Classification - MachineLearningMastery.com. In: *Imbalanced Classification with Python*. [Online]. Available at: https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/ [Accessed: 15 August 2025].

Buda, M. et al. (2018) A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. Elsevier Ltd. [Accessed: 24 August 2025].

C., C. & L.C., T. (2010) On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *The Journal of Machine Learning Research*, 11, 2079–2107. JMLR.orgPUB6573. [Accessed: 15 August 2025].

Chattopadhyay, A. et al. (2018) Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. *Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, 2018-January, 839–847. Institute of Electrical and Electronics Engineers Inc. [Accessed: 15 August 2025].

Chen, T. et al. (2020) Computer-aided diagnosis of gallbladder polyps based on high resolution ultrasonography. *Computer Methods and Programs in Biomedicine*, 185. Elsevier Ireland Ltd. [Accessed: 10 August 2025].

Choi, J.H. et al. (2023) Analysis of ultrasonographic images using a deep learning-based model as ancillary diagnostic tool for diagnosing gallbladder polyps. *Digestive and Liver Disease*, 55(12), 1705–1711. W.B. Saunders. [Accessed: 29 July 2025].

Choi, Y.S. et al. (2016) Prevalence and risk factors of gallbladder polypoid lesions in a healthy population. *Yonsei Medical Journal*, 57(6), 1370–1375. Yonsei University College of Medicine. [Accessed: 28 July 2025].

Cobo, M. et al. (2023) Enhancing radiomics and Deep Learning systems through the standardization of medical imaging workflows. *Scientific Data*, 10(1), 1–7. Nature Research. [Accessed: 13 August 2025].

Cubuk, E.D. et al. (2018) AutoAugment: Learning Augmentation Policies from Data. *Cvpr 2019*, (Section 3), 113–123. [Accessed: 13 August 2025].

Davis, J. and Goadrich, M., 2006, June. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). [Accessed: 26 August 2025]

Dilek, O.N. et al. (2019) Diagnosis and Treatment of Gallbladder Polyps: Current Perspectives. *Euroasian Journal of Hepato-Gastroenterology*, 9(1), 40. Jaypee Brothers Medical Publishing. [Accessed: 28 July 2025].

Elshennawy, N.M. & Ibrahim, D.M. (2020) Deep-Pneumonia Framework Using Deep Learning Models Based on Chest X-Ray Images. *Diagnostics*, 10(9), 649. Multidisciplinary Digital Publishing Institute (MDPI). [Accessed: 13 August 2025].

GeeksforGeeks (2025) *Layers in Artificial Neural Networks (ANN)* . Available at: https://www.geeksforgeeks.org/deep-learning/layers-in-artificial-neural-networks-ann/ [Accessed: 10 July 2025].

Goyal, P. et al. (2017) *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*. [Accessed: 15 August 2025].

Gupta, P. et al. (2024) Deep-learning enabled ultrasound based detection of gallbladder cancer in northern India: a prospective diagnostic study. *The Lancet Regional Health - Southeast Asia*, 24. Elsevier Ltd. [Accessed: 10 August 2025].

Han, K. et al. (2019) GhostNet: More Features from Cheap Operations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1577–1586. IEEE Computer Society. [Accessed: 13 August 2025].

Hand, D.J. & Till, R.J. (2001) A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45(2), 171–186. Springer. [Accessed: 15 August 2025].

He, H. & Garcia, E.A. (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. [Accessed: 21 August 2025].

He, K. et al. (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 770–778. IEEE Computer Society. [Accessed: 18 August 2025].

Housset, C. et al. (2016) Functions of the Gallbladder. *Comprehensive Physiology*, 6(3), 1549–1577. John Wiley and Sons Inc. [Accessed: 23 June 2025].

Howard, A. et al. (2019) Searching for MobileNetV3. *Proceedings of the IEEE International Conference on Computer Vision*, 1314–1324. Institute of Electrical and Electronics Engineers Inc. [Accessed: 13 August 2025].

Ian Goodfellow et al. (2016) *Deep Learning*. [Online]. The MIT Press. Available at: http://alvarestech.com/temp/deep/Deep%20Learning%20by%20Ian%20Goodfellow,%20Yoshua%20Bengio,%20Aaron%20Courville%20(z-lib.org).pdf [Accessed: 10 July 2025].

Inui, K. et al. (2011) Diagnosis of gallbladder tumors. *Internal Medicine*, 50(11), 1133–1136. Intern Med. [Accessed: 28 July 2025].

Ioffe, S. & Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *32nd International Conference on Machine Learning, ICML 2015*, 1, 448–456. International Machine Learning Society (IMLS). [Accessed: 24 August 2025].

Japkowicz, N. & Stephen, S. (2002) The class imbalance problem: A systematic study. *Intelligent Data Analysis*. IOS PressPUB827Amsterdam, The Netherlands, The Netherlands. [Accessed: 21 August 2025].

Jeong, Y. et al. (2020) Deep learning-based decision support system for the diagnosis of neoplastic gallbladder polyps on ultrasonography: Preliminary results. *Scientific Reports*, 10(1), 7700. Nature Research. [Accessed: 29 July 2025].

Kaufman, S. et al. (2012) Leakage in data mining. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4). ACMPUB27New York, NY, USA. [Accessed: 15 August 2025].

Kim, T. et al. (2021) Gallbladder Polyp Classification in Ultrasound Images Using an Ensemble Convolutional Neural Network Model. *Journal of Clinical Medicine*, 10(16), 3585. MDPI. [Accessed: 29 July 2025].

Kingma, D.P. & Ba, J.L. (2014) Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. [Accessed: 15 August 2025].

Kirkpatrick, J. et al. (2017) Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13), 3521–3526. National Academy of Sciences. [Accessed: 22 August 2025].

Kohavi, R. (1995) *A study of cross-validation and bootstrap for accuracy estimation and model selection | Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*. In: 1995. Available at: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf [Accessed: 18 August 2025].

Kornblith, S. et al. (2018) Do Better ImageNet Models Transfer Better? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June, 2656–2666. IEEE Computer Society. [Accessed: 18 August 2025].

Kundu, R. et al. (2021) Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLOS ONE*, 16(9), e0256630. Public Library of Science. [Accessed: 13 August 2025].

Lam, R. et al. (2021) Gallbladder Disorders: A Comprehensive Review. *Disease-a-Month*, 67(7). Mosby Inc. [Accessed: 28 July 2025].

Li, Q. et al. (2023a) *A Bayesian network model to predict neoplastic risk for patients with gallbladder polyps larger than 10 mm based on preoperative ultrasound features*. 37, 5453–5463. [Accessed: 10 August 2025].

Li, Q. et al. (2023b) A Bayesian network prediction model for gallbladder polyps with malignant potential based on preoperative ultrasound. *Surgical Endoscopy*, 37(1), 518–527. Springer. [Accessed: 8 August 2025].

Li, X. et al. (2018) Explicit Inductive Bias for Transfer Learning with Convolutional Networks. *35th International Conference on Machine Learning, ICML 2018*, 6, 4408–4419. International Machine Learning Society (IMLS). [Accessed: 24 August 2025].

Lian, J. et al. (2017) Automatic gallbladder and gallstone regions segmentation in ultrasound image. *International Journal of Computer Assisted Radiology and Surgery*, 12(4), 553–568. Springer Verlag. [Accessed: 10 August 2025].

Lin, H. et al. (2024) Machine learning and human-machine trust in healthcare: A systematic survey. *CAAI Transactions on Intelligence Technology*, 9(2), 286–302. John Wiley and Sons Inc. [Accessed: 10 July 2025].

Lin, T.Y. et al. (2017) Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2), 318–327. IEEE Computer Society. [Accessed: 15 August 2025].

Lin, W.R. et al. (2008) Prevalence of and risk factors for gallbladder polyps detected by ultrasonography among healthy Chinese: Analysis of 34 669 cases. *Journal of Gastroenterology and Hepatology (Australia)*, 23(6), 965–969. Blackwell Publishing. [Accessed: 28 July 2025].

Loshchilov, I. & Hutter, F. (2016) SGDR: Stochastic Gradient Descent with Warm Restarts. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track*

Proceedings. International Conference on Learning Representations, ICLR. [Accessed: 15 August 2025].

Loshchilov, I. & Hutter, F. (2017) Decoupled Weight Decay Regularization. *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR. [Accessed: 15 August 2025].

Luo, X. et al. (2024) Efficient Deep Learning Infrastructures for Embedded Computing Systems: A Comprehensive Survey and Future Envision. *Journal of the ACM*, 1. [Accessed: 13 August 2025].

Ma, N. et al. (2018) ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11218 LNCS, 122–138. Springer Verlag. [Accessed: 13 August 2025].

Martin, H. (2019) *Deep learning explained* . Available at: https://www.proquest.com/docview/2229801791?accountid=9653&parentSessionId=nH0QstWjCAvz8g5wiOUItNfLqKfsEQR0QpNFBPp8fTA%3D&pq-origsite=summon&sourcetype=Trade%20Journals [Accessed: 10 July 2025].

Moinuddin, M. et al. (2022) Medical ultrasound image speckle reduction and resolution enhancement using texture compensated multi-resolution convolution neural network. *Frontiers in Physiology*, 13, 961571. Frontiers Media S.A. [Accessed: 13 August 2025].

Narang, S. et al. (2017) Mixed Precision Training. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. [Accessed: 18 August 2025].

Nia, S.N. & Shih, F.Y. (2024) Medical X-Ray Image Enhancement Using Global Contrast-Limited Adaptive Histogram Equalization. *International Journal of Pattern Recognition and Artificial Intelligence*, 38(12). World Scientific. [Accessed: 24 August 2025].

Obaid, A.M. et al. (2023) Detection of Gallbladder Disease Types Using Deep Learning: An Informative Medical Method. *Diagnostics 2023, Vol. 13, Page 1744*, 13(10), 1744. Multidisciplinary Digital Publishing Institute. [Accessed: 27 May 2025].

Pan, S.J. & Yang, Q. (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345–1359. [Accessed: 15 August 2025].

Pavlidis, E.T. et al. (2023) Contemporary diagnosis and management of gallbladder polyps. *Gastroenterology and Functional Medicine*, 1(1). Scholar Media Publishing. [Accessed: 28 July 2025].

Pertuz, S. et al. (2013) Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5), 1415–1432. Pergamon. [Accessed: 13 August 2025].

Powers, D.M.W. & Ailab (2020) *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*. [Accessed: 15 August 2025].

Prechelt, L. (1998) *Early Stopping - But When?* 55–69. [Accessed: 15 August 2025].

Rajpurkar, P. et al. (2017) *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. [Accessed: 13 August 2025].

Reinhold, J.C. et al. (2019) Evaluating the Impact of Intensity Normalization on MR Image Synthesis. *Proceedings of SPIE--the International Society for Optical Engineering*, 10949, 109493H. SPIE-Intl Soc Optical Eng. [Accessed: 13 August 2025].

Roberts, D.R. et al. (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. Blackwell Publishing Ltd. [Accessed: 22 August 2025].

Roberts, M. et al. (2021) Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3), 199–217. Nature Research. [Accessed: 21 August 2025].

Rumack, C.M. and Levine, D. eds., 2023. Diagnostic ultrasound E-book. Elsevier Health Sciences. [Accessed: 26 August 2025].

Russell, S.J.. et al. (2022) *Artificial intelligence : a modern approach*. 1166. Pearson.

Saito, T. & Rehmsmeier, M. (2015) The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3), e0118432. Public Library of Science. [Accessed: 15 August 2025].

Sevakula, R.K. et al. (2020) State-of-the-Art Machine Learning Techniques Aiming to Improve Patient Outcomes Pertaining to the Cardiovascular System. *Journal of the American Heart Association*, 9(4). American Heart Association Inc. [Accessed: 9 July 2025].

Shorten, C. & Khoshgoftaar, T.M. (2019) A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1), 1–48. SpringerOpen. [Accessed: 13 August 2025].

Singh, J. et al. (2023) Batch-balanced focal loss: a hybrid solution to class imbalance in deep learning. *Journal of Medical Imaging*, 10(5), 051809. SPIE-Intl Soc Optical Eng. [Accessed: 15 August 2025].

Singh, P. et al. (2021) Diagnosing of disease using machine learning. *Machine Learning and the Internet of Medical Things in Healthcare*, 89–111. Academic Press. [Accessed: 15 August 2025].

Smith, L.N. (2015) Cyclical Learning Rates for Training Neural Networks. *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, 464–472. Institute of Electrical and Electronics Engineers Inc. [Accessed: 15 August 2025].

Szegedy, C. et al. (2015) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, 2818–2826. IEEE Computer Society. [Accessed: 24 August 2025].

Takahashi, K. et al. (2024) Recent Advances in Endoscopic Ultrasound for Gallbladder Disease Diagnosis. *Diagnostics 2024, Vol. 14, Page 374*, 14(4), 374. Multidisciplinary Digital Publishing Institute. [Accessed: 9 July 2025].

Tejani, A.S. et al. (2024) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence*, 6(4). Radiological Society of North America Inc. [Accessed: 15 August 2025].

ThinkAutonomous (2024) *19 Machine Learning Types you need to know (Advanced Mindmap)*. Available at: https://www.thinkautonomous.ai/blog/types-of-learning-in-machine-learning/ [Accessed: 10 July 2025].

Turki, A. et al. (2024) *Gallblader Diseases Dataset*. 1. Mendeley Data. [Accessed: 28 May 2025].

Varma, S. & Simon, R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1), 1–8. BioMed Central. [Accessed: 22 August 2025].

Vasilev, Ivan. (2019) *Python deep learning : exploring deep learning techniques and neural network architectures with PyTorch, Keras, and TensorFlow*. Packt Publishing.

Wang, L.F. et al. (2023) Risk stratification of gallbladder masses by machine learning-based ultrasound radiomics models: a prospective and multi-institutional study. *European Radiology*, 33(12), 8899–8911. Springer Science and Business Media Deutschland GmbH. [Accessed: 10 August 2025].

Wang, X. et al. (2024) Global Epidemiology of Gallstones in the 21st Century: A Systematic Review and Meta-Analysis. *Clinical Gastroenterology and Hepatology*, 22(8), 1586–1595. W.B. Saunders. [Accessed: 9 June 2025].

Wang, X. et al. (2025) Current status of artificial intelligence analysis for the diagnosis of gallbladder diseases using ultrasonography: a scoping review. *Translational Gastroenterology and Hepatology*, 10, 12. AME Publishing Company. [Accessed: 27 May 2025].

Wiles, R. et al. (2014) Growth rate and malignant potential of small gallbladder polyps - Systematic review of evidence. *Surgeon*, 12(4), 221–226. Elsevier Ltd. [Accessed: 28 July 2025].

Winslow, T. (2010) *Gallbladder Anatomy: Image Details* . Available at: https://visualsonline.cancer.gov/details.cfm?imageid=9078 [Accessed: 23 June 2025].

Wu, K. et al. (2022) TinyViT: Fast Pretraining Distillation for Small Vision Transformers. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13681 LNCS, 68–85. Springer Science and Business Media Deutschland GmbH. [Accessed: 13 August 2025].

Xu, A. et al. (2017) Gallbladder Polypoid-Lesions: What Are They and How Should They be Treated? A Single-Center Experience Based on 1446 Cholecystectomy Patients. *Journal of Gastrointestinal Surgery*, 21(11), 1804–1812. Springer New York LLC. [Accessed: 28 July 2025].

Xu, A. & Hu, H. (2017) The gallbladder polypoid-lesions conundrum: moving forward with controversy by looking back. *Expert Review of Gastroenterology and Hepatology*, 11(11), 1071–1080. Taylor and Francis Ltd. [Accessed: 28 July 2025].

Xue, F. et al. (2011) Image denoising based on improved non-local algorithm. *Communications in Computer and Information Science*, 152 CCIS(PART 1), 283–289. [Accessed: 24 August 2025].

Xue, L. et al. (2021) Segnet Network Algorithm-Based Ultrasound Images in the Diagnosis of Gallbladder Stones Complicated with Gallbladder Carcinoma and the Relationship between P16 Expression with Gallbladder Carcinoma. *Journal of Healthcare Engineering*, 2021. Hindawi Limited. [Accessed: 10 August 2025].

Yang, J. et al. (2023) MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*, 10(1), 41. Nature Research. [Accessed: 13 August 2025].

Yosinski, J. et al. (2014) How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 4(January), 3320–3328. Neural information processing systems foundation. [Accessed: 18 August 2025].

Yu, Y. & Acton, S.T. (2002) Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11(11), 1260–1270. [Accessed: 13 August 2025].

Yuan, H. xia et al. (2023) Differential diagnosis of gallbladder neoplastic polyps and cholesterol polyps with radiomics of dual modal ultrasound: a pilot study. *BMC Medical Imaging*, 23(1). BioMed Central Ltd. [Accessed: 8 August 2025].

Zech, J.R. et al. (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11). Public Library of Science. [Accessed: 22 August 2025].

Zhang, H. et al. (2017) mixup: Beyond Empirical Risk Minimization. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. [Accessed: 18 August 2025].

Zhou, W. et al. (2021) Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. *Nature Communications*, 12(1), 1259. Nature Research. [Accessed: 11 August 2025].

9. WORD COUNT

21,119 words.

10. APPENDICES

10.1. APPENDIX 1 – Gantt Chart and Research Schedule

Table 14 - Research Schedule.

Task	Start	End	Duration in days
Proposal & planning	02/06/2025	09/06/2025	7
Introduction, Background and Problem Statement	10/06/2025	17/06/2025	7
Literature review & Informing methodology	18/06/2025	04/07/2025	16
Methodology Planning and Chapter Writing	05/07/2025	22/07/2025	17
EDA & data audit	23/07/2025	24/07/2025	1
Preprocessing & QC design	24/07/2025	26/07/2025	2
Preprocessing implementation & caching	26/07/2025	27/07/2025	1
Data splits & leakage controls	27/07/2025	29/07/2025	2
Baseline training (Original methodology)	29/07/2025	03/08/2025	5
Training - First changes	03/08/2025	06/08/2025	3
Training - Second changes	06/08/2025	09/08/2025	3
Training - Final changes (grayscale + repairs)	09/08/2025	12/08/2025	3
Experiments: Grad-CAM++, Confusion/PR, Held-out test	12/08/2025	16/08/2025	4
DEMO (image-level split) & contrast study	16/08/2025	20/08/2025	4
Implementation & Experiments Chapter Writing	20/08/2025	23/08/2025	3
Discussion Chapter Writing	23/08/2025	24/08/2025	1
Conclusion & Future Work Chapter Writing	24/08/2025	25/08/2025	1
Final edit, formatting & submission	26/08/2025	08/09/2025	13

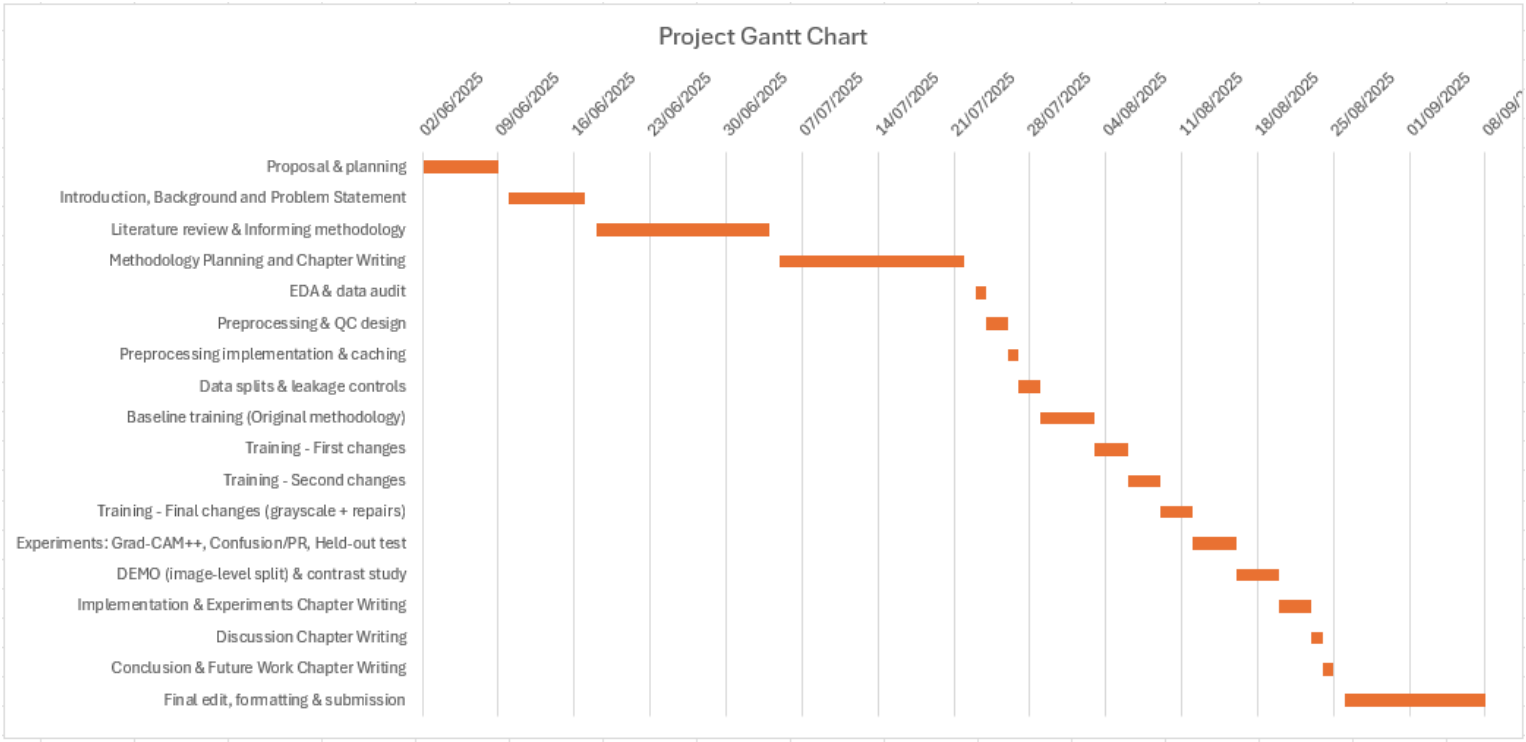


Figure 55 - Gantt Chart.

10.2. APPENDIX 2 – Access Links to Colab Notebooks and Dataset.

The GBD Datataset: https://data.mendeley.com/datasets/r6h24d2d3y/1

The Main Project Colab Notebook:

https://colab.research.google.com/drive/1gDyn_WXGn3repUKswgfRuNbnHzNo2Alf?usp=sharing

The DEMO Training Colab Notebook:

https://colab.research.google.com/drive/1U0C_LyGEpcp49287NNCaH9JUNUlNkmt4?usp=sharing