

博士論文

限られたデータを用いた機械学習の適用法と
その有効性評価

(Development and Evaluation of Machine Learning Applications
under Limited Data Conditions)

2026 年 3 月

中田和真

山口大学大学院 創成科学研究科

目次

1	序論	1
1.1	機械学習	1
1.1.1	機械学習	1
1.1.2	機械学習の社会への応用	3
1.2	超音波画像を用いた肝硬変診断	4
1.2.1	肝硬変	4
1.2.2	肝硬変診断方法	6
1.2.3	超音波画像	6
1.2.4	超音波画像を用いた肝硬変診断	9
1.2.5	超音波画像を用いた医師の目視診断の精度	11
1.3	コンクリート構造物のひび割れ調査	13
1.3.1	コンクリート構造物	13
1.3.2	コンクリートのひび割れ	13
1.3.3	ひび割れ調査	14
1.4	実験で使用する評価指標	14
1.5	本論文の提案および貢献	16
1.5.1	医療画像分野（ラベル情報が限られる状況での機械学習）	16
1.5.2	インフラ分野ひび割れ画像（特徴量が限られる状況での機械学習）	18
1.5.3	本論文の構成	19
2	位置情報を用いた信頼性評価による肝硬変検出の高精度化	20
2.1	はじめに	20
2.2	先行研究	22
2.2.1	手動で設定された ROI	22
2.2.2	自動で設定された ROI	23
2.3	提案手法: 信頼性評価法 (REM)	23
2.3.1	提案システム	24
2.3.2	画像前処理	25

2.3.3	自動 ROI 設定	25
2.3.4	Gabor 特徴量	25
2.3.5	Quadratic Discriminant Analysis	25
2.3.6	Product Rule	26
2.3.7	信頼性評価法	26
2.3.8	ラベル付与方法	27
2.4	実験	28
2.4.1	データセットとパラメーター	28
2.4.2	評価指標	29
2.5	結果と考察	29
2.5.1	予備実験	29
2.5.2	実践的実験	43
2.6	まとめ	46
3	位置情報を用いた信頼性評価と Multiple Instance Learning の統合による CNN モデルの肝硬変検出の精度向上	52
3.1	はじめに	52
3.2	先行研究	55
3.3	提案手法: MIL による学習と REM による信頼性評価の統合	57
3.3.1	CNN モデル構造	58
3.3.2	Multiple Instance Learning(MIL)	58
3.3.3	信頼性評価方法 (Reliability Evaluation Method, REM)	59
3.3.4	複数 ROI を用いた分類方法	59
3.4	予備実験	59
3.4.1	データセットとパラメーター	60
3.4.2	実装の詳細	60
3.4.3	評価指標	60
3.4.4	結果と考察	61
3.5	実験	62
3.5.1	実装の詳細	62
3.5.2	評価指標	63
3.5.3	比較する手法	63
3.5.4	結果と考察	64
3.5.5	制約	67
3.6	まとめ	68

4 複数解像度画像における共変量シフト適応を用いたコンクリートひび割れ抽出の高精度化	69
4.1 はじめに	69
4.2 先行研究	71
4.2.1 深層学習によるひび割れ抽出	71
4.2.2 ひび割れ抽出におけるドメイン適応	71
4.2.3 共変量シフト適応のための重要度重みづけ	72
4.3 提案手法: 1 ステップ複数解像度適応 (MRA)	74
4.3.1 モデル構造	74
4.3.2 学習フレームワーク	75
4.4 実験	75
4.4.1 データセット	76
4.4.2 複数解像度作成方法	76
4.4.3 評価指標	77
4.5 結果と考察	77
4.5.1 IoU 結果	77
4.5.2 学習データ内の低解像度画像が 5% の場合	78
4.5.3 学習データ内の低解像度画像が 1% の場合	78
4.5.4 重要度推定の結果	78
4.5.5 CAM の可視化結果	79
4.5.6 本章の課題	79
4.6 まとめ	81
5 結論	83
5.1 総括	83
5.2 今後の展望	85
付録	88
第 2 章, 第 3 章 付録 超音波画像データセット	88
第 4 章 付録 共変量シフトの例	94
謝辞	95
本論文に関連する発表論文	97
参考文献	98

目次

1.1	肝臓の正常と肝硬変の比較.	5
1.2	フィブロスキャンによる肝臓の硬さと肝硬変ステージ F0 ~ F4 の 対応 [13].	5
1.3	超音波画像による診断のイメージ.	6
1.4	肝生検による診断.	7
1.5	肝硬変患者の超音波画像.	8
1.6	健常者の超音波画像.	8
1.7	トンネルと舗装路面に発生したひび割れの例.	13
2.1	肝硬変疾患クラスの被験者に対する予備実験での分類結果.	22
2.2	提案システムのフロー.	24
2.3	ROI 信頼性評価プロセスフロー. (a) 周辺 ROI/ 領域の定義. (b) 中心 ROI と周辺 ROI の比較.	26
2.4	予備実験と実践的実験のフロー.	28
2.5	信頼性評価法を用いた予備実験における, テストデータ中の肝硬 変患者の分類結果例.	30
3.1	使用する分類器による分類境界の違い.	53
3.2	ImageNet を用いた学習で獲得された AlexNet の特徴量. [53] より 引用, 加筆.	54
3.3	MIL による学習フロー.	57
3.4	REM を用いた信頼性評価フロー.	59
4.1	(a) 共変量シフト適応を用いた場合 (オレンジ) と用いない場合 (青) の, 線形回帰モデルの学習結果を示す. (b) 学習データとテス トデータの入力分布および密度比を示す.	72
4.2	本章で実験に使用した画像とひび割れ画素数の条件により棄却し た画像, およびそれらに対応するマスク画像の例.	76
4.3	ダウンスケーリングとアップスケーリングにより作成された低解 像度画像および, 対応するマスク画像の例.	77

4.4	各圧縮比率 1/1; 1/10; 1/20 における元画像, 入力画像, U-Net 予測出力画像, マスク画像, 重要度推定モデルの CAM 可視化画像の例.	80
1	超音波画像の操作野 (観察領域) の大きさ.	88
2	Gabor フィルタ 25 次元の実部.	91
3	Gabor フィルタ 25 次元の虚部.	91
4	(a) 共変量シフト適応を用いた場合 (オレンジ) と用いない場合 (青) の, 線形回帰モデルの学習結果を示す. (b) 学習データとテストデータの入力分布および密度比を示す.	94

表目次

2.1	予備実験の結果.	31
2.1	予備実験の結果 (続き).	32
2.1	予備実験の結果 (続き).	33
2.1	予備実験の結果 (続き).	34
2.2	提案 (iv): REM を学習時とテスト時に適用した際の Accuracy. . .	35
2.3	提案 (iv): REM を学習時とテスト時に適用した際の Sensitivity. .	36
2.4	提案 (iv): REM を学習時とテスト時に適用した際の Specificity. .	37
2.5	ROI の枚数比較. REM は各 ROI のラベルでなく, 各 ROI の分類 されたクラスに応じて, 周辺領域のパラメーターを適用する. . .	38
2.6	分類性能の比較.	46
2.7	$8 \times 8 = 64$ 通りにパラメーターの範囲を設定した場合.	47
2.7	$8 \times 8 = 64$ 通りにパラメーターの範囲を設定した場合 (続き). . .	48
2.8	$4 \times 4 = 16$ 通りにパラメーターの範囲を限定した場合.	49
2.9	学習データと検証データに REM を適用した際の最適なパラメー ターの選択.	50
2.10	学習データと検証データに REM を適用した際の各パラメーター の F 値.	51
3.1	AUC の比較.	61
3.2	比較手法の実験結果.	63
3.3	クラス別 Drop rate を用いた際のテストデータに対する精度. . .	65
3.3	クラス別 Drop rate を用いた際のテストデータに対する精度 (続 き).	66
3.4	実験結果一覧.	66
4.1	複数解像度画像に対するひび割れ抽出 IoU 精度の比較. 学習・テ ストデータにおける各圧縮比率 ($1/1$, $1/10$, $1/20$) の構成割合をパ ーセンテージで示し, 各設定においてスコアが高かった値を太字 で示す.	78

4.2 提案手法における各解像度に対する重要度推定結果 (Estimated importance). 各構成割合における 1 ステップ共変量シフト適応で算出された重み係数を示す.	79
--	----

論文の要旨

GPUをはじめとする計算資源およびスマートフォンなどの IoT デバイスの普及により大量のデータを取得・分析する取り組みが加速している。特に深層学習を含む機械学習手法はより多くのデータを活用することで精度の良い予測を可能にする。一方で、ラベルが付与されたデータやテスト環境でのデータは収集コストが高いという課題がある。そこで、比較的収集コストが抑えられる、弱教師と呼ばれるあいまいなラベルや、ラベルなしデータを機械学習モデルの学習に活用することで、データ収集コストの抑制と信頼性の高い機械学習手法の実現が注目されている。本研究では、医療画像とひび割れ画像を研究対象とし、データの収集コスト抑制と高精度な予測の実現を通じて、限られたデータしか活用できない状況においても医療現場とインフラ設備外観検査の支援を目的とした。

「第 1 章 序論」では、これまでの機械学習について一般的な手法と実社会での活用についてまとめるとともに、手法だけでなく活用するデータの重要性について整理を行った。そのうえで、本研究の対象である医療分野とインフラ分野に対して、本研究の貢献と提案を整理した。

「第 2 章 位置情報を用いた信頼性評価による肝硬変検出の高精度化」では、超音波画像を用いた肝硬変検出について、疾患部位のマスクが存在せず、画像単位でのラベルのみ与えられている弱教師あり学習の問題設定として扱う。小領域ごとのパッチ画像を周辺領域のパッチ画像と比較し信頼性評価 (Reliability Evaluation Method, REM) を行うことで、画像内での疾患および健常な部位について選定を行い、機械学習モデルの学習およびテスト時の肝硬変検出の高精度化を検討した。従来手法ではラスタスキャンにより自動生成されたパッチ画像をすべて使用しており、超音波ノイズや影・血管を含む肝硬変検出に不要な領域がモデルの学習とテスト時の分類に影響を与えていた。位置情報を用いた信頼性評価によって信頼できるパッチ画像に対象を絞ったモデルの学習とテスト時の分類を行うことが、被験者単位での肝硬変検出性能の向上に有効であると確認した。

「第 3 章 位置情報を用いた信頼性評価と Multiple Instance Learning の統合による CNN モデルの肝硬変検出の精度向上」では、第 2 章で有効性を確認した信頼

性評価法を深層学習に適用し、畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) による画像特徴量抽出の自動化によって人の経験に依存しない特徴量と分類器の設計を検討した。第 2 章において使用した Gabor 特徴と二次分類器はこれまでの経験から選択されている。一方で、そのような人手による疾患検出システムの設計では、検出精度が設計者に属人的になってしまう課題がある。また、CNN を用いた肝硬変検出では、画像特徴に基づく信頼性評価方法として Multiple Instance Learning (MIL) を活用した学習法が提案されていた。従来法では、医師の指導に基づいて手動で Region of Interest (ROI) を設定しており、設定された ROI の品質に肝硬変検出精度が依存するため、属人性が高い。本章では、自動設定された ROI に対して画像特徴抽出と分類器として CNN を使用し、MIL フレームワークに REM の統合を行った。これにより、自動設定された ROI からでも人手に依存せず学習とテスト時の分類に有用な ROI の選択を可能とし、分類性能の向上を検討した。

「第 4 章 複数解像度画像における共変量シフト適応を用いたコンクリートひび割れ抽出の高精度化」では、コンクリート構造物の表面に発生するひび割れのセグメンテーションを対象とし、深層学習モデルの訓練とテスト時の環境で異なる解像度の画像を活用した場合でもひび割れ抽出の性能低下を抑制する手法を検討した。ひび割れ抽出はインフラ構造物の外観検査自動化に活用され、トンネル、橋梁、といった面積の大きい領域を対象とする。現場での画像取得時にはドローンでの撮影により対象物とカメラの位置の距離が一定でない場合や、コスト削減のために低解像度のカメラしか使用できない場合が想定され、深層学習モデルの訓練とテスト時で解像度の変化によるひび割れ抽出性能の低下が課題となる。本研究では、訓練データ内のサンプルに対して、テスト時の環境に出現する確率の高いサンプルを重要サンプルとして自動で判別する重要度重みづけモデルを学習した。実験では、重要サンプルに対して大きい重みを付与することで、ひび割れ抽出モデルの学習をテスト時の環境に適応できたことを確認した。

「第 5 章 結論」では、第 2 章から第 4 章までで得られた検討結果の総括と今後の課題について述べた。第 2 章、第 3 章での本研究の成果は、被験者単位でのラベルのみ与えられた弱教師あり学習により、疾患領域のラベル付けコスト抑制と高精度な疾患の検出を実現する手法を開発したことである。特に分割画像に対する位置情報の活用で信頼できる領域の選択を行い、従来の弱教師あり学習から精度を向上させた。第 4 章での本研究の成果は、共変量シフト適応により、ひび割れ抽出での複数解像度によるドメイン適応を実現したことである。特にテスト時の環境で得られる画像が低解像度であるとき、訓練データの重要サンプルに対して自動での重みづけを可能とし、重みづけしない場合と比較してひび割れ抽出性能を向上させた。

Abstract

The rapid proliferation of computational resources such as GPUs, together with the widespread adoption of IoT devices including smartphones, has significantly accelerated efforts to acquire and analyze large volumes of data. In particular, machine learning techniques—especially those based on deep learning—can achieve highly accurate predictions by leveraging large-scale datasets. However, collecting labeled data and data obtained under controlled test environments is often costly and labor-intensive. To address this challenge, increasing attention has been paid to learning frameworks that exploit weak supervision, such as ambiguous or noisy labels, as well as unlabeled data, which can be acquired at relatively low cost. By utilizing such data for model training, it becomes possible to reduce data collection costs while maintaining reliable machine learning performance. This study investigates medical images and crack images, with the objective of supporting medical practice and visual inspection of infrastructure facilities under data-scarce conditions. By identifying informative samples from limited datasets, we aim to reduce data collection costs while maintaining high prediction accuracy.

Chapter 1, Introduction, provides an overview of general machine learning methodologies and their applications in real-world settings. In addition to summarizing representative techniques, this chapter emphasizes the critical importance of data itself, rather than algorithms alone, in determining system performance. Based on this discussion, the chapter clarifies the contributions and proposals of this research with respect to the medical domain and infrastructure inspection.

Chapter 2, Improving the Accuracy of Cirrhosis Detection via Reliability Evaluation Using Spatial Information, addresses the problem of cirrhosis detection from ultrasound images under a weakly supervised learning setting. In this scenario, no pixel-level or region-level annotations of disease areas are available, and only image-level labels are provided. To address this challenge, the proposed method evaluates the reliability of local patch images by comparing each patch with its surrounding region. Through this reliability evaluation, disease-related and healthy regions within an image are selectively identified, and only reliable patches are used during

model training and testing. This approach enables more accurate cirrhosis detection. Conventional methods rely on all patch images automatically generated by raster scanning, which often include regions irrelevant to cirrhosis detection, such as ultrasound noise, acoustic shadows, and blood vessels. These irrelevant regions can negatively affect both model training and classification performance. Experimental results demonstrate that applying reliability evaluation based on spatial information to restrict learning and classification to reliable patch images is effective in improving subject-level cirrhosis detection performance.

Chapter 3, *Enhancing CNN-Based Cirrhosis Detection by Integrating Spatial Reliability Evaluation and Multiple Instance Learning*, extends the reliability evaluation method validated in Chapter 2 by applying it to deep learning. In this chapter, convolutional neural networks (CNNs) are employed to automate image feature extraction, enabling the design of feature representations and classifiers that do not depend on human expertise. In Chapter 2, Gabor features and Quadratic Discriminant Analysis were selected based on prior experience. However, such manually designed systems suffer from subjectivity, as detection performance can strongly depend on the designer’s expertise. In contrast, CNN-based cirrhosis detection methods have been proposed that employ Multiple Instance Learning (MIL) as a framework for reliability evaluation based on image features. In conventional MIL-based approaches, regions of interest (ROIs) are manually defined according to medical expertise, resulting in high dependency on the quality of the selected ROIs. To overcome these limitations, this chapter integrates the proposed Reliability Evaluation Method (REM) into the MIL framework while using CNNs for both feature extraction and classification. The method operates on automatically generated ROIs, enabling the selection of reliable ROIs for training and testing without human intervention. Experimental results confirm that this integration improves classification performance and reduces dependence on manually defined ROIs.

Chapter 4, *Improving Concrete Crack Segmentation via Covariate Shift Adaptation for Multi-Resolution Images*, focuses on the segmentation of cracks on concrete surfaces. This task is essential for automating visual inspections of large-scale infrastructure such as tunnels and bridges. In practical environments, images are often captured under varying conditions, for example using drones, where the distance between the camera and the target object is not constant. Additionally, cost constraints may require the use of low-resolution cameras. These factors cause discrepancies in image resolution between training and testing data, leading to degraded performance of deep learning models. To address this issue, the proposed method applies covariate shift adaptation by learning an importance-weighting model that

automatically identifies samples in the training dataset that are more likely to appear under test-time conditions. By assigning higher weights to such important samples during training, the crack segmentation model can better adapt to the test environment. Experimental results demonstrate that this approach effectively mitigates performance degradation caused by resolution differences and improves crack extraction accuracy compared to models trained without importance weighting.

Finally, Chapter 5, Conclusion, summarizes the findings from Chapters 2 through 4 and discusses future research directions. The key contributions of Chapters 2 and 3 include the development of methods that reduce labeling costs while achieving high-accuracy disease detection under weakly supervised learning settings, where only subject-level labels are available. In particular, leveraging spatial information to select reliable regions significantly improves detection performance over conventional weakly supervised approaches. The primary contribution of Chapter 4 is the realization of domain adaptation across multiple image resolutions through covariate shift adaptation. By automatically weighting important training samples when test-time images are low resolution, the proposed method achieves superior crack extraction performance compared to unweighted training approaches.

第 1 章

序論

1.1 機械学習

この章では、一般的な機械学習について、手法、応用先と得られるメリットについて述べた後、機械学習モデルの予測精度を向上させるために必要なデータの重要性について述べる。その後、収集コストを削減しつつ精度良いモデルを学習する一般的な手法である弱教師あり学習について述べる。本論文では医療分野とインフラ分野への機械学習の応用に研究対象を絞り、最後に本研究の貢献を述べる。

1.1.1 機械学習

スマートフォンの普及に代表されるデータ爆発により、過去数十年前と比較して私たちは膨大な量のデータを収集、活用することが可能となった。普段の生活の中でスマートフォンは欠かせない存在であり、私たちの属性（性別、生年月日など）だけでなく撮影された画像には容姿、オンライン購買履歴では好みの製品やブランドなど様々な個人情報が詰まっている。GAFAMをはじめとするインターネット企業はこれらの情報から機械学習アルゴリズムを活用することで個人に最適なコンテンツをレコメンドし、よりユーザー満足度の高いサービスを提供している。また、消費者向けと同時に企業間でのビジネスや企業内でのデジタルトランスフォーメーション (DX) にも機械学習は活用されている。製造業における在庫最適化では、小売り店舗と生産工場の間地点として大規模配送センターで一時在庫を保管するが、注文が入ってから配送センター内の在庫を特定し出庫するまでの流れを機械学習を活用した自動化によって省人化している。近年の EC マーケット成長に起因して配送センターの自動化・省人化は急速にニーズが高まっている。これら、個人・企業の利益を最大化する目的の他、オンライン上の安全を守る目的では、スパムメールやサーバーへの不

正ログインの特定・解析にも機械学習が活用されている。

一般的に機械学習とは、特徴量（もしくは説明変数） x と対応するラベル y のペアで構成されるデータを扱い、 $y = f(x)$ の数式で表せられる関数について、 x が入力された際に対応するラベル y を出力する関数 $f()$ を学習データから獲得することをさす [1]。ここで、関数 $f()$ は SVM, ランダムフォレスト, ニューラルネットワークなどといった汎用性の高いモデルがこれまで提案されている。

実応用の場面では、機械学習を活用するシステム設計者の経験にもとづいてヒューリスティックにモデルが決定される。特に比較的大規模なデータが利用可能となった現在では、ニューラルネットワークを多層化した深層学習が特徴量とラベルの複雑な関係性を自動で学習し、高精度な予測を実現可能なため、モデル選択肢の中で有用であるといわれている。近年の研究では、データ量を増加させていくごとにモデル精度も増加していくことが示されている [2]。

これまで述べた機械学習では、入力データとなる特徴量 x に対して対応するラベル y が必ずモデルの学習に活用できる教師あり学習に着目した。一方で、大規模なデータセットの活用においてすべてのデータに対して正確なラベル付けは人的・時間的コストが非常に高い。例えば、ソーシャルメディア上でのユーザーの広告クリック行動を機械学習で予測する際は、画面上のクリックされなかった広告がユーザーにとって魅力的でなかったのか、そもそも広告を認知するまえに他のコンテンツをクリックして画面遷移したのかをログから判断してラベル付けすることは困難である。衛星画像を用いたリモートセンシングでは、同じ地表面の座標を対象としても、画像を撮影する季節や年代の差異によって対応する地表面被覆クラスは変化する。これらの問題設定では、入力となる特徴量に対して完全ではないが弱いラベル情報が与えられる設定を想定し、弱教師あり学習という枠組みで精度良い予測を実現するモデルの学習手法が研究されてきた。

画像認識分野で最も広く研究されている弱教師あり学習の一つが Multiple Instance Learning (MIL) である [3]。教師あり学習では各画像に対して一つのラベルが付与されていたが、MIL では複数のパッチ画像の集合に対して一つのラベルを付与する。ここで、パッチ画像をインスタンスと呼び、インスタンスで構成される画像集合をバッグと呼ぶ。2 クラス（正/負）の分類において MIL では、正クラスのインスタンスを最低 1 枚含むバッグに対して正クラスのラベルを付与し、すべてのインスタンスが負クラスである場合にバッグに対して負クラスのラベルを付与する。モデル学習時にはインスタンス単位の詳細なラベルを必要とせず、バッグ単位のあいまいなラベル情報のみ付与することで学習を可能にし、インスタンスのラベル付けコスト削減を可能にしている。

また、学習データ全体の中で少量の正クラスにラベル付けされたデータと多量のラベルなしデータを用いた学習法も PU 分類として研究されている。PU 分

類ではあらかじめラベルなしデータに占める正クラスデータの割合を既知とする．ラベルなしデータを負クラスと考え学習を行うと同時に，ラベル無しデータ内に存在する正クラスへの学習をラベル付き正クラスデータを用いて打ち消すことで，正負の適切な分類境界を獲得している [4, 5, 6]．

1.1.2 機械学習の社会への応用

本節では，先の節でふれた活用例のほかに，医療分野およびインフラ分野における機械学習の活用について，画像データに着目して述べる．

医療分野

医療画像を用いた機械学習では，入力された画像を用いた疾患の分類や領域分割（セグメンテーション）が最も取り上げられる領域であり，その他にも画像再構成，説明可能な AI(xAI)，大規模言語処理モデルとの統合などがある．

機械学習を活用した医療分野での研究に頻繁に使用される画像形式として，CT 画像，MRI 画像，超音波画像がある．CT 画像は X 線吸収率の差を利用し，複数方向から得られた投影データを再構成することで生成された断層画像である．空間分解能が高く，画像取得速度が速いことが特徴である．一方で，短所として放射線被ばくを伴う．MRI 画像は磁場と電波を利用し，体内組織の陽子を揃えてから元の状態に戻るまでの時間差を利用する撮影方法である．MRI 画像と比較して撮影に時間がかかることが短所であるが，放射線被ばくのリスクがない．超音波画像は高周波音波を体内に送信し，組織表面からの反射波を受信して強度変化を画像化する．撮影から画像化までのリアルタイム性が高く，心臓や胎児など動的構造を観察することが可能である．一方で，音波の減衰・反射によって深部や臓器の背後は画像化が困難である．また，CT 画像や MRI 画像と比較して，ノイズが多いため観察結果が定量的になりやすい．

疾患の分類では，これら医療画像を機械学習モデルに入力し，各画像もしくは被験者ごとに疾患の存在する確率を予測する．ここで，教師あり学習を用いたモデルの学習では，学習データの各画像に対して医師によるラベル付けが必要であり，人的コストが大きくなる傾向にある．解決策として，畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）を用いた場合に少量の学習データからでも精度良い分類を可能にするため，データ拡張による学習データの増しや事前学習による重みの転用が用いられる [7, 8]．データ拡張を行う際は，分類対象となる疾患の画像内所見が変化しラベル情報と画像特徴の不一致が発生しないように注意する必要がある．事前学習を用いた CNN モデルの活用では，多量に収集することができる自然画像を活用することが多く，これらの画像から得られる画像特徴と疾患の分類に有用な画像特徴のギャップを

考慮した学習が求められる。

医療画像を用いたセグメンテーションでは、医療画像内の特定の臓器や疾患部位を画素単位で予測・抽出する。セグメンテーションモデルの教師あり学習を行う際は、画素単位でのラベル付けが必要となり、分類問題と比較してよりラベル付けのコストが高い。疾患の分類と同様にセグメンテーションにおいてもデータ拡張や事前学習による重みの転用が用いられる。また、近年ではセグメンテーションの基盤モデルが提案され、ゼロショットでのセグメンテーションを実現している [9, 10]。

インフラ分野

インフラ分野への機械学習の応用では、設備の劣化や損傷のモニタリングが広く議論される領域の一つである。対象とするインフラ設備は建物、橋梁、ダムを含むコンクリート構造物やアスファルト舗装路面の他に、エネルギーインフラを支える化学プラントや電力インフラを支える送電線、変電設備など様々である。化学プラントや送電線、変電設備を対象とした機械学習によるモニタリングでは、インフラ設備が正しく機能しているかを判定するため、配管に設置された流量計や電線に設置された振動センサで得られるデータを用いる。多くの場合で正常な値からの差異に着目し、設備の劣化兆候や機能不全を検出する。コンクリート構造物やアスファルト舗装路面を対象とした機械学習による劣化や損傷のモニタリングでは、最も研究対象として扱われるものの一つが疲労や過荷重によって表面に発生するひび割れである。ひび割れの検査を行う際、画像は RGB やグレースケールの可視光画像が一般的に用いられる。

1.2 超音波画像を用いた肝硬変診断

本節では、医師による超音波画像を用いた目視での肝硬変診断について説明し、機械学習を用いた医師の支援の有効性を述べる。

1.2.1 肝硬変

肝硬変は、線維化（結合組織の生成と沈着）と呼ばれる過程の異常な継続によって引き起こされる肝疾患である [11]。図1.1は肝臓の健常と肝硬変の状態を比較するイメージ図である。線維化を経た肝臓組織では、局所的な組織の巨大化（結節性と呼ばれるしこりができた状態）や隔壁の発生、肝臓組織の硬化などによって肝機能が低下する。ここで、肝臓表面に発生する結節性はその大きさによって、Micronodular (<3mm, 微小結節性) と Macronodular (>3mm, 大結節性) に分類される [12]。肝硬変は一般的に、フィブロスキャンと呼ばれる肝臓組

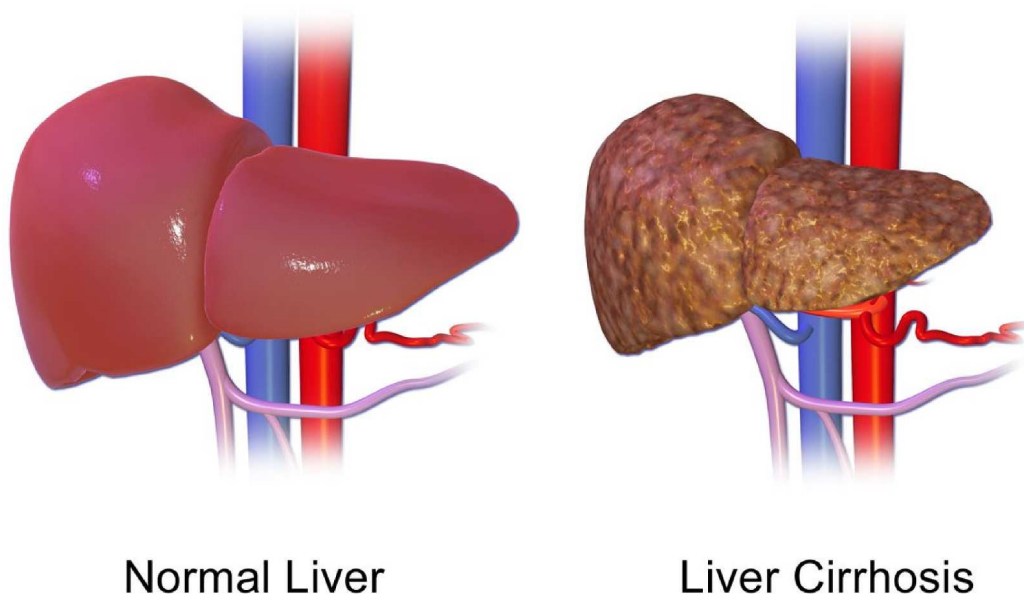


図 1.1: 肝臓の正常と肝硬変の比較.
 出典: BruceBlaus, “Liver Cirrhosis.png”, CC BY-SA 4.0, Wikimedia Commons,
https://commons.wikimedia.org/wiki/File:Liver_Cirrhosis.png

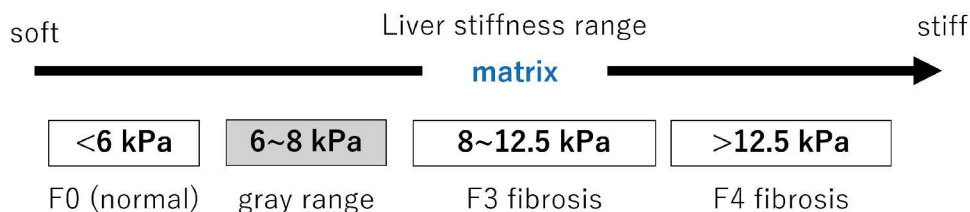


図 1.2: フィブロスキャンによる肝臓の硬さと肝硬変ステージ F0 ~ F4 の対応 [13].

組織の硬さを計測する手法を用いて、健常 (F0) を含めた F0~F4 の 5 段階で肝硬変の進行がステージ分けされる [13]. 図1.2にフィブロスキャンの値と肝硬変ステージ F0~F4 の対応を示す. ここで、6~8kPa の範囲では線維化がなくても炎症など可逆的な要因によって観測される可能性があり、純粋な線維化状態を反映していない可能性があるため、6~8kPa を単純なステージ F1~F2 と区別している. ステージ F1~F3 では肝臓組織の一部で健常な領域が存在するが、F4 では肝臓組織全体にわたって結節性が形成されている. また、ステージ F4 の肝臓で、肝機能が正常に働いている肝臓を **Compensated cirrhosis** (代償期)、肝機能が働いていない肝臓を **Decompensated cirrhosis** (非代償期) と分類する. 肝硬変の進行により肝がんの発生リスクが非常に高くなるため、早期発見が不可欠である.



図 1.3: 超音波画像による診断のイメージ.

出典: waketechcc / Wikimedia Commons, CC0 1.0,

[https://commons.wikimedia.org/wiki/File:](https://commons.wikimedia.org/wiki/File:Sonography_Machine_Donation_from_Raleigh_Radiology-35.jpg)

[Sonography_Machine_Donation_from_Raleigh_Radiology-35.jpg](https://commons.wikimedia.org/wiki/File:Sonography_Machine_Donation_from_Raleigh_Radiology-35.jpg)

1.2.2 肝硬変診断方法

肝硬変の診断には、肝生検、血液検査、フィブロスキャン、超音波画像診断など、様々な方法がある。早期の臨床的疑いの段階では、非侵襲的な検査であるフィブロスキャンや図1.3に示す超音波画像診断が用いられる。血液検査は一定程度侵襲的な診断方法であり、非侵襲的診断方法と比較して信頼性が高い。肝生検は図1.4に示すように肝臓組織を直接検査するため、肝硬変の確定的な診断法として知られている。肝生検の結果は信頼性が高い一方で、手術によって肝細胞を採取する必要があるため、患者への負担は大きくなる。さらに、肝臓組織全体でみると 1/50000 しか診断根拠としてカバーできていないため、採取する組織の品質に診断結果が大きく左右される [14]。

1.2.3 超音波画像

図1.5,1.6に肝硬変患者と健常者の超音波画像を示す。超音波画像は一般的な可視光画像の撮影方法と異なり、プローブと呼ばれる機器から発信された超音波が人体組織のインピーダンスの違いによって反射し、プローブで観測される際の強度によって画像撮影される。プローブから発信される超音波は周波数、

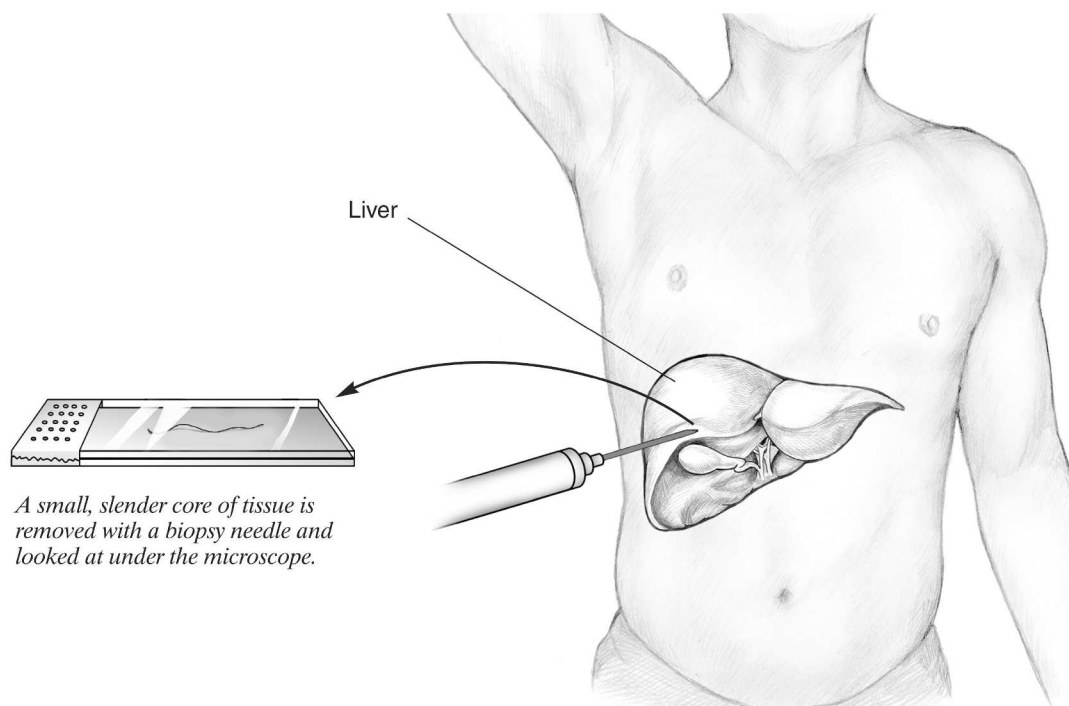


図 1.4: 肝生検による診断.

出典 : Wikimedia Commons, Public domain,

https://commons.wikimedia.org/wiki/File:Human_liver_biopsy_sample.jpg

表示深度, といったパラメーターで決定される. 周波数は超音波の周波数を表す. 周波数が高いほど撮影される超音波画像の分解能が高くなるが, 超音波減衰の影響を受けやすく人体深部まで到達せず画像化できない領域が増加する. 一方で周波数が低いほど分解能は低下するが, 超音波減衰の影響を受けづらく人体深部まで画像化が可能となる. 表示深度はプローブ表面からの最大表示距離を表し, 観察したい臓器の最遠端を少しだけ超える位置に設定される. 設定された表示深度によって, **Pulse Repetition Frequency** (超音波パルス送信頻度) や **Transmit Focus** (送信フォーカス) といった, 画像化に最適なパラメーターが自動で設定される [15]. また, 超音波画像撮影装置は連続波ではなく短いパルスを送信するため, 1 パルスあたりに含まれる正弦波の数をサイクル数と呼び, 1 パルスあたりの長さ (パルス長) を **Spatial Pulse Length (SPL)** と呼ぶ.

超音波画像の特徴として, 実際の撮影対象とは一致しない見かけ上の像であるアーチファクトの発生がある. アーチファクトは以下に例として示すように複数種類存在し, それぞれ異なる発生原理や見え方が特徴である.

- **スペckルノイズ**: 超音波が散乱体 (光や波が当たると, その進行方向を変えて四方八方に広がる (散乱する) 原因となる物質や構造) に反射して発生. 見え方は, 均一性のあるランダムな画素値.
- **後方エコー増強**: 線維化の中に存在する脂肪変性や小嚢胞様構造は減衰が

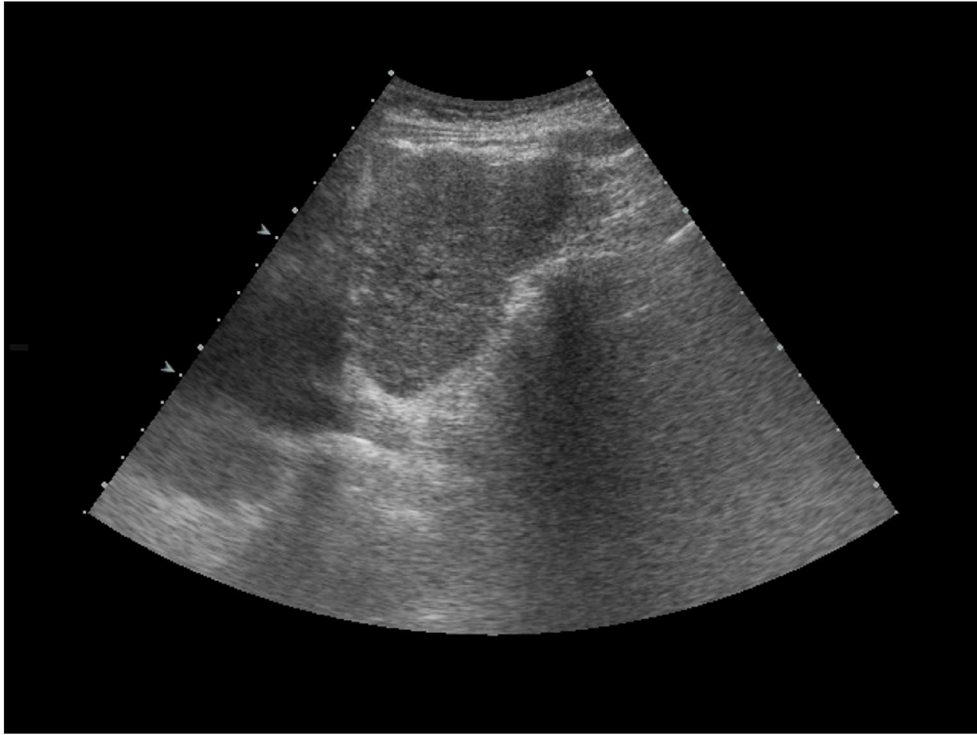


図 1.5: 肝硬変疾患の超音波画像.



図 1.6: 健常者の超音波画像.

少なく、その後方で音圧が増加。結節の背後だけが異常に明るく、偽の低エコー結節が浮き上がる見た目。

- 音響陰影：線維化や石灰化した再生結節で音の減衰が増大。見え方は、結節背後に暗い帯で、偽の欠損像。
- リバーブレーション：2つの強い反射面の間で超音波が多重反射する。往復時間が長くなるたびに、より深部に“偽の像”として表示される。見え方は、等間隔に並ぶ平行な高輝度エコー（横線状）。

これらの中でもスペックルノイズは画像全体の鮮明性を大きく低下させるため、最も影響が大きいとされる。

1.2.4 超音波画像を用いた肝硬変診断

超音波画像を用いた肝硬変診断で医師は、以下の所見をもとに診断を行う [16, 17]。一般的に、最も上段の肝臓表面の結節性が肝硬変検出に効果的であるとされる [18]。

- 肝臓表面：健常な領域では表面が面状の特徴を持つ。線維化による肝臓表面の粗さ、小さな突起状の結節性と隔壁の形成、不規則な肝臓表面の特徴が肝硬変の領域の特徴とされる。
- 肝臓の縁：健常な肝臓の下縁は V 字型に近い形状だが、肝硬変によって丸みを帯びた鈍い形状に変化する。
- 肝臓サイズ：健常な肝臓と比較して、小さい（進行した肝硬変に相当）もしくは大きい（初期の肝硬変に相当）場合に肝硬変とされる。
- 門脈圧亢進症の兆候：肝静脈/門脈の拡張、脾臓の腫大、および腹水。

ここで注意すべきなのは、超音波画像撮影時に使用する周波数や肝硬変の進行度合いによって肝臓表面で確認できる特徴が変化する点である。撮影時の周波数が高いほど分解能が高くなり結節性などの線維化をはっきりと表示できる一方で、肝臓といった腹部深部に存在する臓器の撮影では超音波減衰によって所望の臓器まで適切に超音波が到達せず、画像として表示が困難な可能性がある。反対に撮影時の周波数が低いほど超音波減衰が抑制されるが、分解能が低下するため線維化を鮮明に表示できなくなる。そのため、画像の鮮明性と画像として表示できる臓器領域の広さにはトレードオフの関係があり、診断時には医師による経験をもとに撮影が行われる。

また、工学的観点で肝硬変の目視診断における所見を研究した論文では、画像上で確認できる特徴が肝硬変の構造（実態）を表したものであるとする立場

から、スペックルノイズによって構造が直接画像化されずスペックルノイズの歪みに着目して診断を行っているとする立場まで様々な見方が報告されている。Di Lelio らは 1989 年に、周波数 5MHz、短焦点の small-parts プローブを用いた肝表面の観察によって評価される肝表面の不整は、肝硬変の肉眼的外観と直接相関する客観的所見であると報告している [19]。Colli らは 2003 年に、肝臓表面の結節性は、5~12 MHz のトランスデューサーを使用して肝表面全体と肝実質の外側 2~3 cm を調べることで評価した。超音波画像では、肝臓表面の結節は再生結節と線維性隔壁の存在を表現しており、これらは肝硬変の診断に不可欠な組織学的所見であると報告している [20]。Filly らは 2002 年に、超音波画像の肝臓領域の腹壁側（画像上部、プローブ直下）よりも深部（画像下部、脂肪に接する肝臓表面）で観察した方が、高感度で肝硬変を検出できると報告している。深部の方が結節性をより見えやすい理由として Filly らは、隣接する脂肪組織が高いコントラストを提供していることを挙げている。一方で、腹壁側では近傍界アーチファクト（near-field artifacts）の影響により結節の判別が制限され、観察部位という物理条件が診断精度に大きく影響することを示した [21]。また Shen らは 2006 年の報告で、超音波画像を用いた目視診断について肝臓表面の結節性や肝実質内におけるエコーの不均一な分布が進行した線維化と関連するとしている。つまり、超音波における肝表面の結節性は、病理学的な再生結節や線維性隔壁の存在を反映したものであるとする一方で、周波数の低さやアーチファクトの影響によって超音波画像でみられる所見は病理診断と完全には一致しないとして [22]。また Toyoda らは 2009 年に、超音波画像に含まれるスペックルノイズは多数の微小散乱体からの反射波が干渉して生じるものであり、正常な肝臓では RF エコー振幅の統計分布が Rayleigh 分布に近似されることを報告している。一方で、線維化が進行した肝臓では、結節性や隔壁などの構造によってエコー振幅分布が Rayleigh 分布から逸脱し、統計的に有意差 (F_2, F_3 vs F_0, F_1) が生じる。Toyoda らは Rayleigh 分布からの乖離を修正カイ二乗統計量 C^2_m (観測されたエコー振幅の修正分散と、Rayleigh 分布を仮定して平均値から理論的に計算される分散との比) として定量化し、線維化の進行度評価に有用であることを示した。また、結節性や隔壁が小さい初期段階においては、スペックルノイズと統計的に区別することが困難であることも報告している [23]。これに対し Igarashi らは 2010 年に、Toyoda らが用いた Rayleigh 分布からの乖離を発展させ、線維化におけるエコー振幅分布を、正常組織由来成分と線維化組織由来成分からなる 2 つの Rayleigh 分布の混合としてモデル化した。この混合 Rayleigh モデルに基づき、エコー振幅分布の 2 次および 3 次中心モーメントを用いることで、線維化組織の量を表す混合率と、線維組織の音響的特性を反映する分散比を同時に推定可能であることを示した。さらに、これらの統計量から線維化の進行度を逆問題として定量的に評価できる評価チャートを構築し、臨床デー

タにおいて線維化ステージとの良好な対応関係が得られることを報告している [24]. Gerstenmaier らは 2014 年に、超音波画像内においてスペックルとは多数の散乱体の干渉によって形成される強度パターンであり、基礎となる肝臓組織の構造を直接表現したものではないと報告している。一方で、スペックルの局所的な明るさは、対応する散乱体のエコー強度を反映しており、肝臓超音波の場合、その散乱体はおそらく肝小葉である。スペックルパターンは脂肪化および線維化の両方によって変化し、微細な組織構造の変化は、従来の B モード超音波画像では肉眼では知覚できない場合があると報告している [25].

これらから、肝硬変診断時の根拠として結節性や隔壁といった線維化の所見を超音波画像上で直接観察している場合と、結節性や隔壁が直接表現されていないためスペックルノイズの均一性の差異を観察している場合の 2 パターンが考えられ普遍的な診断根拠の同定は困難である。そのため医師は、超音波画像を用いた目視診断時にこれらの複合的な要素を考慮していると考えられる。

1.2.5 超音波画像を用いた医師の目視診断の精度

超音波画像を用いた肝硬変の目視診断の精度は医師の経験や診断に使用する機材と撮影の条件設定、被験者の進行度合いなど様々な要因に左右される。以下、この節での識別率（正診率）を、正しく診断した被験者数/全被験者数、感度を、正しく診断した肝硬変患者数/全肝硬変患者数、特異度を、正しく診断した健常者数/全健常者数、とする。Huber らでは、超音波画像を用いた医師の目視診断の精度は識別率 64–79%、感度 52–69%、特異度 74–89% と報告されている [17]. Choong らでは、156 名 (F0 (健常) : 57 名, F1 : 34 名, F2 : 20 名, F3 : 22 名, F4 : 23 名) の被験者について、健常 vs 肝硬変のカットオフを F1-F4 の 4 段階で設定し、それぞれでの診断精度を検証した [18]. F1 (軽度の線維化) -F4 のすべてを肝硬変クラスとした場合 (括弧は 95% 信頼区間)、識別率 56 (48-64)%, 感度 28 (20-38)%, 特異度 88 (76-95)%, F2 以上を肝硬変クラスとした場合、識別率 59 (51-67)%, 感度 32 (21-45)%, 特異度 85 (76-91)%, F3 以上を肝硬変クラスとした場合、識別率 62 (54-70)%, 感度 40 (26-56)%, 特異度 85 (77-91)%, F4 以上を肝硬変クラスとした場合、識別率 66 (58-73)%, 感度 48 (27-69)%, 特異度 82 (74-88)%, であった。F1 から F4 へ肝硬変と判定するカットオフを移行するにつれて識別率と感度が上昇している。線維化以外の肝臓表面の炎症や脂肪肝といった複雑な所見が混在することから、初期の肝硬変 (F1-F3) では超音波画像を用いた正確な診断が困難であるとしている。Ong らでは、88 名の被験者で肝生検による診断後に超音波画像での診断精度を計測した [26]. 識別率が 76.1%, 感度 37.5%, 特異度 84.7% であり、感度が低くなった。この原因として、使用した超音波の周波数が 3.5~5Mhz の低周波プローブを使用しており周波数が 7.5Mhz 以上の高周波の方が肝臓表面をより詳細に描出で

きるためであるとしている。また、診断結果が医師の経験に依存することに加えて、事前に行った肝生検も原因の一つとしている。この論文では経皮肝生検 (Percutaneous biopsy, 体表から盲目的または超音波画像を頼りに針を刺す) を行っており、腹腔鏡肝生検 (Laparoscopic biopsy, 腹腔鏡で肝臓表面を直接観察する) と比較して、結節性を狙って採取できる確率が低下することによる偽陰性の被験者が一定数想定される。そのため、肝生検による診断結果が健常であり、超音波画像による診断では偏った肝硬変患者を母数としたことで、感度が低下したとしている。Hetland らでは、315 名の被験者 (肝硬変患者: 115 名, 健常者: 200 名) を対象とし、医師による超音波画像の診断精度が感度 71 (62-79)%, 特異度 94 (90-97)%, 対象を代償期の被験者に限定した際は、感度 62 (49-74)%, 特異度 94 (90-97)% であった [16]。特に、症状のない初期の肝硬変 (代償期) では肝硬変の見落としが多い傾向にあり、肝臓表面の所見と合わせて、血流速度を診断根拠とすることで 82% まで診断精度が向上することを報告している。

これらのように、超音波画像では肝硬変の所見が (特に初期の F1-F3, および F4 の非代償期において) 不明瞭で診断は困難であり、医師の経験によって診断結果が異なる可能性がある。そのため、超音波画像と合わせて他の診断手法を用いることで、精度の良い診断が望まれている。目視での診断を支援するために、機械学習による肝硬変患者の検出も活用されている。機械学習による肝硬変の検出では、技術的には多量のデータをもとに機械学習モデルを学習可能であり、ベテラン医師の育成と比較して時間的コストが低減できる可能性がある。また、学習したデータにもとづき判定を行うため、医師の疲労や主観的バイアスの影響を受けない。これらの利点から、目視診断の際はセカンドオピニオンとして活用されている。

第 2 章、第 3 章で使用するデータセットについては、付録に詳細を記載している。肝硬変患者か健常者かの判定は、医師によって行われた。ここでフィブロスキャンの値によると、データセットを構成する肝硬変患者 12 名は全員がステージ F4 まで進行しているが、代償期か非代償期かの確定的な判定はできない。代償期であった場合には健常な領域が存在していない場合でも、結節性が小さいため超音波画像上では滑らかな健常領域に見える可能性がある。また、健常者の中にも脂肪肝の被験者が存在し、不均一な脂肪沈着によって、肝硬変患者にみられるような超音波画像上での粗さや結節性の所見に見える可能性がある。これらの要因より、本データセットでは肝硬変患者の肝臓表面にも健常な領域が存在し、健常者の肝臓表面にも肝硬変の特徴を持つ領域が存在するように見える可能性が高く、機械学習による分類の精度低下の原因となる。



(a) トンネルに発生したひび割れ。
出典：Xu ら (2023) [30] より引用 (CC BY 4.0)



(b) 舗装路面に発生したひび割れ。
出典：Andy Beecroft / A Cracking Road There / CC BY-SA 2.0
(https://commons.wikimedia.org/wiki/File:A_Cracking_Road_There_-_geograph.org.uk_-_1460260.jpg より引用)

図 1.7: トンネルと舗装路面に発生したひび割れの例。

1.3 コンクリート構造物のひび割れ調査

本節では、コンクリート構造物のひび割れ調査について説明し、機械学習を用いた支援の有効性を述べる。

1.3.1 コンクリート構造物

日本では 1950 年代後半から 1970 年代にかけての高度経済成長の過程で、舗装道路、トンネル、橋梁といったインフラ設備が集中的に整備された [27, 28]。図1.7にトンネルと舗装路面に発生したひび割れの例を示す。国土交通省は建設後 50 年を老朽化の一つの目安としている [29]。日本は近年高齢化の影響により 65 歳以上人口の割合が 2025 年時点で 29.4% に達したことが示されており、2070 年には総人口が約 30% 減少し、65 歳以上が約 40% を占める見通しが示されている。インフラの老朽化と同時に維持管理を行う人員の減少が進行するため、機械学習をはじめとするデジタル技術による自動化が求められる。

1.3.2 コンクリートのひび割れ

コンクリートは引張に弱く、内部に生じた引張応力 (引き延ばす力) が引張耐力 (引き延ばす力に対して破断せずに耐えられる限界値) を超えると、表面にひび割れが発生する。ひび割れの定義は様々な指標で行われる。

- ひび割れ幅：最も一般的な指標 [31]
 - RC (鉄筋コンクリート) 構造物：ひび割れ幅 0.3mm 以上を「大」, 0.2mm 0.3mm 未満を「中」と定義.
 - PC (プレストレストコンクリート) 構造物：RC 構造物よりも厳しく, 0.2mm 以上を「大」, 0.1mm 0.2mm 未満を「中」と定義.
- ひび割れ長さ：延長, 連続性
- 方向：鉄筋方向に沿う (腐食起因の疑い), 斜め (せん断など) の推定
- ひび割れ密度：単位面積当たりの本数・総延長
- 貫通性・深さ：表面のみか, かぶり～鉄筋近傍まで達するか (耐久性に影響)

1.3.3 ひび割れ調査

目視でのひび割れ調査を支援するため, 画像データを用いたひび割れ自動抽出が近年提案されている. ひび割れの抽出を行う際, 画像は RGB, グレースケールの可視光画像が一般的に用いられる. 機械学習によるひび割れ自動抽出は, 必ずしも信頼できるものではなく, あくまで目視検査の支援として使用される. 重大なひび割れを見逃しが最も損失が大きいため, 感度 (Recall) を評価指標とした見逃しの少ない出力が求められる.

また, 対象物のロケーションが環境によって異なることから, 各対象物の環境を考慮した手法の設計が必要である. 例えば, ダムや橋梁といった地表面から一定の高さに存在する対象物を検査対象とする場合, 人手による画像の撮影は困難であり, ドローンや壁面走行ロボットを用いた撮影が行われる. また, 人手による画像撮影が可能な場面でも, トンネル壁面や梁裏では空間的制約があり正面からの画像撮影ができない場合がある. 斜め方向からの画像撮影では, インフラ構造物やひび割れなどの検査対象物に幾何歪みが発生したり, 焦点ずれや照明ムラの発生したりする. そのため, ひび割れの実寸と異なる見え方や一部箇所がぼやけた見え方になり, 検出・抽出の精度低下を招く.

1.4 実験で使用する評価指標

この節では, 2~4 章で使用する評価指標として, 感度 (Sensitivity, または Recall), 特異度 (Specificity), 識別率 (Accuracy), 適合率 (Precision), F 値 (F-measure), Area Under the Curve (AUC)[32], Intersection over Union (IoU) を定義する. 本論文では一貫して機械学習における 2 クラス分類を扱うため, 陽性

クラスを Positive (P), 陰性クラスを Negative (N) として表現する. また, モデル出力が実際のラベルと一致している場合を True (T), 一致していない場合を False (F) として表現する.

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1.1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (1.2)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \quad (1.3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1.4)$$

$$\text{F 値} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \quad (1.5)$$

$$\text{AUC} = \frac{1}{\text{Num}_P \times \text{Num}_N} \sum_i \text{Trapezoid_Area}(\text{FP}_i; \text{TP}_i) \quad (1.6)$$

$$\text{IoU} = \frac{|\mathbf{Y}_{\text{pred}} \cap \mathbf{Y}_{\text{true}}|}{|\mathbf{Y}_{\text{pred}} \cup \mathbf{Y}_{\text{true}}|} \quad (1.7)$$

ここで, TP は Positive クラスと正しく予測されたサンプルの数, FN は Negative クラスと誤って予測されたが実際には Positive クラスのサンプルの数, TN は Negative クラスと正しく予測されたサンプルの数, FP は Positive クラスと誤って予測されたが実際には Negative クラスのサンプルの数を表す. 感度は Positive クラスのサンプルを正しく識別する割合を示し, 特異度は Negative クラスのサンプルを正しく識別する割合を示す.

AUC は受信者動作特性曲線 (Receiver Operating Characteristic curve, ROC) 下の面積を表し, 値が 1 に近いほどモデルの分離性能が高いことを示す. ここで, $\text{Num}_P, \text{Num}_N$ はそれぞれ, Positive クラスと Negative クラスのサンプル数を示す. $\text{Trapezoid_Area}(\cdot)$ は ROC 曲線上の隣接する 2 点 $(\text{FP}_{i-1}; \text{TP}_{i-1})$ と $(\text{FP}_i; \text{TP}_i)$ が FPR 軸と形成する台形的面積を表す. すべての台形的面積を合計すると, ROC 曲線下の総面積 A が得られる. これを $\text{Num}_P \times \text{Num}_N$ で除算することにより, 座標系 $(\text{FP}; \text{TP})$ を $(\text{FPR}; \text{TPR})$ へ変換し, 単位正方形上に正規化された AUC 値が求められる.

$\mathbf{Y}_{\text{pred}}$ と $\mathbf{Y}_{\text{true}}$ は Positive クラスと予測された領域と, 実際の Positive クラス領域を示す. 分子は正しく Positive クラスと予測された領域の画素数を表し, 分母は Positive クラスと予測された領域と実際のひび割れ領域の和集合の画素数を表す.

1.5 本論文の提案および貢献

この論文での貢献は、データを構成する特徴量とラベルの片方が十分に収集できない状況においても機械学習を活用したシステムの実装現場において高精度な予測が可能なモデルの学習を目的とする。特に、全体の提案において、限られた情報からでも機械学習モデルの学習やテスト時の予測に有効なサンプルの推定を自動で行い、人手による情報の追加を削減しつつ精度良い予測の実現を提案する。

データを構成する要素のうち「ラベル情報」では、画像データに対してラベル情報の根拠となる領域が詳細にマーキングされておらず、被験者単位でのラベルのみ与えられている問題に対して、画像を分割したパッチ画像ごとに「画像空間上での他のパッチとの位置情報に基づいた信頼性や重要度を推定する手法」を提案する。これにより、モデルの学習とテスト時の予測に寄与しない領域の影響を削減し、画像全体の特徴量を入力した場合と比較して高精度な予測が可能となる。データを構成する要素のうち「特徴量」では、モデルの学習に使用する学習データと学習済みモデルを適用するテストデータで入力される特徴量の分布が変化する、共変量シフトと呼ばれる問題に対して、各学習データに対する「テスト時での出現度合いを重要度とし、重要度推定を自動で行う手法」を提案する。これにより、学習データの中でテストデータに出現する可能性の高い画像を重要度が高い画像として誤分類時の損失を高く設定し、よりテストデータの環境に適応した学習を実現することで高精度な予測が可能となる。

1.5.1 医療画像分野（ラベル情報が限られる状況での機械学習）

医療画像を用いた機械学習において、画像内での疾患部位のマーキングを伴うラベル付けは、人的・時間的コストが非常に高い。特に医療分野は人命にかかわる観点から、一度誤ったラベル付きデータセットが流用されてしまうと多数の患者の生命に危険を及ぼす可能性がある。さらにラベル付けには医学的専門知識を高いレベルで必要とし、ラベル付けを行う能力が十分な医師の安定的な確保も困難である。

1.1.1の節で述べたように、ラベル情報が限られる環境での精度良いモデルの学習を可能にする弱教師あり学習がこれまで提案されてきた。本論文では被験者単位での疾患の有無がラベル情報として付与されている場合に、画像全体からモデルの学習とテスト時の予測に有効な領域をパッチ画像単位で推定する手法を提案する。具体的には、被験者単位でのラベル情報を用いた学習の中でラベル情報に起因する領域を、各パッチ画像の特徴量と位置情報（画像空間上での距離）を参照して推定する。

第2章では超音波画像を用いた肝硬変検出に対して位置情報に基づく信頼性

評価法を提案する．超音波画像は CT や MRI といった他の画像モダリティと異なり，超音波ノイズによる画像内の鮮明性の低下が顕著である．そのため深層学習による特徴抽出器の自動最適化は困難であることが知られており，本論文では先行研究において有効性が確認された，人手で設計された画像特徴量と分類器を使用した肝硬変検出システムをベースとする．肝硬変検出システムは画像前処理，関心領域（Region of Interest, ROI）自動設定，特徴抽出，分類器の学習とテスト時の予測，ROI 単位での予測の統合による被験者単位の予測の出力，で構成される．従来では，入力された超音波画像からラスタスキャンによって自動設定される ROI すべてを使用して分類器の学習とテスト時の予測を行っていた．提案手法では一度分類器の学習を行った後，訓練データとテストデータに対しての分類器の予測結果と位置情報を用い，信頼できる特徴量を持つ ROI を選別する．ここで信頼できる ROI とは，超音波ノイズによる影響で実際の肝臓表面のテクスチャを正しく表現できていない領域や，影・血管などの肝硬変検出に不要な領域を含む ROI を指す．提案手法では，信頼できる小領域の周辺には似た画像特徴量を持つ領域が一定の広さで広がっており，分類器の各 ROI に対する出力（疾患 / 健常）を周辺領域の ROI に対する出力の統合結果（疾患 / 健常）と比較し，同じ出力結果である場合にのみ ROI を信頼できる特徴を持つ ROI として選択した．信頼できる ROI のみを用いた分類器の学習とテスト時の予測を行った結果，実験において肝硬変検出精度の向上を確認し，提案手法の有効性を確認した．また，提案手法の信頼性評価法において参照する周辺領域の広さについて，肝硬変領域と健常領域での最適な広がりの違いをとらえ，システム内で検証データを用いた自動での広さの最適化を行い，有効性を確認した．

第 3 章では第 2 章で有効性を確認した位置情報にもとづく信頼性評価のアイデアを深層学習モデルへ拡張する．まず，第 2 章で提案した肝硬変検出システムの特徴抽出部と分類器を畳み込みニューラルネットワーク（CNN）で代替する．CNN は被験者単位でのラベル情報に基づく学習を行うため，Multiple Instance Learning (MIL) 学習フレームワークを使用する．MIL では各 ROI に対するモデルの出力順で ROI をソートし，各クラスに対する出力が高い上位の任意の枚数の ROI のみを使用して，モデルのパラメーター更新を損失の値にもとづく勾配降下法で行う．その際，各 ROI に対するモデルの出力結果について，第 2 章で提案した位置情報に基づく信頼性評価法を適用する．CNN を活用することで人の経験に依存しない画像特徴量の抽出が可能となると同時に，位置情報に基づく信頼性評価を行うことで精度良いモデルの学習とテスト時の予測が可能となる．実験結果では，超音波画像のデータ数の不足から位置情報に基づく信頼性評価法と MIL フレームワークの統合による予測精度向上は見られなかったが，学習時の MIL 適用とテスト時の位置情報に基づく信頼性評価法適用により肝硬

変検出の精度向上を確認した。

1.5.2 インフラ分野ひび割れ画像（特徴量が限られる状況での機械学習）

機械学習を扱う際、一般的に学習データとテストデータは同一の確率分布に従うことが仮定されている。一方で実社会で機械学習モデルを活用する際、ひび割れの外観検査においては、学習データと実装環境で得られる画像データの分布が異なる場面が多い。要因として、画像撮影場所の違いによるひび割れ形状の違い、画像撮影機材の違い、撮影者の撮影方法の違い、といった様々な点があげられる。機械学習モデルは学習データに対しての損失を最小化するように学習が行われるため、学習データとテストデータで分布が異なる際、テストデータに対する精度は低下することが知られている。一方で、変化するテスト環境に対応する学習データを毎回個別に用意するコストは非常に高い。

本論文では共変量シフトと呼ばれる、学習データとテストデータの間に特徴量の分布が変化する環境において、深層学習モデルをテストデータに適応した学習を行う手法をひび割れ抽出モデルに適用する。具体的には、解像度が高いひび割れ画像が学習データに多く存在し、解像度が低いひび割れ画像がテストデータに多く存在するケースを扱う。ひび割れ抽出を行うセグメンテーションモデルのほかに学習データの各画像に対して重要度を推定する重要度推定モデルを用意し、両者の学習を同時に行う。重要度推定モデルは統計的学習理論の枠組みにおいてはパラメーターに対して線形なモデルを対象とし、 $r(\mathbf{x}) = p_{te}(\mathbf{x}) / p_{tr}(\mathbf{x})$ で定義される密度比の推定を行う。推定された密度比を使用して学習データに対する機械学習モデルの損失に重みづけを行うことで、学習データでのモデル訓練結果がテストデータでの損失の最小化になることが証明されている。本論文では大規模なデータの活用と複雑な特徴表現の獲得を目的として線形モデルではなく深層学習モデルを使用する。そのため、学習結果が大域的最適解ではなく局所的最適解になる可能性があることや、モデル出力の安定性のために出力層の範囲を限定させる必要があることから必ずしも密度比を正確に推定できる保証はないが、学習データの中でどの画像がテストデータに出現しそうかといった重要度を推定することを狙いにする。実験では、学習データ内の低解像度画像に対して重要度推定モデルが高い値を出力することで、ひび割れ抽出モデルが低解像度画像に対しての損失の最小化を優先的にを行い、テストデータでのひび割れ抽出精度の向上を確認した。

1.5.3 本論文の構成

これまで、一般的な機械学習とその応用、本論文の貢献と提案について述べてきた。本章以降の構成は次の通りである。

第 2 章では、位置情報に基づく信頼性評価によって、超音波画像を用いた肝硬変検出の精度向上を提案し実験結果を述べる。従来手法では自動設定されたすべての ROI を分類器の学習とテスト時の予測を行っていたため、予測に寄与しない信頼性の低い領域が学習時とテスト時に存在し、肝硬変検出精度の低下の原因であった。実験では、位置情報に基づく信頼性評価により、肝硬変検出精度の向上を示す。

第 3 章では、第 2 章で有効性を確認した位置情報に基づく信頼性評価法の適用を深層学習モデルへ拡張し、MIL フレームワークとの併用による肝硬変検出の精度向上を検討する。MIL を用いた肝硬変検出では、目視で設定された ROI を使用しており、分類器の性能が設定される ROI の品質に強く依存する。また、第 2 章で提案した人手による設計の画像特徴と分類器についても、設計者の医学および工学的知見に分類性能が強く依存する。そのため、自動設定された ROI を使用することで、ROI 設定の属人生を抑制しつつ、通常の教師あり学習と比較して MIL フレームワークを用いた学習と、テストデータに対して REM を適用することで各 ROI の位置情報を信頼性評価に取り入れ、分類性能の向上を示す。ここで、MIL と REM におけるパラメーター (Drop rate, 周辺領域の広さ) はクラス別に設定することで、各クラスで最適な ROI の信頼性評価による分類性能向上を期待する。

第 4 章では、重要度推定モデルを用いて、学習データとテストデータでの特徴量の分布が異なる環境での共変量シフト適応学習の有効性を検証する。具体的には、コンクリート構造物の表面に発生するひび割れの抽出を目的とし、学習データでは高解像度画像の占める割合が大きく、テストデータでは低解像度画像の占める割合が大きい場合を想定する。実験では、学習データ内の低解像度画像に対して重要度を高く推定することで、テストデータの環境に適応した学習を実現し、ひび割れ抽出精度の向上を示す。また、重要度推定における注目領域を可視化することで、重要度推定モデルがひび割れ領域ではなく背景領域の特徴に基づいて画素の違いを判別していることを確認する。

第 5 章では、本論文で扱った課題と各章での提案、実験結果について総括を行い、最後に今後の展望を述べる。

第 2 章

位置情報を用いた信頼性評価による肝硬変検出の高精度化

肝硬変は、線維化の異常な継続によって引き起こされ、肝がんのリスクが非常に高いため、早期発見が不可欠である。本章では、超音波画像を用いた肝硬変診断の支援システムを検討し、医師の経験に依存しない肝硬変検出を目的とする。特に、超音波画像上に自動的に設定される Region of Interest (ROI) 集合から信頼性の高い ROI を決定するために、信頼性評価法を提案する。信頼性の高い領域では肝硬変領域または健常領域が一定の広さで支配的に観察できると仮定し、事後確率と各 ROI の周辺 ROI との関係に基づいて信頼性評価法を設計する。実験では、提案手法によって信頼できる ROI の選択を可能とし、それらを用いた学習と分類による肝硬変検出の精度向上を示す。

2.1 はじめに

肝硬変は、線維化（結合組織の生成と沈着）の異常な継続によって引き起こされる肝疾患である [11]。肝硬変では肝がんのリスクが非常に高いため、早期発見が不可欠である。

肝硬変の診断には、肝生検、血液検査、線維化スキャン、超音波画像診断など、様々な方法がある。中でも、肝生検は肝臓組織を直接検査するため、肝硬変の確定的な診断法として知られている。肝生検の結果は信頼性が高い一方で、手術によって肝細胞を採取する必要があるため、患者への負担は大きくなる。一方、超音波画像は非侵襲性であるため、医師による肝硬変診断に広く用いられている。しかし、超音波画像では肝硬変の所見が不明瞭なため、診断は困難であり、医師の経験によって、診断結果が異なる可能性がある。

近年、画像処理と機械学習を用いたコンピュータ診断支援システムが開発され、「セカンドオピニオン」として医師が肝硬変を診断するのを支援している

[33]. 高い検出精度と検出プロセスの広範囲な自動化の両立が望まれるが、最近の研究では、前者を人間によるラベル付けで実現することに重点が置かれている。これは、特に超音波画像において、肝臓組織の中で明瞭で効果的な画像特徴を持つ関心領域を選択することが、高い検出精度を達成する上で不可欠だからである。例えば、一部のシステムでは医師が手動で設定した Region of Interest (ROI) を入力する。しかし、検出精度は ROI の品質に依存するため、医師の作業負荷が増大する。

自動設定された ROI を使用する場合、血管、影、診断に関係のない肝臓領域など、不要な ROI が含まれる可能性がある。一般的に、これらの ROI を分類器の学習や被験者の分類に使用すると、誤った分類結果につながる可能性が高い。そのため、信頼性の高い ROI を選択することは、この肝硬変検出精度向上において非常に重要となる。

本研究では、超音波画像上に自動的に設定される ROI 集合から信頼性の高い ROI を決定するために、各 ROI の信頼性を評価する信頼性評価法 (Reliability Evaluation Method, REM) を提案する。提案では、超音波ノイズの影響を受けた領域や影・血管領域と異なり、信頼性の高い領域では肝硬変領域または健常領域 (肝硬変ではない領域) が一定の広さで支配的に観察できると仮定した。この仮定に基づき、提案手法は事後確率と各 ROI の周辺 ROI との関係に基づいて設計される。各 ROI の事後確率は、学習済み分類器の出力として与えられ、特徴空間における ROI の信頼性を評価する。周辺 ROI との関係では、画像空間における各 ROI 間の距離を考慮する。

これまでも、超音波画像上に設定された ROI 集合から信頼性の高い ROI を推定、機械学習モデルの学習に使用する手法が提案されてきた。一方で、それらの手法では ROI の画像特徴と ROI が取得された被験者に付与されたラベルをもとに信頼性評価を行っており、画像空間上での ROI 同士の関係を考慮できていなかった。本章で提案する REM は、ROI 同士の関係を画像空間上での位置情報をもとに考慮する点が、大きな違いである。

提案手法の有効性を示す具体例を図2.1に示す。ここで、(a) と (b) は、自動的に設定されたすべての ROI の分類結果を示している。肝硬変クラスと健常者クラスに分類された ROI は、互いに重なり合って存在するため、結果の信頼性が低いことがわかる。反対に、(d) と (e) は、REM によって選択された信頼性の高い ROI を示している。肝硬変と正常に分類された ROI は、それぞれ独立して配置されているため、結果の信頼性が高くなっていることがわかる。

以下に本章における貢献を述べる。

- 肝硬変と正常領域を示す信頼性の高い ROI を選択する信頼性評価法 (Reliability Evaluation Method, REM) を提案する。
- REM を活用した肝硬変診断支援システムを設計し、検出精度の高精度化

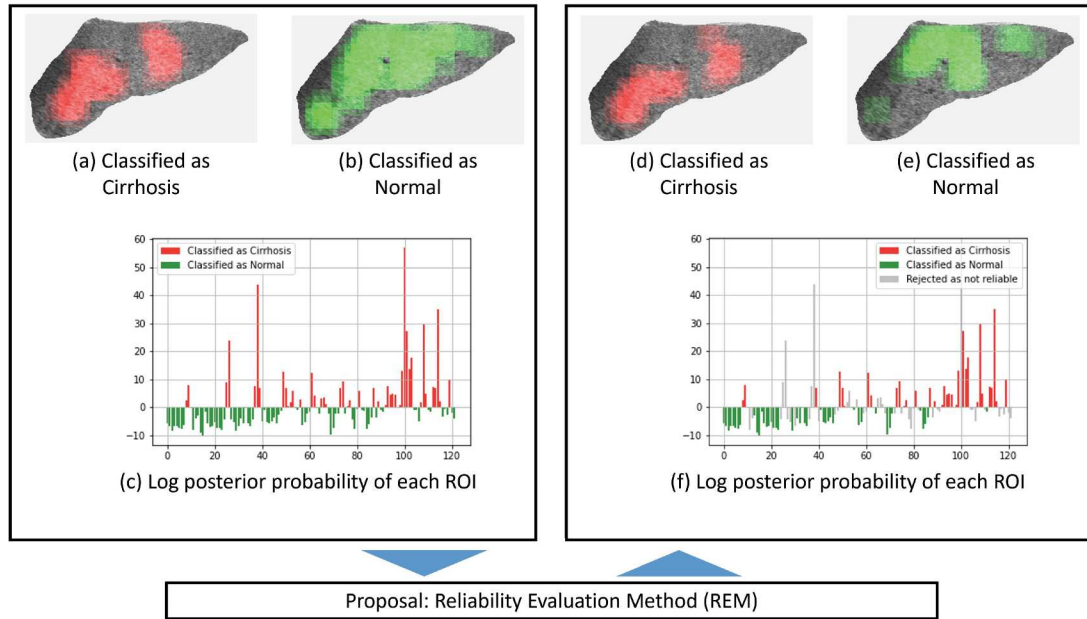


図 2.1: 肝硬変疾患クラスの被験者に対する予備実験での分類結果。テストデータの信頼できる ROI を獲得するために REM を適用した。(a), (b) は、それぞれ肝硬変クラスと健常者クラスに分類されたすべての ROI を示す。(d), (e) は、(a), (b) のすべての ROI のうち、信頼できると判断された ROI のみを示す。(c), (f) は、(a), (b), (d), (e) の対数事後確率をヒストグラムで示す。(f) において、信頼性がなく棄却された ROI は灰色で表示されている。プロダクトルール（積の法則）によって各 ROI の事後確率を統合することで、被験者単位の予測を行う。

と検出プロセスの自動化を実現する。

- 実用化に向けてパラメーターを最適化することで提案システムの有効性を検証する。

2.2 先行研究

2.2.1 手動で設定された ROI

先に述べたように、超音波画像の不鮮明さにより信頼できる ROI の選択は困難である。この不鮮明さを軽減するために、先行研究では医師が手動で設定した ROI をベースにした手法が提案されてきた。Gaber らは、有識者が定めた ROI を使用し、遺伝的アルゴリズム（Genetic Algorithm: GA）を使用してより良い特徴、ROI のサイズと量、閾値を決定する脂肪肝検出法を発表した [34]。また、CNN の学習で放射線科医が作成した注意機構に基づくアプローチが Schmauch らによって提案された [35]。Feng らは近年、CNN で空間ピラミッドプーリングによるピラミッド構造と、放射線科医が描画した ROI を使用したデータ拡張を使用した [36]。28 のテクスチャ特徴と加重投票を使用したアンサンブルモデルが Das らによって報告された [37]。これらの手法は高い精度を達成した一方、

医師が各 ROI にラベルを付ける作業が必要であった。つまり、分類器の学習と患者の分類において、システムへの人的介入が必要であり、このようなシステムの診断精度は医師の専門知識や経験値に依存する。実際、Vicas らは、医師間の ROI 設定のばらつきが分類結果に影響を与えることを実証している [38]。

2.2.2 自動で設定された ROI

手動設定 ROI だけでなく、自動設定 ROI を用いた研究もおこなわれてきた。Wu らは特徴選択に GA を使用し、SIFT 特徴点を用いた ROI 配置を提案した [39]。Acharya らは、形態学的オペレータ (morphological operator) を用いてコントラストを強調し、連結成分解析 (connected component analysis) によってバウンディングボックスを ROI として抽出した [40]。Zhang らは、超音波画像内の有効領域を強調するために、局所ヒストグラム適応均等化とアンシャープマスク画像による減算を使用し、ROI を定義する必要がない手法を提案した [41]。Minhas らでは、異なるスケールの連続ウェーブレット変換 (Continuous Wavelet Transform) と One-Class SVM によって得られた最大特徴値に基づいて ROI を定義する手法が提案された [42]。

我々の研究室はこれまで、Gabor 特徴量、二次分類器 [43]、Voting Rule を用いて各 ROI の事後確率を統合し、肝硬変検出 [44] を行う手法を提案してきた。さらに、画像処理による影および血管領域のマスクを適用することで、自動設定された ROI の信頼性を高め、Voting Rule ではなく Product Rule を適用することで事後確率を算出した [45, 46]。学習データにおいてより信頼性の高い ROI を効率的に使用するため、MIL Boost [47] と Self-training の適用を提案し、元の学習データよりも高い分類精度を実現した [48, 49, 50]。

これまでの研究では特徴量に基づいて、信頼性の高い ROI の選択に注力してきた。一方で、各 ROI 間の画像空間上での関係性は ROI の信頼性の尺度として使用されていなかった。さらに、信頼性の高い ROI を用いて分類器を単純に学習させるだけでは、テストデータにも信頼性の低い ROI が含まれており、正確な分類にはテストデータに対しても信頼性評価を行う必要がある。その際、学習済みモデルを用いた分類の段階において患者のラベルを使用せずにテストデータから信頼性の高い ROI を獲得し、個人レベルの出力の信頼性を向上させる必要がある。

2.3 提案手法：信頼性評価法 (REM)

画像認識では、各小領域は近傍と似た画像特徴量を持つため、特定の物体に関する情報は画像内の限定された範囲で一定の広がり支配的に観察される。超音波画像を用いた肝硬変検出の場合、肝硬変領域と正常領域はそれぞれ一定の

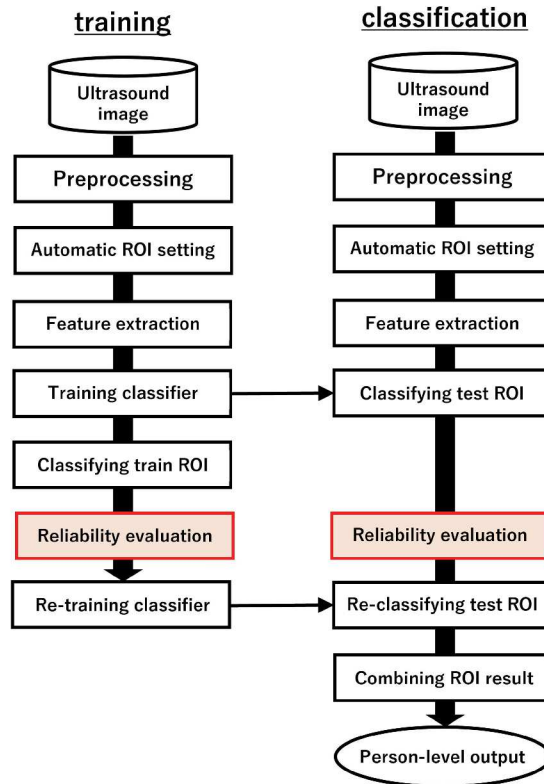


図 2.2: 提案システムのフロー.

広がりを持って分布する．一方で，超音波画像の不鮮明さの影響を受けた領域は肝硬変領域や健常領域とは異なり，不規則な見え方をする．これを活用するために，本章では自動設定された各 ROI の事後確率および位置情報を使用した信頼性評価法 (Reliability Evaluation Method, REM) を提案する．

2.3.1 提案システム

超音波画像を用いた肝硬変/健常者の 2 クラス分類問題を扱う．図2.2は提案するシステムの流れを示す．提案システムは特徴抽出フェーズと分類フェーズから構成される．特徴抽出フェーズでは，超音波画像内の肝臓領域に ROI を自動的に定義し，Gabor フィルタを用いて画像特徴量を抽出する．分類フェーズでは，二次分類器を学習し，テストデータの ROI を分類する．各 ROI の分類結果は，各クラスに対する事後確率として表される．最後に，被験者単位の出力を得るために，各 ROI の分類結果を Product Rule で統合する．これ以降，図2.2に示す提案システムで使用される各プロセスについて詳しく説明する．

2.3.2 画像前処理

肝臓領域における超音波ノイズや血管などの不要な領域の影響を抑制するために前処理が行われる [44]. 前処理では, 以下のように定義される各ブロックの画素値の平均値と標準偏差を用いる.

$$\mu = \frac{1}{X \times Y} \sum_{j=0}^{Y-1} \sum_{i=0}^{X-1} f(i;j) \quad (2.1)$$

$$\sigma = \sqrt{\frac{1}{X \times Y} \sum_{j=0}^{Y-1} \sum_{i=0}^{X-1} (f(i;j) - \mu)^2} \quad (2.2)$$

ここで, X と Y はブロックの高さと幅を表し, $f(i;j)$ は座標 $(i;j)$ におけるピクセル値を表す. が閾値より低い場合, ブロックは影領域としてマスクされる. が閾値より高い場合, ブロックは血管としてマスクされる.

2.3.3 自動 ROI 設定

ROI はラスタスキャン [44] によって重なりを許しながら自動的に定義される. この際, マスクされた領域を含む ROI は拒否される.

2.3.4 Gabor 特徴量

異なる周波数と方向を持つ Gabor フィルタのセットは, 分類やセグメンテーションなどの様々な分野で画像特徴を抽出するために使用される. サンプリング点を原点 $(x;y) = (0;0)$ に設定すると, フィルタは次のように定義される.

$$G(x;y;k_x;k_y,\sigma^2) = \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\} \exp\{i(k_x x + k_y y)\} \quad (2.3)$$

ここで, x と y は 2 次元空間の座標を表し, k_x と k_y は周波数座標を表す. フィルタの周波数と方向は k_x と k_y によって設定される. σ はガウス包絡関数 (Gaussian envelope) の標準偏差である.

2.3.5 Quadratic Discriminant Analysis

肝臓超音波画像の ROI から抽出した Gabor 特徴を用いて, ROI を分類する二次分類器 (Quadratic Discriminant Analysis, QDA)[43] を学習する. 二次分類器は特徴空間内で各クラスの特徴量が正規分布するという仮定をおいており, 分

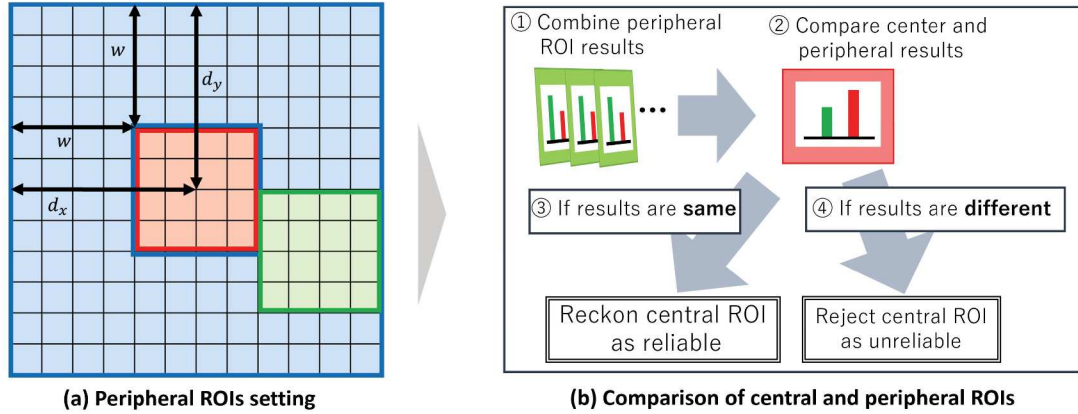


図 2.3: ROI 信頼性評価プロセスフロー. (a) 周辺 ROI/領域の定義. (b) 中心 ROI と周辺 ROI の比較. 左側の図は周辺 ROI で構成される周辺領域を示す. 中央の赤色で示される領域は注目 ROI を示す. 緑色で示される領域は周辺 ROI の一例を示す. 青で塗りつぶされた領域は, 周辺 ROI となりうるすべての領域を示す. 右側の図は中心 ROI と周辺 ROI の比較プロセスを示す. 各注目 ROI に対して, 周辺領域はパラメーター dy , dx および w で定義される. 周辺領域に存在する周辺 ROI の事後確率を Product Rule で統合した後, 提案手法は中心の注目している ROI と周辺 ROI の分類結果を比較し, 分類結果が一致する場合にのみ中心 ROI を信頼できる ROI と判断する.

類結果は超音波画像の不鮮明さの影響を受けにくくなる. 対数事後確率は以下のように記述される.

$$f_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)\Sigma^{-1}(\mathbf{x} - \mu_i)^T - \frac{1}{2}\log|\Sigma_i| + \text{Cst} \quad (2.4)$$

ここで, μ_i と Σ_i はクラス k の標本平均ベクトルと共分散行列である. Cst は各クラスに共通する定数値である.

2.3.6 Product Rule

一人の被験者から得られた複数の ROI を用いて, Product Rule[45, 46] により肝硬変クラスと健常者クラスの分類を行う. Product Rule は以下のように記述される.

$$P(k|\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_n) = \prod_j P(k|\mathbf{x}_j): \quad (2.5)$$

分類器の出力は対数事後確率に変換されるため, Product Rule においては合計で算出, 比較を行い, 被験者のクラスを決定するために使用される.

2.3.7 信頼性評価法

図2.3は, 各 ROI が信頼できるか信頼できないかを決定するプロセスを示しており, そのアルゴリズムは Algorithm 1で説明する. N 個の ROI のセットを持

Algorithm 1: Reliability evaluation method.

Input : an ultrasound image I with N ROIs; a trained classifier f ;
peripheral parameters of cirrhosis cir and normal nor ; a width w .
Output: a set of reliable ROIs S .

for $i \leftarrow 1$ to N do
 Select an ROI R_i from I (where $i \in [1; N]$ indexes the ROI locations).
 Obtain a classification result $y^0 \in \{cirrhosis; normal\}$ of R_i using f .
 Set the peripheral parameters $d_y, d_x \in \{cir; nor\}$ to be the same as y^0 .
 Select peripheral ROIs based on the coordinates of R_i within d_y, d_x pixels
 around R_i , excluding the inner region of width w .
 Obtain the classification results of each peripheral ROI using f .
 Aggregate the peripheral results by the Product Rule to get
 $y \in \{cirrhosis; normal\}$.
 if $y = y^0$ then
 Append R_i to S .
 else
 Reject R_i .

つ超音波画像 I , 学習済みの分類器 f , 肝硬変 cir と健常者 nor の周辺パラメーター, および周辺領域の幅 w を与えると, REM は信頼できる ROI のセット S を出力する. 各 ROI の信頼性は, 関心のある ROI の事後確率と, 周辺領域の周辺 ROI を使用して Product Rule で統合したものを比較することで評価される. 両方の分類結果が一致する場合, 関心のある ROI を信頼できると判断する. これらの分類は, Gabor 特徴と二次分類器によって行う.

REM は適用対象となる被験者から得られる全ての ROI に適用し, 信頼性の高い ROI 集合 S のみを後続の処理に用いる.

REM は, 信頼できる肝硬変領域と健常領域が一定の広さで分布していると仮定している. そのため, 実際の肝硬変領域や健常領域が狭く分布している被験者への適用は困難となる可能性がある.

2.3.8 ラベル付与方法

被験者単位のラベルは, 医師の臨床診断に基づいて手動で付与される. 学習データ内の各 ROI のラベルは, その ROI が得られた被験者のラベルに基づいて付与される. そのため, 肝硬変クラスの学習データの場合, 肝硬変患者の健康な領域を示す ROI が肝硬変クラスとしてラベル付けされることがある. 同様に, 健常者クラスの学習データの場合, 超音波ノイズによって健常者の肝臓にみられる肝硬変に似た特徴を示す ROI が健常者クラスとしてラベル付けされることがある. この処理により, 自動的にラベル付けされた ROI を用いて分類器を学習することができ, 医師によるラベル付けの負担を軽減できる一方で, ROI に

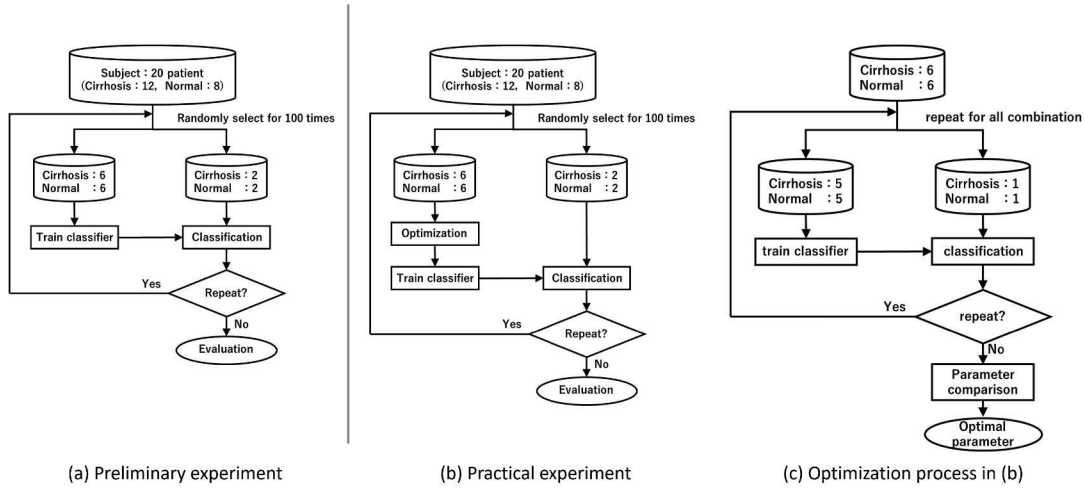


図 2.4: 予備実験と実践的実験のフロー. (a) 予備実験. (b) 実践的実験. (c) (b) での最適化プロセス. (c) では REM のパラメーターを最適化し、最適化した値を (b) で活用する.

付与される各ラベルの信頼性は低下する.

2.4 実験

この節では、予備実験と実践的実験について説明する. 両実験の流れを図2.4に示す. 予備実験は REM による高信頼 ROI 抽出が、モデル分類精度の向上に有効であることを検証するために行う. 実践的実験は、REM での周辺領域パラメーターの自動最適化を含む、提案システムの実用性を評価するために実施する.

2.4.1 データセットとパラメーター

実験では、付録で述べたデータセットを使用した. 前処理で用いるパラメーターは、先行研究 [45, 46] に基づき決定した. ブロックサイズを 4×4 ピクセルに設定し、 μ の閾値を 25, σ を 65 に設定した. 各 ROI は 32×32 ピクセルで 24 ピクセルが重なり合うように自動的に定義される. Gabor 特徴については、 $k_y = 0$ のときは $\sigma = 50$, $k_x = \{0; 1; 2; 3\}$, $k_y = \{1; 2; 3\}$ のときは $k_x = \{-3; \dots; 3\}$ を使用して Gabor 特徴を抽出した.

REM については、周辺領域のパラメーター $dy; dx$ の選択候補として以下に示す 2 パターンを用意した.

(1) $cir; nor \in \{24; 32; 40; 48; 56; 64; 72; 80\}$

(2) $cir; nor \in \{24; 32; 40; 48\}$

(1) について、本章で使用するデータセットにおいて肝臓領域の大きさ (縦幅) の平均が約 160 画素であるため、周辺領域の外径の最大値が 160 を超えない範

囲として、 $dy; dx = 80$ を最大値としている。(2) について、最も広い周辺領域の設定でも、中心 ROI と周辺領域が接する広さに限定している。周辺領域のパラメーター (1) は (2) の候補を包含しているため、予備実験では (1) を使用し、各パラメーターにおける分類性能への影響を分析する。一方で実践的実験では周辺領域のパラメーター (1), (2) を両方使用し、パラメーター候補の多さが分類性能に与える影響を分析する。周辺領域の幅 w は、各周辺パラメーターにおける分類性能の詳細な比較を行うために、1 つの ROI が一列に並ぶように 32 ピクセルに設定した。

本実験に使用した画像の枚数が再現性の観点で不十分である可能性があることは留意すべきである。一方で、可能な限り最大枚数の ROI を収集し、画像特徴の周波数を固定し、特徴空間において正規分布を仮定することで分器器の過学習を防ぎ、実験の再現性を向上させている。

2.4.2 評価指標

評価指標として、第 1 章で定義した感度 (Sensitivity, または Recall), 特異度 (Specificity), 識別率 (Accuracy), 適合率 (Precision), F 値, Area Under the Curve (AUC)[32] を使用した。

2.5 結果と考察

2.5.1 予備実験

予備実験では、以下の 4 つのケースで評価指標による比較を行った：

ベースライン：学習と分類の両方で REM を使用しない

- (i)：分類に REM を使用した提案システム
- (ii)：学習に REM を使用した提案システム
- (iii)：学習と分類の両方で REM を使用した提案システム（訓練時とテスト時で共通のパラメーター）
- (iv)：学習と分類の両方で REM を使用した提案システム（訓練時とテスト時で個別のパラメーター）

ケース (iv) では周辺領域のパラメーター候補を (1) $cir; nor \in \{24; 32; 40; 48; 56; 64; 72; 80\}$ とすると、各クラスにおいて訓練時とテスト時で $8 \times 8 = 64$ 通り、2 クラスを通して $64 \times 64 = 4096$ 通りの候補があり実験時間が膨大になってしまう。そのため、周辺領域を (2) $cir, nor \in \{24; 32; 40; 48\}$

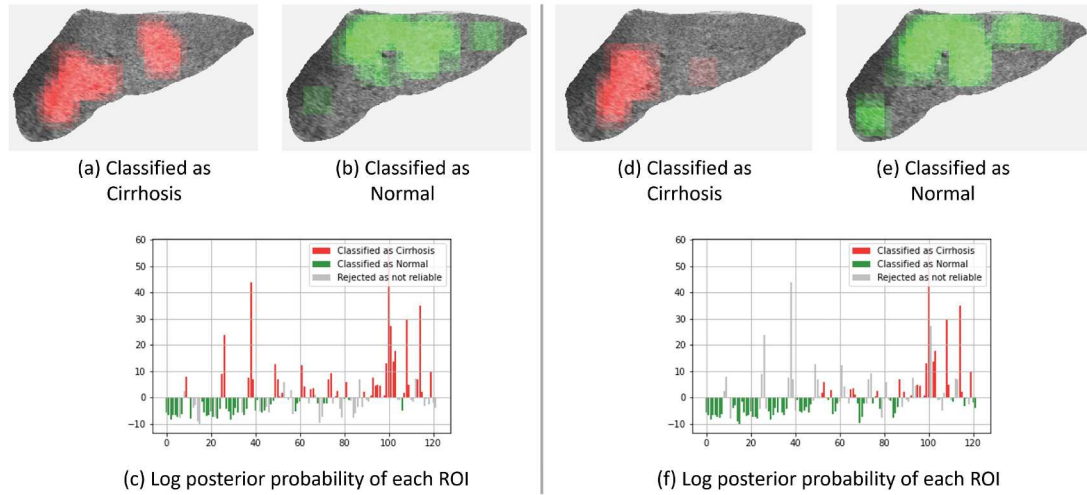


図 2.5: 信頼性評価法を用いた予備実験における, テストデータ中の肝硬変患者の分類結果例. (a),(b),(c) は, 周辺領域パラメーター $\text{cir}(\text{dy} = \text{dx} = 24); \text{nor}(\text{dy} = \text{dx} = 48)$ を用いた分類結果を示す. (d),(e),(f) は, 周辺領域パラメーター $\text{cir}(\text{dy} = \text{dx} = 48); \text{nor}(\text{dy} = \text{dx} = 24)$ を用いた分類結果を示す. 被験者単位の予測を得るために, 各 ROI で得られた事後確率を Product Rule で統合した. (a) 肝硬変クラスと分類された信頼できる ROI. (b) 健常者クラスと分類された信頼できる ROI (c) REM によって信頼できると判断された ROI の各対数事後確率. (d) 肝硬変クラスと分類された信頼できる ROI. (e) 健常者クラスと分類された信頼できる ROI. (f) REM によって信頼できると判断された ROI の各対数事後確率.

に限定し, 合計で $4(\text{訓練時}) \times 4(\text{テスト時}) \times 2(\text{クラス数}) = 256$ 通りでの実験を行った. 上記のすべてのケースで, REM が使用されなかった学習およびテストデータでは, 自動設定される ROI すべてを分類器の学習と分類に使用した.

1 回の試行において, 学習データには各クラス 6 名, テストデータには 2 名の被験者を使用した. 学習データとテストデータの選択はランダムに 100 回繰り返され, 各試行において各周辺パラメーターの分類性能を評価した.

表 2.1(a) は, REM の各パラメーターの中で最も高い精度を記録したパラメーターでの分類性能を各ケース (i) ~ (iv) について示す. REM を用いた提案システムの精度 (i); (ii); (iii); (iv) はいずれもベースライン (REM を用いない) よりも高く, 超音波画像を用いた肝硬変検出において REM の有効性を示している. 特に, 訓練時とテスト時で REM を適用する提案システム (iii) よりも, 提案システム (ii) の方が識別率が高くなっている. これは, 提案システム (iii) において訓練時とテスト時で共通の周辺領域パラメーターを使用しているためであり, 個別のパラメーターを使用した提案システム (iv) がすべてのケースの中で最も識別率が高くなっていることから, 訓練時とテスト時の両方に REM を適用することの有効性を確認した.

表 2.1: 予備実験の結果.

(a): ベースラインと各ケース (i)~(iii) の最も良い精度の比較.

Proposed	Reliability evaluation method	Sensitivity	Specificity	Accuracy
Baseline	Not used	0.9000	0.7900	0.8450
(i)	Used on test data	0.8650	0.9650	0.9150
(ii)	Used on training data	0.8800	0.9700	0.9250
(iii)	Used on training and test data	0.9700	0.8700	0.9200
(vi)	Used on training and test data	0.8950	0.9700	0.9325

(b): 提案 (i): REM をテスト時に適用した際の Accuracy.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	0.8375	0.7925	0.8000	0.7925	0.8025	0.7675	0.7675	0.7875
	32	0.9025	0.9000	0.9050	0.9050	0.9075	0.8900	0.8825	0.8750
	40	0.8725	0.8925	0.9100	0.9125	0.9050	0.8975	0.9050	0.9000
	48	0.8350	0.8775	0.9075	0.9150	0.9100	0.9000	0.9075	0.8750
	56	0.8375	0.8725	0.8850	0.9000	0.9050	0.9000	0.8875	0.8550
	64	0.8475	0.8700	0.8900	0.9000	0.8900	0.8750	0.8750	0.8700
	72	0.8575	0.8650	0.8750	0.8875	0.8800	0.8550	0.8525	0.8175
	80	0.8450	0.8850	0.8950	0.8975	0.8950	0.8750	0.8750	0.8775

(c): 提案 (i): REM をテスト時に適用した際の Sensitivity.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	0.8950	0.9350	0.9800	0.9700	0.9550	0.9300	0.9400	0.9700
	32	0.8600	0.9000	0.9350	0.9400	0.9250	0.9100	0.9150	0.9600
	40	0.7700	0.8300	0.8750	0.8750	0.8850	0.8850	0.8900	0.8900
	48	0.6950	0.7850	0.8450	0.8650	0.8500	0.8400	0.8550	0.8450
	56	0.7000	0.7750	0.8150	0.8400	0.8350	0.8450	0.8500	0.8400
	64	0.7300	0.7950	0.8450	0.8650	0.8500	0.8500	0.8500	0.8600
	72	0.7300	0.8000	0.8400	0.8500	0.8350	0.8250	0.8300	0.8450
	80	0.6900	0.7800	0.8050	0.8100	0.8000	0.7950	0.8050	0.8200

表 2.1: 予備実験の結果 (続き).

(d): 提案 (i): REM をテスト時に適用した際の Specificity.

		nor							
$dy; dx$ ($dy = dx$)		24	32	40	48	56	64	72	80
cir	24	0.7800	0.6500	0.6200	0.6150	0.6500	0.6050	0.5950	0.6050
	32	0.9450	0.9000	0.8750	0.8700	0.8900	0.8700	0.8500	0.7900
	40	0.9750	0.9550	0.9450	0.9500	0.9250	0.9100	0.9200	0.9100
	48	0.9750	0.9700	0.9700	0.9650	0.9700	0.9600	0.9600	0.9050
	56	0.9750	0.9700	0.9550	0.9600	0.9750	0.9550	0.9250	0.8700
	64	0.9650	0.9450	0.9350	0.9350	0.9300	0.9000	0.9000	0.8800
	72	0.9850	0.9300	0.9100	0.9250	0.9250	0.8850	0.8750	0.7900
	80	1.0000	0.9900	0.9850	0.9850	0.9900	0.9550	0.9450	0.9350

(e): 提案 (ii): REM を学習時に適用した際の Accuracy.

		nor							
$dy; dx$ ($dy = dx$)		24	32	40	48	56	64	72	80
cir	24	0.8200	0.8850	0.9100	0.9225	0.9225	0.9150	0.9200	0.9250
	32	0.7475	0.7825	0.7975	0.8000	0.7925	0.8000	0.8100	0.8150
	40	0.7200	0.7400	0.7475	0.7600	0.7675	0.7525	0.7625	0.7700
	48	0.7350	0.7525	0.7825	0.7825	0.7825	0.7850	0.7800	0.7800
	56	0.7375	0.7600	0.7800	0.7850	0.7850	0.7850	0.7900	0.7850
	64	0.7425	0.7600	0.7850	0.7850	0.7825	0.7875	0.7825	0.7950
	72	0.7325	0.7475	0.7650	0.7800	0.7800	0.7850	0.7875	0.7900
	80	0.7150	0.7375	0.7450	0.7525	0.7575	0.7525	0.7550	0.7650

(f): 提案 (ii): REM を学習時に適用した際の Sensitivity.

		nor							
$dy; dx$ ($dy = dx$)		24	32	40	48	56	64	72	80
cir	24	0.9300	0.8950	0.8950	0.8950	0.8950	0.8800	0.8850	0.8800
	32	1.0000	0.9900	0.9650	0.9500	0.9550	0.9600	0.9550	0.9350
	40	1.0000	1.0000	1.0000	0.9950	1.0000	0.9950	0.9950	0.9800
	48	1.0000	0.9950	0.9900	0.9850	0.9850	0.9850	0.9750	0.9700
	56	1.0000	0.9950	0.9950	0.9850	0.9850	0.9850	0.9850	0.9700
	64	1.0000	0.9950	0.9950	0.9900	0.9850	0.9850	0.9800	0.9700
	72	1.0000	0.9950	0.9950	0.9950	0.9950	0.9950	0.9950	0.9800
	80	1.0000	1.0000	0.9950	0.9950	0.9950	0.9950	0.9950	0.9950

表 2.1: 予備実験の結果 (続き).

(g): 提案 (ii): REM を学習時に適用した際の Specificity.

		nor							
	$\frac{dy}{dx}$ ($\frac{dy}{dx} = \frac{dx}{dx}$)	24	32	40	48	56	64	72	80
cir	24	0.7100	0.8750	0.9250	0.9500	0.9500	0.9500	0.9550	0.9700
	32	0.4950	0.5750	0.6300	0.6500	0.6300	0.6400	0.6650	0.6950
	40	0.4400	0.4800	0.4950	0.5250	0.5350	0.5100	0.5300	0.5600
	48	0.4700	0.5100	0.5750	0.5800	0.5800	0.5850	0.5850	0.5900
	56	0.4750	0.5250	0.5650	0.5850	0.5850	0.5850	0.5950	0.6000
	64	0.4850	0.5250	0.5750	0.5800	0.5800	0.5900	0.5850	0.6200
	72	0.4650	0.5000	0.5350	0.5650	0.5650	0.5750	0.5800	0.6000
	80	0.4300	0.4750	0.4950	0.5100	0.5200	0.5100	0.5150	0.5350

(h): 提案 (iii): REM を学習時とテスト時に適用した際の Accuracy.

		nor							
	$\frac{dy}{dx}$ ($\frac{dy}{dx} = \frac{dx}{dx}$)	24	32	40	48	56	64	72	80
cir	24	0.8200	0.8075	0.8100	0.8075	0.8325	0.7875	0.7725	0.7775
	32	0.9150	0.9050	0.9050	0.9150	0.9200	0.8625	0.8725	0.8600
	40	0.8800	0.8875	0.8975	0.9075	0.9000	0.8825	0.8850	0.8775
	48	0.9025	0.8900	0.8875	0.9000	0.9200	0.9000	0.8875	0.8575
	56	0.8825	0.8725	0.8725	0.8800	0.9050	0.8775	0.8725	0.8625
	64	0.8350	0.8500	0.8500	0.8600	0.8675	0.8575	0.8600	0.8425
	72	0.8600	0.8450	0.8400	0.8475	0.8600	0.8425	0.8300	0.8175
	80	0.8475	0.8500	0.8650	0.8725	0.8625	0.8550	0.8575	0.8550

表 2.1: 予備実験の結果 (続き).

(i): 提案 (iii): REM を学習時とテスト時に適用した際の Sensitivity.

		nor							
	$\frac{dy}{dx}$ ($\frac{dy}{dx} = \frac{dx}{dx}$)	24	32	40	48	56	64	72	80
cir	24	0.9150	0.9300	0.9500	0.9400	0.9250	0.9100	0.9100	0.9250
	32	0.9550	0.9650	0.9700	0.9750	0.9700	0.9450	0.9500	0.9700
	40	0.9000	0.9000	0.9000	0.9000	0.9050	0.9050	0.9050	0.9050
	48	0.8600	0.8700	0.8850	0.8900	0.8950	0.9000	0.9000	0.9050
	56	0.8400	0.8600	0.8750	0.8750	0.8750	0.8800	0.8800	0.9000
	64	0.8500	0.8700	0.8750	0.8800	0.8900	0.8900	0.8950	0.9050
	72	0.8750	0.8850	0.9000	0.9000	0.8850	0.8900	0.8950	0.8950
	80	0.8400	0.8450	0.8700	0.8800	0.8650	0.8550	0.8600	0.8500

(j): 提案 (iii): REM を学習時とテスト時に適用した際の Specificity.

		nor							
	$\frac{dy}{dx}$ ($\frac{dy}{dx} = \frac{dx}{dx}$)	24	32	40	48	56	64	72	80
cir	24	0.7250	0.6850	0.6700	0.6750	0.7400	0.6650	0.6350	0.6300
	32	0.8750	0.8450	0.8400	0.8550	0.8700	0.7800	0.7950	0.7500
	40	0.8600	0.8750	0.8950	0.9150	0.8950	0.8600	0.8650	0.8500
	48	0.9450	0.9100	0.8900	0.9100	0.9450	0.9000	0.8750	0.8100
	56	0.9250	0.8850	0.8700	0.8850	0.9350	0.8750	0.8650	0.8250
	64	0.8200	0.8300	0.8250	0.8400	0.8450	0.8250	0.8250	0.7800
	72	0.8450	0.8050	0.7800	0.7950	0.8350	0.7950	0.7650	0.7400
	80	0.8550	0.8550	0.8600	0.8650	0.8600	0.8550	0.8550	0.8600

表 2.2: 提案 (iv): REM を学習時とテスト時に適用した際の Accuracy.

		nor (train)											
		8				16				24			
		nor (test)				nor (test)				nor (test)			
cir (train)	cir (test)	8	16	24	32	8	16	24	32	8	16	24	32
8	8	0.8200	0.7850	0.7925	0.7875	0.8750	0.8075	0.8125	0.8025	0.8975	0.8175	0.8100	0.8075
	16	0.9150	0.9050	0.9075	0.9025	0.9100	0.9150	0.9300	0.9225	0.8975	0.9125	0.9275	0.9275
	24	0.8850	0.9050	0.9150	0.9200	0.8625	0.8975	0.9225	0.9250	0.8450	0.8975	0.9175	0.9275
	32	0.8550	0.8825	0.9025	0.9150	0.8225	0.8650	0.8950	0.9025	0.8050	0.8550	0.8875	0.8950
16	8	0.7675	0.7500	0.7475	0.7500	0.7850	0.7650	0.7650	0.7625	0.7850	0.7825	0.7775	0.7750
	16	0.9150	0.8450	0.8275	0.8300	0.9200	0.9050	0.8800	0.8800	0.9250	0.9075	0.9050	0.9150
	24	0.9175	0.8750	0.8675	0.8625	0.9175	0.9175	0.9100	0.9125	0.9100	0.9300	0.9225	0.9225
	32	0.8950	0.8775	0.8550	0.8650	0.8825	0.8975	0.8925	0.8975	0.8775	0.9025	0.9100	0.9175
24	8	0.7375	0.7325	0.7325	0.7325	0.7525	0.7375	0.7375	0.7375	0.7800	0.7525	0.7450	0.7450
	16	0.8775	0.7950	0.7875	0.7875	0.9200	0.8625	0.8350	0.8375	0.9225	0.8850	0.8525	0.8650
	24	0.8800	0.8525	0.8425	0.8425	0.9200	0.8875	0.8875	0.8825	0.9300	0.9000	0.8975	0.9000
	32	0.9050	0.8425	0.8325	0.8350	0.9025	0.8900	0.8625	0.8700	0.8950	0.9050	0.8900	0.9000
32	8	0.7475	0.7400	0.7400	0.7400	0.7750	0.7600	0.7550	0.7525	0.7750	0.7675	0.7650	0.7650
	16	0.8950	0.8125	0.8075	0.8125	0.9025	0.8625	0.8375	0.8400	0.9125	0.8875	0.8625	0.8625
	24	0.8750	0.8625	0.8475	0.8500	0.9175	0.8875	0.8850	0.8800	0.9225	0.8975	0.8925	0.8950
	32	0.9025	0.8475	0.8400	0.8375	0.8950	0.8900	0.8650	0.8725	0.8925	0.9000	0.8875	0.8975

表 2.3: 提案 (iv): REM を学習時とテスト時に適用した際の Sensitivity.

		nor (train)											
		8				16				24			
		nor (test)				nor (test)				nor (test)			
cir (train)	cir (test)	8	16	24	32	8	16	24	32	8	16	24	32
8	8	0.9150	0.9600	0.9950	0.9950	0.8950	0.9300	0.9750	0.9700	0.8950	0.9100	0.9500	0.9450
	16	0.8750	0.9200	0.9600	0.9500	0.8400	0.8800	0.9250	0.9150	0.8150	0.8700	0.9050	0.8950
	24	0.7950	0.8500	0.8750	0.8900	0.7450	0.8200	0.8700	0.8750	0.7100	0.8150	0.8550	0.8750
	32	0.7350	0.7900	0.8500	0.8700	0.6650	0.7550	0.8150	0.8300	0.6300	0.7350	0.8000	0.8150
16	8	0.9850	1.0000	1.0000	1.0000	0.9400	0.9950	1.0000	1.0000	0.9200	0.9900	1.0000	1.0000
	16	0.9550	0.9800	0.9850	0.9850	0.9000	0.9650	0.9700	0.9750	0.8900	0.9550	0.9700	0.9750
	24	0.8950	0.8950	0.8950	0.9000	0.8600	0.8950	0.8950	0.8950	0.8450	0.8950	0.8950	0.8950
	32	0.8300	0.8850	0.8850	0.8950	0.7900	0.8500	0.8750	0.8750	0.7800	0.8400	0.8750	0.8750
24	8	1.0000	1.0000	1.0000	1.0000	0.9850	1.0000	1.0000	1.0000	0.9650	1.0000	1.0000	1.0000
	16	0.9800	0.9850	1.0000	1.0000	0.9550	0.9800	0.9900	0.9900	0.9350	0.9800	0.9850	0.9850
	24	0.9000	0.9050	0.9050	0.9050	0.8950	0.9000	0.9000	0.9000	0.8950	0.9000	0.9000	0.9000
	32	0.8800	0.8950	0.8950	0.9000	0.8400	0.8900	0.8950	0.8950	0.8200	0.8900	0.8950	0.8950
32	8	0.9900	1.0000	1.0000	1.0000	0.9650	1.0000	1.0000	1.0000	0.9450	0.9950	1.0000	1.0000
	16	0.9800	0.9800	1.0000	1.0000	0.9350	0.9750	0.9800	0.9750	0.9200	0.9750	0.9750	0.9750
	24	0.9000	0.9050	0.9050	0.9050	0.8950	0.9000	0.8950	0.9000	0.8800	0.9000	0.8950	0.9000
	32	0.8600	0.8950	0.8900	0.8950	0.8300	0.8700	0.8850	0.8900	0.8150	0.8650	0.8850	0.8900

表 2.4: 提案 (iv): REM を学習時とテスト時に適用した際の Specificity.

		nor (train)											
		8				16				24			
		nor (test)				nor (test)				nor (test)			
cir (train)	cir (test)	8	16	24	32	8	16	24	32	8	16	24	32
8	8	0.7250	0.6100	0.5900	0.5800	0.8550	0.6850	0.6500	0.6350	0.9000	0.7250	0.6700	0.6700
	16	0.9550	0.8900	0.8550	0.8550	0.9800	0.9500	0.9350	0.9300	0.9800	0.9550	0.9500	0.9500
	24	0.9750	0.9600	0.9550	0.9500	0.9800	0.9750	0.9750	0.9750	0.9800	0.9800	0.9800	0.9800
	32	0.9750	0.9750	0.9550	0.9600	0.9800	0.9750	0.9750	0.9750	0.9800	0.9750	0.9750	0.9750
16	8	0.5500	0.5000	0.4950	0.5000	0.6300	0.5350	0.5300	0.5250	0.6500	0.5750	0.5550	0.5450
	16	0.8750	0.7100	0.6700	0.6750	0.9400	0.8450	0.7900	0.7850	0.9600	0.8600	0.8400	0.8350
	24	0.9400	0.8550	0.8400	0.8250	0.9750	0.9400	0.9250	0.9300	0.9750	0.9650	0.9500	0.9450
	32	0.9600	0.8700	0.8250	0.8350	0.9750	0.9450	0.9100	0.9200	0.9750	0.9650	0.9450	0.9450
24	8	0.4750	0.4650	0.4650	0.4650	0.5200	0.4750	0.4750	0.4750	0.5950	0.5050	0.4900	0.4900
	16	0.7750	0.6050	0.5750	0.5750	0.8850	0.7450	0.6800	0.6850	0.9100	0.7900	0.7250	0.7450
	24	0.8600	0.8000	0.7800	0.7800	0.9450	0.8750	0.8750	0.8650	0.9650	0.9000	0.8950	0.9000
	32	0.9300	0.7900	0.7700	0.7700	0.9650	0.8900	0.8300	0.8450	0.9700	0.9200	0.8900	0.9050
32	8	0.5050	0.4800	0.4800	0.4800	0.5850	0.5200	0.5100	0.5050	0.6050	0.5400	0.5300	0.5250
	16	0.8100	0.6450	0.6150	0.6250	0.8700	0.7500	0.6950	0.7050	0.9050	0.8000	0.7500	0.7500
	24	0.8500	0.8200	0.7900	0.7950	0.9400	0.8750	0.8750	0.8600	0.9650	0.8950	0.8900	0.8900
	32	0.9450	0.8000	0.7900	0.7800	0.9600	0.9100	0.8450	0.8550	0.9700	0.9350	0.9000	0.9050

表 2.5: ROI の枚数比較. REM は各 ROI のラベルでなく, 各 ROI の分類されたクラスに応じて, 周辺領域のパラメーターを適用する.

(a): REM on test data.

Labeled class	Classified class	# of ROI w/ o REM	# of ROI w/ REM cir: 48 nor: 24	# of ROI w/ REM cir: 24 nor: 48
Cirrhosis	Cirrhosis	28,054	17,758 (63.3%)	23,638 (84.3%)
	Normal	38,413	28,028 (73.0%)	17,291 (45.0%)
Normal	Cirrhosis	12,960	4,022 (31.0%)	8,155 (62.9%)
	Normal	42,802	37,994 (88.8%)	32,778 (76.6%)

(b): REM on training data.

Labeled class	Classified class	# of ROI w/ o REM	# of ROI w/ REM cir: 48 nor: 24	# of ROI w/ REM cir: 24 nor: 48
Cirrhosis	Cirrhosis	92,772	65,155 (70.2%)	81,059 (87.4%)
	Normal	105,606	73,547 (69.6%)	40,014 (37.9%)
Normal	Cirrhosis	29,787	4,098 (13.8%)	14,714 (49.4%)
	Normal	148,951	137,716 (92.5%)	126,662 (85.0%)

提案システム (i)

具体的には、REM を用いた各ケースの識別率、感度、特異度を表2.1(b) ~ 2.1(j), 表2.2 ~ 2.4に示す. 表2.1(c), 2.1(d) (REM: テスト時) では、パラメーターが小さいクラスでは各クラスへの分類性能が向上し、パラメーターが大きいクラスでは各クラスへの分類性能が低下している.

提案システム (ii)

逆に、表2.1(f), 2.1(g) (REM: 学習時) では、クラスのパラメーターが小さいほど各クラスへの分類性能が低下し、クラスのパラメーターが大きいほど各クラスへの分類性能が向上した. これは、表2.1(c), 2.1(d) (REM: テスト時) とは逆の結果である.

提案システム (i) と比較して提案システム (ii) では感度と特異度の差が大きい. これは、提案システム (i) では REM の対象がテストデータの各クラス 2 名に限定され、QDA で仮定する各クラスの正規分布は変化しないのに対して、提案システム (ii) では、REM の対象が学習データの各クラス 6 名であり、QDA で仮定する各クラスの正規分布が高信頼 ROI の選択によって変化したため、分類性能に大きく影響が出たことが原因であると考ええる.

提案システム (iii)

さらに、表2.1(i), 2.1(j) (REM: 学習時とテスト時に共通のパラメーター) では、各クラスのパラメーターが小さいほど分類性能が向上し、パラメーターが大きいほど分類性能が低下している. これは表2.1(c), 2.1(d) (REM: テスト時) と同じ傾向である. また、表2.1(c), 2.1(d) (REM: 学習時) における分類能力の低下は表2.1(i), 2.1(j) (REM: 学習時とテスト時) では抑制されており、訓練時にも REM を適用することが有効であると分かる.

提案システム (iv)

さらに、表2.3, 2.4(REM: 学習時とテスト時に個別のパラメーター) では、提案システム (i)(REM: テスト時) と提案システム (ii)(REM: 学習時) の両方の傾向が確認できる. まず、訓練時の各パラメーターの設定 $\text{cir}(\text{train})$, $\text{nor}(\text{train}) \in \{24; 32; 40; 48\}$ に注目するとテスト時に設定する各クラスのパラメーター $\text{cir}(\text{test})$, $\text{nor}(\text{test}) \in \{24; 32; 40; 48\}$ が小さいほど分類性能が向上し、パラメーターが大きいほど分類性能が低下している. これは表2.1(c), 2.1(d) (REM: テスト時) と同じ傾向である. また、テスト時の各パラメーターの設定 $\text{cir}(\text{testtest})$, $\text{nor}(\text{testtest}) \in \{24; 32; 40; 48\}$ に注目すると学習時に設定す

る各クラスのパラメーター $\text{cir}(\text{train})$, $\text{nor}(\text{train}) \in \{24; 32; 40; 48\}$ が大きいほど分類性能が向上し、パラメーターが小さいほど分類性能が低下している。これは表2.1(f), 2.1(g) (REM: 学習時) と同じ傾向である。

REM による分類バイアス

REM 適用時に選択されるパラメーターに依存してモデル出力が偏った結果になることを分類バイアスとして考察する。予備実験の分類結果の例を図2.5に示す。表2.5(a)は、テストデータに REM を適用した際に信頼できると判断されたテストデータ内の ROI の総数を示し、表2.5(b)は、学習時に REM を適用した際に信頼できると判断された学習データ内の ROI の総数を示す。表2.5(a), 2.5(b)はどちらも予備実験における 100 回のクロスバリデーションの結果を示している。簡単のため、一方のクラスのパラメーターが大きく、反対のクラスのパラメーターが小さい 2 つの設定 $\text{dy}; \text{dx} = 24(\text{cir}); 48(\text{nor})$ と $\text{dy}; \text{dx} = 48(\text{cir}); 24(\text{nor})$ に焦点を当てる。

(a) テスト時の適用 表2.5(a) (REM: テスト時) に示すように、より大きな周辺パラメータークラスに分類された ROI は、ラベル付けされたクラスに関わらず、信頼できないと判断される傾向があった。この理由の一つとして、テストデータに REM を適用した場合、より小さな周辺パラメータークラスに分類された ROI がテストデータに残る傾向があることが挙げられる。そのため、被験者単位の出力は、より小さな周辺パラメータークラスに偏る傾向があり、図2.5でも確認できる。

(b) 学習時の適用 表2.5(b) (REM: 学習時) からわかるように、より小さな周辺パラメータークラスに分類された ROI は、ラベル付けされたクラスに関わらず、信頼できると判断される傾向がある。これらの ROI は、ラベル付けされたクラスと分類されたクラスが異なる場合でも、学習データとして使用されたことに注意する必要がある。このため、より大きな周辺パラメーターが割り当てられた学習データは、反対のクラスに分類された ROI によって支配される可能性があり、その中には反対のクラスの画像特徴を持つ ROI も含まれる。このような状況では、被験者単位の出力は、より大きな周辺パラメータークラスに偏ることになる。テストデータへの適用と同様に、図2.5でも確認できる。

例えば、正常の周辺パラメーターが肝硬変のパラメーターよりも小さい場合、正常と分類された ROI が肝硬変の学習データに残る可能性がある。この結果、正常な特徴を持つ ROI が肝硬変クラスの学習データで多く学習されるため、テストデータの出力が正常になる傾向があり、対策を行わない場合、偽陰性率が増加する可能性がある。

(c) 学習時とテスト時の適用（共通パラメーター） 前述のように，REM の精度は，学習時とテスト時の両方に共通パラメーターを使用した場合，テスト時のみに適用した場合と同様の特性を示す傾向があった．これは，学習時とテスト時で最適なパラメーターが異なり，テスト時の REM 使用でテストデータから信頼性の低い ROI を棄却することで，被験者単位の識別性能全体を直接的に向上させることができるためである．逆に，学習時に REM を使用すると，学習データから信頼性の低い ROI を棄却することで，被験者単位の最低識別性能を向上させることができる．

(d) 学習時とテスト時の適用（個別パラメーター） [学習時とテスト時の適用（共通パラメーター）] ではテスト時の REM 適用が学習時と比較して分類性能に影響を与えていたが，学習時とテスト時の両方に個別パラメーターを使用した場合，学習時に最適なパラメーターを使用することでさらなる分類性能向上がみられた．よって，[学習時とテスト時の適用（共通パラメーター）] と比較して柔軟に周辺領域パラメーターを選択可能になり，パラメーター選択におけるバイアスが低減されている．

周辺領域の広さと画像所見の関係

ここでは REM 適用時における周辺領域の広さについて，超音波画像上での結節性の大きさなど医学的観点との関係性を分類性能の変化を軸に考察する．考察ではまず REM によるテスト時の周辺領域設定を対象とし，分類性能向上に有効な周辺領域の広さから各クラスの超音波画像上での分布について検討を行う．さらにこの検討をもとに，REM による学習時の周辺領域設定についても，分類性能向上に有効な周辺領域の広さについて検討を行う．

(a) テスト時における最適な周辺領域の検討 表2.1(c)，2.1(d) (REM: テスト時) について，肝硬変クラスへ分類された ROI の REM における周辺領域パラメーターを $dy; dx = 48 \rightarrow 40$ もしくは $40 \rightarrow 32$ と設定したとき感度が大きく増加し， $dy; dx = 32 \rightarrow 24$ としたとき特異度が顕著に低下した．[テスト時の適用] の考察より，肝硬変クラスに分類された ROI の周辺領域 $dy; dx$ を小さく設定するほど，肝硬変クラス分類 ROI がテストデータで信頼性があると判断され，Product Rule での統合に使用されるため感度が増加し，特異度が低下する傾向がみられていた．一方で， $dy; dx = 48 \rightarrow 40$ と $40 \rightarrow 32$ で感度が大幅に増加したことから，REM で信頼性があると判断された肝硬変クラス分類 ROI が，被験者単位のラベルと多く一致していたと考えられる．そのため，本章で使用したデータセットの画像空間上での肝臓領域において，肝硬変の所見として見られる線維化が $dy; dx = 40$ もしくは 32 で作られる周辺領域の範囲内にある可能性が高いと考えられる．また表2.1(d) (REM: テスト時) について，周辺領域のパラメーターを

dy;dx = 32 以上と設定した際には、健常者の肝臓領域に存在する健常な ROI が周辺領域に含まれる確率が増加し、肝硬変クラスに分類された ROI が棄却され、特異度が一定の値を保っていた。一方で、dy;dx = 24 と設定することで周辺領域に健常な ROI が含まれる確率が急激に低下し、他の周辺領域の設定では棄却されていた肝硬変クラスに分類された ROI が棄却されずに Product Rule での統合に使用された結果、被験者単位のラベル (= 健常者クラス) と一致していないため特異度が大幅に低下したと考えられる。そのため、本章で使用したデータセットの画像空間上での肝臓領域において、肝硬変クラスに分類された ROI に設定する周辺領域は、dy;dx = 32 または 40 が適切であると考えられる。(実際の周辺領域のスケール: 64(px) or 80(px) → 縦軸方向では 30(mm) or 38(mm)。非代償期では、広範囲の線維化により肝臓が委縮し、表面に明確な大結節が数 cm 単位で形成・巨大化される。それらを分厚い繊維が取り囲むため、30(mm) や 38(mm) の広さは一定の合理性を有すると考えられる。) 一方で、超音波画像上での肝硬変と健常な領域はあくまでも画像上での見え方であり、超音波のノイズや撮影条件の変化の影響により実際の肝臓表面の状態とは異なる。また、REM による ROI の信頼性評価では学習済みモデルの出力を使用しており、使用するモデルや学習データのラベル付けの品質によって、肝硬変や健常とされる領域は変化することに注意する必要がある。

(b) 周辺領域拡大による特異度優位傾向の要因 次に、表2.1(c), 2.1(d) (REM: テスト時) の中段では、感度よりも特異度が高い傾向にある。これは、肝硬変クラスへ分類された ROI の周辺領域に存在する健常な ROI の数が増加し、REM によって棄却される確率が増加したためである。一方で、dy;dx = 56 以上の広さを設定した場合は中心 ROI と周辺領域が隣接せず離れるため、健常な ROI が肝硬変クラスに分類された ROI に隣接していても周辺領域として考慮されず、離れた位置に存在する肝硬変クラスに分類された ROI が周辺領域として含まれる場合があり、感度が一定の精度を保っている。

(c) 広すぎる周辺領域設定による特異度増加の要因 また表2.1(c), 2.1(d) (REM: テスト時) の一番下の段では、肝硬変クラスに分類された ROI に設定する周辺領域のパラメーターを dy;dx = 72→80 としたときに特異度が大幅に増加した。この周辺領域の設定では、超音波画像上での肝臓領域の端にある ROI を中心 ROI としたとき、肝臓領域の反対を周辺領域として含むようになり、中央の ROI と周辺領域の間に 32 画素分の隙間ができる。また、周辺領域の範囲が肝臓領域の外側にはみ出すケースも増加する。健常な領域は広く分布していると仮定し、この設定でも高信頼であると判定されやすい。一方で、肝硬変クラスに分類された ROI は、中央の ROI に近い (重なっている、もしくは隣接している) ROI が肝硬変クラスに分類された ROI であっても、周辺領域をそれらより広く設定して周辺領域として参照しないため、もともと広く分布していると仮定す

る健常な領域によって低信頼であるとされ棄却されやすい。これらの要因から、この周辺領域の設定において特異度が急激に増加したと考えられる。

(d) 学習時における狭い周辺領域設定の医学的妥当性 次に、表2.1(f), 2.1(g) (REM: 学習時) について、肝硬変クラスへ分類された ROI の周辺領域を $dy; dx = 24$ としたときの特異度の増加が激しい。これは、肝硬変クラスに設定する周辺領域を $dy; dx = 32$ 以上としたときに棄却されていた ROI が棄却されずに学習データとして使用されたためである。そのため、この設定において $dy; dx = 24$ の狭い範囲で存在する肝硬変クラスの特徴を持った ROI を健常者クラスの学習データとして学習しており、そのような ROI がテストデータの健常者にも多く含まれたため、特異度が急激に増加した。一方で感度については、周辺領域パラメーターのすべての設定において高い値となっている。これは肝硬変クラスの学習データで肝硬変クラスに分類された ROI と比較して、健常者クラスに分類された ROI が広く存在しており REM によって棄却されずに肝硬変クラスの学習データとして学習されたため、周辺領域の設定によらず感度が高くなったと考えられる。健常な領域を肝硬変クラスとして学習していることに加えて、肝硬変クラスに分類された ROI の周辺領域を $dy; dx = 24$ としたときにも感度が高く維持しているのは、この周辺領域の設定で REM によって高信頼であるとされる肝硬変の特徴を持つ ROI がもともと肝硬変クラスの学習データにも多く存在しているからであると考えられる。これらの結果を踏まえ、仮に健常者であっても、脂肪肝や超音波ノイズによる影響で $dy; dx = 24$ の広さで肝硬変クラスに似た特徴を持つ ROI が存在する可能性があると考えられる。このような ROI を健常者クラスとして学習すること自体は妥当である可能性がある。一方で、大きさが 3(mm) 以下の微小結節性をはじめとする局所的な結節性と隔壁が健常者クラスとして学習されている場合には、医学的観点からは不適切なモデルの学習が生じている可能性がある。この点についての明確な判断は、REM によって高信頼と判断された ROI に対して、医師による観察と評価が必要である。

2.5.2 実践的実験

実験結果

実践的実験では、予備実験と同じベースラインを用いて、REM による高信頼 ROI を用いた分類器の実用性を検証する。実験における各試行は、(i) 周辺領域パラメーターの自動最適化(図2.4の (c)), (ii) テストデータに対する分類性能の評価(図2.4の (b)), で構成される。各試行における学習データ被験者数とテストデータ被験者数は、各クラス 6 名と 2 名とした。試行は 100 回繰り返され、各試行において周辺領域の全候補パターンにおける分類性能を評価した。各クラ

ス 6 名の被験者の学習データにおいて、モデル訓練用に 5 名、検証データとして各クラス 1 名を選び、合計で $6 \times 6 = 36$ 通りの組み合わせすべてを用いて周辺パラメーターを最適化した。(i) 周辺領域パラメーターの自動最適化(図2.4の(c))の方法として、REM を検証データ 1 名のみ適用するケース、学習データ 5 名のみ適用するケース、検証データ 1 名と学習データ 5 名の両方に適用するケース、を検討した。それぞれのケースにおいてパラメーターを比較するために、まず検証データに対する F 値が高いパラメーターを最適なパラメーターとした。複数のパラメーターが同じ F 値であった場合は、Product Rule によって統合された出力値の 2 クラス間の差が小さいパラメーターを最適なパラメーターとした。すべての組み合わせを試した後、最適な周辺パラメーターを用い、(ii) テストデータに対する分類性能の評価(図2.4の(b))を、提案システム(i);(ii);(iii);(iv) のテストデータに対する分類性能で評価した。ここでは、検証データの被験者を含む 6 名の訓練データの被験者を用いて、分類器を再学習した。また、最適化によって設定される周辺領域パラメーターの選択候補として、(1) cir, nor $\in \{24; 32; 40; 48; 56; 64; 72; 80\}$, (2) cir, nor $\in \{24; 32; 40; 48\}$ の 2 パターンを比較し、パラメーター候補を限定することによる汎化性能への影響を分析した。

表2.6はベースライン、提案システム(i)~(iv) のテストデータの被験者における AUC, 識別率, 感度, 特異度, 適合率, F 値の結果を示す。表の中段 3 行は周辺領域のパラメーター候補を $8 \times 8 = 64$ 通り ($\{24, 32, 40, 48, 56, 64, 72, 80\} \times \{24, 32, 40, 48, 56, 64, 72, 80\}$), 下 3 行は $4 \times 4 = 16$ 通り ($\{24, 32, 40, 48\} \times \{24, 32, 40, 48\}$) とする。また、表2.7(a) ~ 2.7(c), 表2.7(d) ~ 2.7(f)は表2.6の提案(i) ~ (iii) おいて選択された周辺領域パラメーターの分布と検証データに対する F 値を示す。

識別率に注目すると、提案システム(ii), (iv) が識別率 0.8925, F 値に注目すると提案システム(ii) が 0.8938 で最も良い精度であった。ここで、表2.7(b), 2.8(c)に注目すると、肝硬変クラスに分類された ROI の周辺領域として $dy; dx = 24$, 健常者クラスに分類された ROI の周辺領域として表2.7(b) : $dy; dx = 80$, 表2.8(c) : $dy; dx = 48$ が最も選ばれている。これは予備実験の考察[学習時における狭い周辺領域設定の医学的妥当性]と同様の傾向を示しており、最適化によって選ばれたこの周辺領域が医学的に肝硬変の所見を捉えているか、医師による観察と評価が必要である。提案システム(iv) では、予備実験の考察[学習時とテスト時の適用(個別パラメーター)]より、学習時とテスト時で個別の最適なパラメーターを選択することで、学習時とテスト時で共通のパラメーター使用時よりも分類性能向上を実現していた。まず表2.8(c)より、学習データには肝硬変クラスに分類された ROI の周辺領域として $dy; dx = 24$, 健常者クラスに分類された ROI の周辺領域として $dy; dx = 48$ が最も選ばれており、表2.8(a)より、検証データには肝硬変クラスに分類された ROI の周辺領域として $dy; dx = 32$, 健常者ク

ラスに分類された ROI の周辺領域として $dy; dx = 40$ が最も選ばれていた。提案システム (iv) では表2.9より、学習データには肝硬変クラスに分類された ROI の周辺領域として $dy; dx = 24$ が 87 回、健常者クラスに分類された ROI の周辺領域として $dy; dx = 48$ が 46 回選択されている。また、検証データでは肝硬変クラスに分類された ROI の周辺領域として $dy; dx = 32$ が 62 回、健常者クラスに分類された ROI の周辺領域として $dy; dx = 40$ が 49 回選択されており、学習時とテスト時に適した周辺領域を個別に選択できていることが分かる。

最適な周辺領域の広さ

周辺領域パラメーターの自動最適化において検討する広さは、表2.6の提案システム (i) ~ (iv) 3 行ずつを比較すると $4 \times 4 = 16$ 通りに範囲を制限した方が精度 (AUC, 識別率, F 値) が高くなったといえる。特に、提案システム (i); (iii) でその傾向にある。これは、予備実験の提案システム (i) の特異度の値が表2.1(d)の一番したの行において急激に増加したことに起因する。周辺領域の広さを $\{24; 32; 40; 48; 56; 64; 72; 80\}$ の 8 通り (2 クラスでは $8 \times 8 = 64$ 通り) から選ぶ場合、最も広い設定では肝臓領域の端や反対側を周辺 ROI とし、広く分布していると仮定される健常領域が周辺領域に存在する確率が高くなる。そのような広すぎる周辺領域の設定は、中心 ROI の信頼性評価に適切でない可能性が高く、表2.7(c)では肝硬変クラスに分類された ROI の周辺領域として $dy; dx = 80$ が 32 回選択されていたのに対し、表2.8(e)では $\{24; 32; 40; 48\}$ の狭い範囲で最適化を実現している。

医師による目視診断との精度比較

第 1 章 超音波画像を用いた医師の目視診断の精度での、目視診断の精度と本章の肝硬変検出の精度を比較する。本章で対象とした肝硬変ステージ F4 を対象とした場合、目視診断の精度では、識別率 66(58–73)%, 感度 48(27–69)%, 特異度 82(74–88)% (括弧内は 95% 信頼区間) であった [18]。また、他の肝硬変ステージを含めた文献でも特異度が約 80% – 90% の値に対して感度が約 40% – 70% と低い傾向にあった [17, 26, 16]。もともと超音波画像を用いた肝硬変診断は、肝硬変の早期疑いの段階で非侵襲的に用いられるため、肝硬変の所見を断定的に見つけることが困難であり、超音波画像のみを根拠とした判定は健常となりやすい。一方で、本章で提案したシステムでは表2.6において、特異度と比較したときの感度が低くなりやすいか、特異度よりも感度が高くなる傾向にある。そのため、特異度が低くなる傾向にある医師の目視診断の支援として本章の提案システムを使用する場合は、機械学習による定量的な観点から肝硬変疾患の検出を支援できる可能性が高い。いずれにしても、機械学習による肝硬変検出

表 2.6: 分類性能の比較.

周辺領域 の選び方	手法	AUC	識別率	感度	特異度	適合率	F 値
-	ベースライン	0.9492	0.8450	0.900	0.7900	0.8108	0.8531
$8 \times 8 = 64$ 通り	提案 (i)	0.9033	0.8375	0.7850	0.8900	0.8771	0.8285
	提案 (ii)	0.9469	0.8925	0.8800	0.9050	0.9026	0.8911
	提案 (iii)	0.9461	0.8375	0.8950	0.7800	0.8027	0.8463
$4 \times 4 = 16$ 通り	提案 (i)	0.9284	0.8775	0.8550	0.9000	0.8953	0.8747
	提案 (ii)	0.9509	0.8925	0.9050	0.8800	0.8829	0.8938
	提案 (iii)	0.9531	0.8600	0.9150	0.8050	0.8243	0.8673
	提案 (iv)	0.9465	0.8925	0.8550	0.9300	0.9243	0.8883

の精度は学習データの規模やラベル付けの品質などに大きく依存するため、検証データを用いた汎化性能の把握など機械学習モデルの分類性能の監視が必要である。

2.6 まとめ

本章では、超音波画像を用いた肝硬変検出のための信頼性の高い ROI 選択手法として REM を提案した。各 ROI の事後確率と位置情報を用いて、検証データの F1 スコアを最大化するように周辺パラメーターの最適化を行った。実験結果では、REM を用いて信頼性の高い ROI を選択することで、提案システムの分類性能と周辺領域パラメーターの自動最適化による実用性が向上することが示された。

REM のさらなる改良のためには、肝臓領域の形態的な大きさや形状情報を活用して、周辺パラメーターをより個別に最適化することが今後の課題である。本研究に関しては、今後の方向性として 2 つの観点がある。まず第 1 に、本研究では ROI の選択と分類結果の統合的性能に基づいて REM を評価しているため、より大規模なデータセットに対応できるよう、多数のパラメーターを持つ CNN などの他の分類器を導入する必要がある [51]。第 2 に、ROI の分類結果と病理学的所見との関係性を検討し、より高い解釈性と信頼性を確保することが不可欠である。また、本研究で使用した超音波画像データセットについては、超音波画像から肝臓領域を自動抽出できるようにすることで、診断支援プロセスのさらなる自動化が可能となり、より信頼性の高い提案と性能評価が可能となる。

表 2.7: $8 \times 8 = 64$ 通りにパラメーターの範囲を設定した場合.

(a): 検証データに REM を適用した際の最適なパラメーターの選択.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	-	-	1	-	-	-	-	-
	32	4	5	15	9	4	-	-	1
	40	6	1	1	1	1	-	-	1
	48	1	-	1	-	-	-	-	1
	56	1	5	1	-	1	-	-	1
	64	-	-	-	-	-	-	-	-
	72	2	1	3	2	2	-	-	1
	80	4	6	3	4	2	1	4	3

(b): 学習データに REM を適用した際の最適なパラメーターの選択.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	2	-	1	15	3	5	8	58
	32	-	-	-	1	1	2	-	-
	40	-	-	-	-	-	1	-	-
	48	-	-	-	-	-	-	1	-
	56	-	-	-	-	-	1	-	-
	64	-	-	-	-	-	-	-	-
	72	-	-	-	-	-	-	-	-
	80	-	-	-	-	-	-	-	-

(c): 学習データと検証データに REM を適用した際の最適なパラメーターの選択.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	-	1	1	1	-	-	-	-
	32	1	5	2	12	13	1	-	3
	40	-	-	-	-	-	-	-	-
	48	-	-	-	-	5	-	-	2
	56	-	-	6	2	5	3	-	1
	64	-	1	-	-	-	-	-	-
	72	-	-	1	2	-	-	-	-
	80	-	2	9	15	3	1	-	2

表 2.7: $8 \times 8 = 64$ 通りにパラメーターの範囲を設定した場合 (続き).

(d): 検証データに REM を適用した際の各パラメーターの F 値.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	0.8222	0.7932	0.7951	0.7939	0.7921	0.7676	0.7596	0.7679
	32	0.8868	0.8790	0.8924	0.8893	0.8804	0.8579	0.8592	0.8506
	40	0.8596	0.8792	0.8918	0.8939	0.8872	0.8753	0.8806	0.8728
	48	0.8432	0.8714	0.8957	0.9022	0.8975	0.8869	0.8867	0.8647
	56	0.8469	0.8679	0.8890	0.8951	0.8989	0.8910	0.8821	0.8593
	64	0.8363	0.8517	0.8738	0.8744	0.8760	0.8567	0.8589	0.8467
	72	0.8604	0.8669	0.8782	0.8849	0.8897	0.8647	0.8626	0.8456
	80	0.8601	0.8838	0.8944	0.8993	0.8982	0.8760	0.8817	0.8797

(e): 学習データに REM を適用した際の各パラメーターの F 値.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	0.8051	0.8500	0.8765	0.8865	0.8854	0.8921	0.8942	0.9015
	32	0.7372	0.7676	0.7799	0.7874	0.7874	0.7876	0.7874	0.7990
	40	0.7064	0.7301	0.7439	0.7554	0.7550	0.7568	0.7524	0.7647
	48	0.7218	0.7494	0.7635	0.7710	0.7715	0.7717	0.7693	0.7749
	56	0.7296	0.7528	0.7682	0.7742	0.7760	0.7750	0.7733	0.7782
	64	0.7253	0.7499	0.7651	0.7739	0.7758	0.7756	0.7756	0.7806
	72	0.7182	0.7464	0.7636	0.7715	0.7728	0.7724	0.7707	0.7782
	80	0.7046	0.7340	0.7515	0.7585	0.7596	0.7614	0.7610	0.7664

(f): 学習データと検証データに REM を適用した際の各パラメーターの F 値.

		nor							
dy; dx (dy = dx)		24	32	40	48	56	64	72	80
cir	24	0.7985	0.7974	0.8092	0.8108	0.8218	0.7953	0.7843	0.7868
	32	0.8349	0.8331	0.8425	0.8544	0.8615	0.8190	0.8047	0.8257
	40	0.7585	0.7708	0.7804	0.8015	0.7944	0.7758	0.7643	0.7750
	48	0.8142	0.8197	0.8317	0.8436	0.8542	0.8403	0.8203	0.8318
	56	0.8304	0.8325	0.8475	0.8522	0.8706	0.8500	0.8296	0.8383
	64	0.7983	0.8081	0.8101	0.8222	0.8246	0.8135	0.8014	0.7904
	72	0.8183	0.8278	0.8297	0.8394	0.8436	0.8361	0.8246	0.8250
	80	0.8696	0.8789	0.8844	0.8865	0.8867	0.8772	0.8778	0.8704

表 2.8: $4 \times 4 = 16$ 通りにパラメーターの範囲を限定した場合.

(a): 検証データに REM を適用した際の最適なパラメーターの選択.

		nor			
dy; dx (dy = dx)		24	32	40	48
cir	24	1	-	3	-
	32	10	6	23	10
	40	10	5	4	4
	48	1	4	12	7

(b): 検証データに REM を適用した際の各パラメーターの F 値.

		nor			
dy; dx (dy = dx)		24	32	40	48
cir	24	0.8222	0.7932	0.7951	0.7939
	32	0.8868	0.8790	0.8924	0.8893
	40	0.8596	0.8792	0.8918	0.8939
	48	0.8432	0.8714	0.8957	0.9022

(c): 学習データに REM を適用した際の最適なパラメーターの選択.

		nor			
dy; dx (dy = dx)		24	32	40	48
cir	24	5	2	5	79
	32	-	2	1	4
	40	-	-	-	2
	48	-	-	-	-

(d): 学習データに REM を適用した際の各パラメーターの F 値.

		nor			
dy; dx (dy = dx)		24	32	40	48
cir	24	0.8051	0.8500	0.8765	0.8865
	32	0.7372	0.7676	0.7799	0.7874
	40	0.7064	0.7301	0.7439	0.7554
	48	0.7218	0.7494	0.7635	0.7710

(e): 学習データと検証データに REM を適用した際の最適なパラメーターの選択.

		nor			
dy; dx (dy = dx)		24	32	40	48
cir	24	4	3	3	5
	32	7	8	9	25
	40	-	1	-	4
	48	3	2	7	19

(f): 学習データと検証データに REM を適用した際の各パラメーターの F 値.

		nor			
dy; dx (dy = dx)		24	32	40	48
cir	24	0.7985	0.7974	0.8092	0.8108
	32	0.8349	0.8331	0.8425	0.8544
	40	0.7585	0.7708	0.7804	0.8015
	48	0.8142	0.8197	0.8317	0.8436

表 2.9: 学習データと検証データに REM を適用した際の最適なパラメーターの選択.

		nor train															
		8				16				24				32			
		nor test				nor test				nor test				nor test			
cir train	cir test	8	16	24	32	8	16	24	32	8	16	24	32	8	16	24	32
8	8	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-
	16	-	-	8	1	2	-	11	2	3	1	5	1	5	3	10	3
	24	-	-	-	-	2	-	2	4	-	1	2	-	3	1	3	4
	32	-	-	-	-	-	-	1	1	-	-	-	2	-	-	3	2
16	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	16	-	-	-	-	1	-	-	-	-	-	-	-	-	4	2	-
	24	-	-	-	-	-	-	-	-	-	1	-	1	1	1	-	-
	32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
24	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	32	-	-	-	-	-	-	-	-	-	-	1	-	-	-	-	-
32	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

表 2.10: 学習データと検証データに REM を適用した際の各パラメータの F 値.

nor train															
8				16				24				32			
nor test				nor test				nor test				nor test			
cir	cir	train	test	8	16	24	32	8	16	24	32	8	16	24	32
8	8	8	8	0.8219	0.8077	0.8131	0.8099	0.8540	0.8210	0.8284	0.8210	0.8750	0.8334	0.8334	0.8249
	16	8	16	0.9003	0.8841	0.8884	0.8855	0.9009	0.9097	0.9209	0.9175	0.8950	0.9141	0.9281	0.9258
	24	8	24	0.8844	0.8927	0.8944	0.8931	0.8418	0.8877	0.9151	0.9194	0.8175	0.8715	0.9129	0.9198
	32	8	32	0.8569	0.8799	0.8945	0.8978	0.7930	0.8411	0.8848	0.8951	0.7658	0.8229	0.8727	0.8908
16	8	8	8	0.7892	0.7803	0.7783	0.7768	0.7983	0.7941	0.7917	0.7906	0.8027	0.8021	0.8021	0.7987
	16	8	16	0.8558	0.8075	0.8026	0.7968	0.8853	0.8557	0.8435	0.8394	0.9000	0.8717	0.8633	0.8596
	24	8	24	0.8517	0.8041	0.7882	0.7873	0.8969	0.8635	0.8522	0.8488	0.9113	0.8902	0.8846	0.8805
	32	8	32	0.8681	0.8256	0.8204	0.8168	0.8899	0.8710	0.8588	0.8610	0.8903	0.8894	0.8814	0.8881
24	8	8	8	0.7743	0.7679	0.7653	0.7654	0.7836	0.7755	0.7735	0.7728	0.7882	0.7792	0.7777	0.7761
	16	8	16	0.8038	0.7778	0.7761	0.7744	0.8559	0.8068	0.7996	0.7953	0.8695	0.8254	0.8158	0.8105
	24	8	24	0.7923	0.7579	0.7495	0.7496	0.8516	0.8007	0.7856	0.7839	0.8712	0.8253	0.8079	0.8076
	32	8	32	0.8166	0.7841	0.7797	0.7740	0.8563	0.8225	0.8177	0.8137	0.8766	0.8405	0.8318	0.8320
32	8	8	8	0.7820	0.7713	0.7674	0.7678	0.7944	0.7858	0.7832	0.7815	0.7961	0.7914	0.7894	0.7869
	16	8	16	0.8166	0.7897	0.7863	0.7839	0.8575	0.8237	0.8160	0.8104	0.8667	0.8410	0.8341	0.8279
	24	8	24	0.8033	0.7707	0.7609	0.7598	0.8592	0.8174	0.8030	0.8021	0.8819	0.8390	0.8251	0.8231
	32	8	32	0.8294	0.7944	0.7924	0.7855	0.8655	0.8334	0.8286	0.8260	0.8836	0.8530	0.8430	0.8443

第 3 章

位置情報を用いた信頼性評価と Multiple Instance Learning の統合 による CNN モデルの肝硬変検出の 精度向上

第 2 章では，超音波画像を用いた肝硬変診断支援のために，Gabor 特徴と二次分類器を用いた肝硬変検出システムを提案した．また，検出精度向上のために学習済み分類器を用いて得られる事後確率と位置情報に基づいて各関心領域 (Region of Interest, ROI) の信頼性を評価し，信頼できる ROI を用いた分類器の学習とテスト時の分類による肝硬変検出精度向上を実現した．これらの画像特徴と分類器は，人の知見に基づいて設計されており，設計者の経験に分類性能が強く依存する．本章では，肝硬変検出に対して CNN モデルを適用し，設計者の知識と経験に依存しない特徴抽出と分類の実現を目的とする．また，各 ROI に対して位置情報を用いた信頼性評価法を適用し，CNN モデルにおいても学習やテスト時の分類に有効な ROI を選択することで，肝硬変検出性能の向上を目的とする．

3.1 はじめに

第 2 章では，超音波画像を用いた肝硬変診断支援のために，Gabor 特徴と二次分類器を用いた肝硬変検出システムを提案した．また，検出精度向上のために学習済み分類器を用いて得られる事後確率と位置情報に基づいて各関心領域 (Region of Interest, ROI) の信頼性を評価し，信頼できる ROI を用いた分類器の学習とテスト時の分類による肝硬変検出精度向上を実現した．

Gabor 特徴や二次分類器は，人の視覚に基づくフィルタの設計や特徴空間で

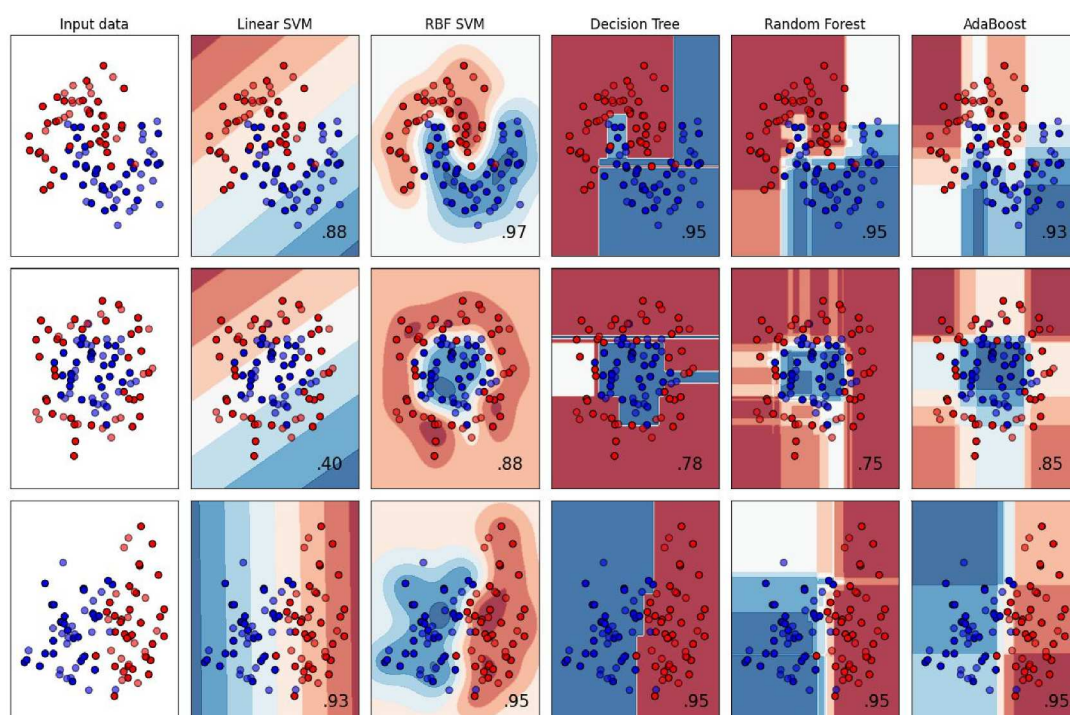


図 3.1: 使用する分類器による分類境界の違い。

のデータの正規分布の仮定といった、人の知見に基づいて設計されたものである。そのため、設計者の仮定に沿った特徴量やデータ分布に従う問題設定では一定の分類性能を実現可能である。一方で、そのように設計された分類システムは設計時に固定された特徴量の構造や分布の仮定に分類性能が強く依存してしまう。

例として、同じ分布を持つデータに対して分類器の選択の違いによる特徴空間での分類境界の例を図3.1に示す [52]。

Support Vector Machine (SVM) は、学習データ内で分類境界に最も近いサポートベクトルと呼ばれるデータ点と分類境界の間のマージンを最大化することで学習を行う。カーネルトリックという特徴変換を用いることで Linear SVM や RBF SVM のように線形または高次元の異なる分類境界を決定する。図3.1では、RBF カーネルを用いた SVM がなめらかな分類境界を学習した一方で、Linear カーネルを用いた SVM は線形の分類境界を学習することからムーンデータ（上段）やサークルデータセット（中段）に対してクラス間の分離が不十分であることが分かる。

また、Decision Tree(決定木), Random Forest(ランダムフォレスト), AdaBoost(アダブースト) はそれぞれ木構造に基づく分類モデルである。決定木はデータ入力から予測出力までの流れを木の構造に見立て、各枝におけるデータの分割を繰り返していき、最終的にたどり着いた葉においてクラスの予測を行う。ランダ

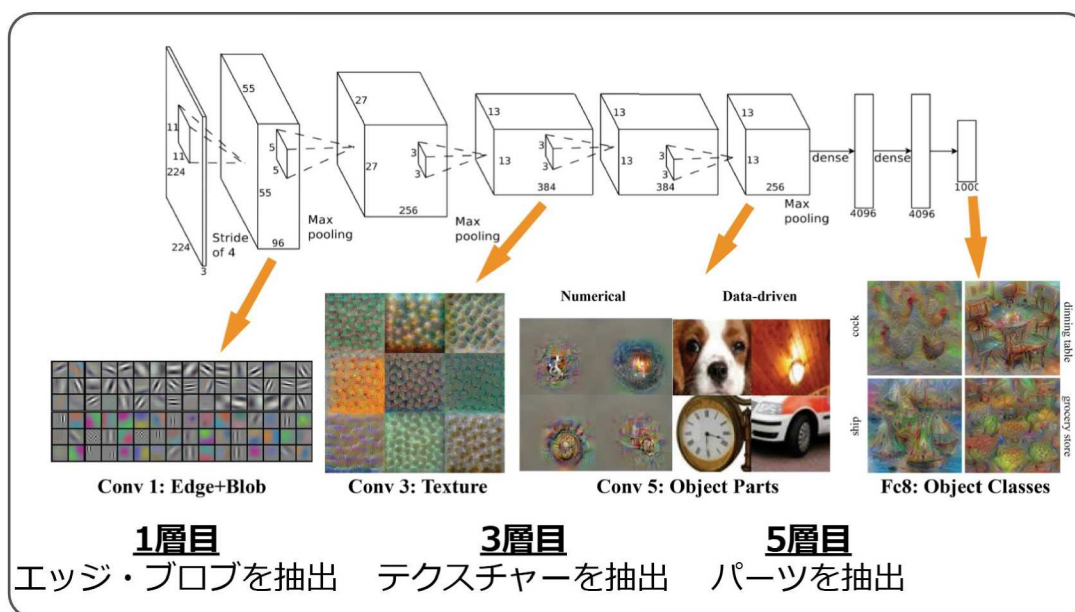


図 3.2: ImageNet を用いた学習で獲得された AlexNet の特徴量。
[53] より引用，加筆。

ムフォレストは複数の決定木を独立に学習させ，それらの出力を統合することで汎化性能を向上させる手法である．アダブーストはある決定木が誤分類したサンプルの重みを増やして再学習を行い，これを継続することで誤分類サンプルに対する学習に重点をおいた学習を実現する．木構造モデルでは各分岐において 1 次元の特徴量に基づく閾値判定を行うため，分類境界が可視化されたようにジグザグ状の形状を呈する．このため，ムーンデータセットとサークルデータセットにおいて画像中央の青色クラスの予測領域が画像の端まで広がってしまっていることが分かる．

これらのように，データが生成された過程を十分に考慮せず経験と勘により特徴量と分類器を選択すると，本来得られる分類性能を大きく低下させる原因となる可能性がある．そのため，機械学習を活用したシステムを応用場面で利用する際は，システム設計者による適切な特徴量と分類器の設計・選択が必要である．

ILSVRC 2012 と呼ばれる画像認識世界大会でそれまでの人手による特徴量設計手法と比較して，AlexNet[51] が大幅な精度向上を達成したことをきっかけに，様々な領域において深層学習を用いた画像認識の研究が行われてきた．深層学習，特に畳み込みニューラルネットワーク (Convolutional Neural Network, CNN) を用いることが一般的である．CNN は従来の画像認識システムにおける人手での特徴設計を，画像の入力から出力までの End-to-End で行うことができる．

図3.2は，ImageNet データセット [54] を用いて学習した AlexNet[51] で獲得された特徴量の例を示す [53]．ここで，一般的にモデル内のパラメーターはモデル

の予測と実際のラベルの乖離を示す損失値を確率的勾配降下法で最小化することで行う。従来の人手による特徴設計と比較して End-to-End による学習では、モデルの高い表現力から多量の学習データを必要とする。一方で、人手による学習への介入を基本的には必要とせず、設計者によるバイアスがモデルの分類性能低下を引き起こす可能性を抑制できる。また、SVM など従来の機械学習手法では、特徴抽出後にあらかじめ設計された分類器へ一度だけ入力することで出力を得ていたため、複雑な画像特徴や分類境界の学習が困難であった。一方で、CNN ではショートカットコネクション [55] を用いた多層化により、勾配消失を抑制しつつフィルタ適用の回数を大幅に増やすことができ、結果として複雑な画像特徴を抽出し分類へ活用することができる。

CNN を用いた画像認識技術は、cifar-10[56] などの自然画像だけでなく、医療画像にも広く活用され、従来の機械学習と比較してより精度の良い予測により疾患の発見に貢献している。特に第 2 章で述べたように、機械学習技術を医療画像へ適用する際は詳細なラベル情報を医師が付与する必要があるが、同様に人手による特徴設計や分類器の選択では、検出対象とする疾患の画像内所見や画像モダリティへの理解といった医学と工学の双方の知識を必要とするため、人的コストが非常に高くなる。

本章では、第 2 章で扱った超音波画像を用いた肝硬変検出に対して CNN モデルを適用し、設計者の経験と勘に依存しない特徴抽出と分類の実現を目的とする。また、各 ROI に対して位置情報を用いた信頼性評価法 (Reliability Evaluation Method, REM) を適用し、CNN モデルにおいても学習やテスト時の分類に有効な ROI を選択することで、肝硬変検出性能の向上を目的とする。

以下に本章における貢献を述べる。

- 超音波画像を用いた肝硬変検出に CNN モデルを使用し、人の経験に依存しない特徴抽出で分類性能の向上を検討する。
- 自動設定された ROI を用いた学習において、Multiple Instance Learning (MIL) フレームワークを適用し、学習データの中で学習に不要な ROI を棄却することで、分類性能の向上を検討する。
- MIL フレームワークでの学習とテスト時の REM による高信頼 ROI の選択により、分類性能向上を実現する。

3.2 先行研究

第 2 章で述べたように、近年、画像処理と機械学習を用いたコンピュータ診断支援システムが開発され、「セカンドオピニオン」として医師が肝硬変を診断するのを支援している [33]。それらの中で、畳み込みニューラルネットワーク

(CNN) は超音波画像を用いた高精度な肝硬変検出を実現し、広く活用されている [57]. CNN モデルは多数のパラメーターを持つため、適切な学習と汎化能力の獲得には多量の画像が必要となる傾向がある. これまでの研究では、研究者たちはデータ量とモデルの複雑さを考慮しながら、最適な CNN アーキテクチャを設計することに取り組んできた. 応用場面を考慮すると、もう一つの重要な側面は、高い分類性能を維持しつつ、ラベリング作業の負担を軽減することである.

Wang らは、肝臓表面の硬さを色として可視化できる超音波画像技術の発展版の Shear Wave Elastography とディープラーニング手法を組み合わせた [58]. この研究では、398 名の患者から得られた 1990 枚のマルチセンター画像データを収集することが、分類性能の向上に重要な役割を果たした. 一方、モデル学習に利用できる超音波画像の枚数が比較的少ない場合には、転移学習を用いて他のデータセットから得られた知識を活用することが有効である. たとえば Meng らでは、正常から S1~S4 までの線維化ステージ分類を目的に 279 名の患者の画像を収集しており、これは単純な肝硬変／健常の二値分類よりも困難な課題である [59]. 著者らはデータ拡張を行い、事前学習済み CNN モデルから特徴を抽出することで、学習データの不足による影響を低減した. さらに、より少ないデータ (55 名の被験者) を用いた Byra ら [60] では、CNN モデルを特徴抽出器としてのみ利用し、分類器としてサポートベクターマシン (SVM) を採用している. この構成により、少量データでもより高い汎化性能を得ることを目指している. 転移学習以外にも、超音波画像内で局所的に重要な領域を特定することが、分類性能の向上に有用である. Biswas ら [61] らは、63 名の患者の肝臓超音波画像に対し、背景領域を除去するクロッピング処理によって肝臓組織のみを含む関心領域 (Region of Interest, ROI) を生成し、これにより分類性能が向上したことが示されている. また、文献 [62] では、肝臓組織を構造的に支え、保護する役割を持つ肝被膜に着目した分類手法が提案された. 肝硬変が進行すると、超音波画像上で肝被膜が視覚的に厚くなるため、この特徴に基づき CNN モデルが 91 枚の画像から正常肝と肝硬変を分類できることを示した.

第 2 章で述べたように、我々の研究室では超音波画像内に自動設定された ROI を用い、Gabor 特徴と二次識別器 [43] による肝硬変検出をこれまで提案してきた [44]. MILBoost と Self-training を用いたアプローチでは、分類器の学習に有用な ROI をそれぞれの特徴量と被験者単位のラベルをもとに選択・重みづけすることで、信頼できるデータに基づく学習を行ってきた [48, 47, 49]. また、特徴抽出後の分類フェーズでは、各 ROI を分類器に入力して得られる予測確率を Product Rule[46] によって統合することで被験者単位のラベルを予測してきた [45].

上記に示す、自動設定された ROI と Gabor 特徴、二次識別器を用いた肝硬変

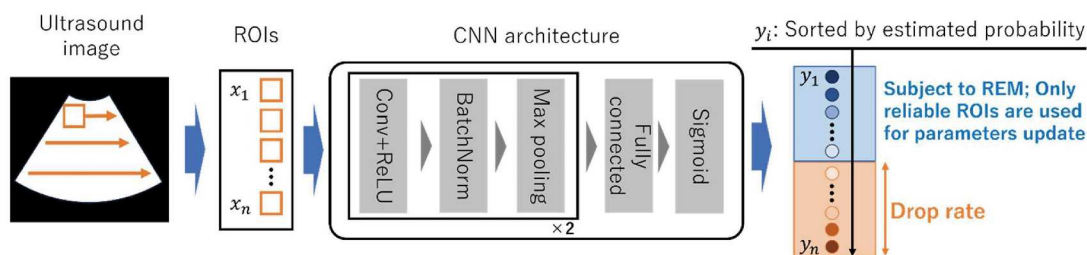


図 3.3: MIL による学習フロー.

検出の発展として、第 2 章で述べた位置情報に基づく信頼性評価法 (Reliability Evaluation Method, REM) がある。また、ROI が超音波画像内に手動で設定される場合も想定されるシナリオの一つである。先行研究 [63] では、医師の指導のもと手動で設定された ROI を使用し Multiple Instance Learning (MIL) による被験者単位のラベル情報のみを用いた弱教師あり学習が肝硬変検出に提案された。MIL による ROI 選択は肝硬変検出に対して一定の有効性を示したが、手動での ROI 設定シナリオでは高品質な ROI を人手で収集する必要があり、分類性能が医師の経験に強く依存してしまう課題がある。

高品質な ROI を手動で設定する労力を削減するためには、各 ROI のもつ画像特徴やラベル情報の品質を妥協しながらも、ラスタスキャンで自動設定された ROI の使用が効果的な代替案となる。藤田ら [63] で提案された MIL フレームワークは実験に先立って手動設定された ROI を使用しており、各 ROI の画像空間上での関係性を考慮していない。自動設定された ROI を使用し、MIL フレームワークに REM を統合することで、より信頼性の高い ROI の選択を自動設定 ROI からでも実現できる可能性がある。また、MIL フレームワークは弱教師あり学習の問題設定で使用される学習法である。藤田ら [63] において学習済みモデルを用いたテスト時の分類では、テストデータのすべての ROI 分類結果を統合して被験者単位の分類結果を出力していた。ここではテストデータにも手動設定された ROI を使用している。そのため、テストデータにおいても自動設定された ROI を使用することで、各 ROI の位置情報を参照することができるようになる。MIL フレームワークで未対応のテスト時の ROI 信頼性評価を REM によって実現することが可能となり、分類性能の向上が期待できる。

3.3 提案手法: MIL による学習と REM による信頼性評価の統合

本章では、第 2 章で提案した REM と MIL フレームワークを統合し、自動設定された ROI からでも CNN モデルの肝硬変検出精度の向上を実現することを目的とする。

図3.3に、MILによる学習フロー図を示す。提案する学習フローは、自動設定された ROI の CNN による分類と MIL による ROI 信頼性評価、および REM による信頼性評価で構成される。ここで、CNN には自動設定された ROI がすべて入力され、その出力に基づいて MIL による信頼性評価が行われる。その後、MIL によって信頼できると判断された ROI のみが REM の入力として使用され、REM によって信頼できると判断された ROI のみを使用して CNN モデルのパラメータ更新を行う。これ以降、図3.3に示す CNN モデル構造、MIL 学習フレームワーク、REM について詳しく説明し、その後テスト時の複数 ROI を用いた信頼性評価と被験者単位の分類方法についても説明を行う。

3.3.1 CNN モデル構造

CNN モデルへの入力超音波画像上に自動設定されたすべての ROI を用いる。CNN モデルの出力は、各 ROI が肝硬変クラスに属する確率 $[0; 1]$ の予測値である。本章では、超音波画像の不規則なノイズによる画像特徴の乱れの影響を低減するために、比較的軽量の CNN モデルとして 2 層の畳み込み層（フィルタサイズ 3×3 、ストライド 1、パディング 1）と 1 層の全結合層で構成される CNN モデルを使用する。この構造は先行研究 [63] を参考にし、実験的にモデル構造について検討した結果選択されたものである。各畳み込み層の後には ReLU 活性化関数とバッチ正規化、最大値プーリング（フィルタサイズ 2×2 、ストライド 2）を適用する。畳み込み層での特徴抽出に加えて、全結合層と Sigmoid 関数による出力層を設け、ROI 単位での出力を得る。

3.3.2 Multiple Instance Learning(MIL)

超音波画像ノイズによって画像特徴が不規則に変化したり、被験者単位でのラベルと一致しない画像特徴を持つ ROI の CNN モデルの学習への影響を低減するために MIL を活用する。MIL では、分類モデルによって予測された出力確率順に各 ROI をソートし、出力確率の降順から任意の割合で ROI を棄却する。ここで、棄却する割合を Drop rate と呼び、Drop rate = None であるときすべての ROI を棄却せずに学習に使用する。Drop rate の範囲に該当する ROI を棄却したのち、残りの ROI を用いて出力確率と被験者単位のラベルに基づく損失の算出と、モデルパラメータの更新を行う。MIL による ROI 信頼性評価の後、後述の REM による信頼性評価を行う場合、Drop rate によって棄却されなかった ROI のみを REM の対象とする。

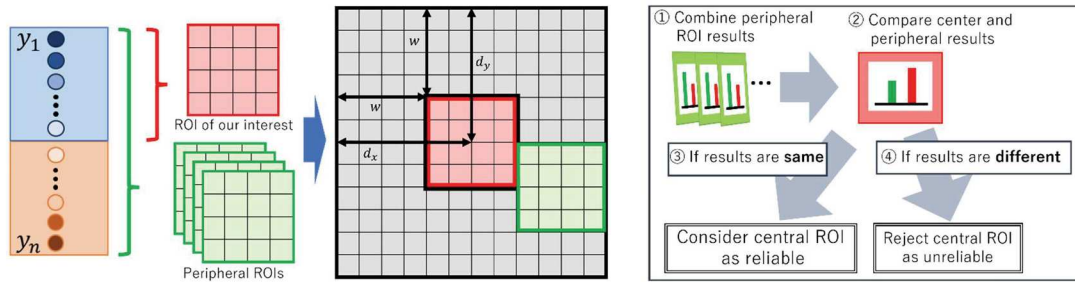


図 3.4: REM を用いた信頼性評価フロー.

3.3.3 信頼性評価方法 (Reliability Evaluation Method, REM)

図3.4は、REM を用いた各 ROI の信頼性評価のフローを示す。REM の詳細なアルゴリズムは第 2 章「信頼性評価法」に記載している。必要に応じて、そこらを参照されたい。REM に入力される ROI に対して位置情報に基づいて、周辺領域 $dy = dx$ および w で決定される周辺領域を定義する。ここで、MIL による信頼性評価では画像特徴のみ考慮しており画像空間上での他の ROI との関係性は考慮していない。そのため、MIL によって信頼できないと判断された ROI も、画像空間上有用な ROI である可能性があるため周辺領域の対象として使用する。周辺領域に存在する周辺 ROI の事後確率を Product Rule もしくは多数決法によって統合した後、中心の注目している ROI と周辺領域の分類結果を比較する。分類結果が一致する場合にのみ中心 ROI を信頼できる ROI と判断する。

MIL は学習フレームワークであるため学習データにのみ適用可能であるが、REM は被験者単位のラベル情報を必要とせず分類モデルが出力した値を使用するため、モデル学習時だけでなくテスト時にも適用可能である。

3.3.4 複数 ROI を用いた分類方法

テスト時の複数 ROI を用いた被験者単位の分類では、自動設定されたすべての ROI を使用する。REM がテストデータに適用される場合、REM によって信頼できると判断された ROI ののみを使用して ROI に対する分類結果の Product Rule もしくは多数決による分類結果の統合と被験者単位の予測結果の出力を行う。

3.4 予備実験

本章では、MIL による学習と REM による信頼性評価の統合による肝硬変分類性能の向上を検討する。予備実験では MIL に共通の Drop rate を用い、以下

の 3 ケースについて検討する.

- (i) MIL(学習時) のみ使用した場合
- (ii) REM(学習時 / テスト時 / 学習時とテスト時) のみ使用した場合
- (iii) MIL(学習時) と REM(学習時 / テスト時 / 学習時とテスト時) を統合した場合

MIL では ROI の信頼性評価をモデル出力値で行うため, 特徴空間での ROI の表現を基準としているのに対し, REM では画像空間上での ROI の位置関係を基準とする. これらの統合によって (iii) MIL と REM を統合した場合の精度が向上することを期待する.

3.4.1 データセットとパラメーター

本章では, 第 2 章で実験に使用したデータセットを使用し, 同一の手順で影・血管領域の前処理と ROI の自動設定を行った.

3.4.2 実装の詳細

CNN モデルの学習では, クロスエントロピー損失を用い, Adam[64] 最適化手法によってモデルのパラメーターを更新した. モデル学習率は 0.001 から開始し, 30 エポックごとに半減させた. 100 エポック学習後, 学習済みモデルの分類性能評価を行った. 学習済みモデルの汎化性能を向上させるために, データ拡張として水平および垂直方向の反転を確率 $p = 0.5$ でそれぞれ行い, コントラストと明るさ調整をランダムに係数 $[0.7; 1.3]$ の範囲で適用した. 学習, 検証, テストの被験者は各クラスそれぞれ 5; 1; 2 名ずつとした. 被験者の選び方をランダムに 20 回行い, 平均の Area Under the Curve (AUC) で最終的な学習済みモデルの分類性能を評価した.

MIL による学習データの信頼性評価では, 2 クラスで共通の Drop rate を $\{\text{None}; 25\%; 50\%; 75\%\}$ から選択した. REM については, 第 2 章実践的実験で分類性能が高いとされた $4 \times 4 = 16$ 通り ($\{24, 32, 40, 48\} \times \{24, 32, 40, 48\}$) のパラメーター設定を用いた.

複数 ROI を用いた分類方法では, 各 ROI に対する出力を Product Rule で統合し被験者単位での予測を行った.

3.4.3 評価指標

評価指標として, 第 1 章で定義した, Area Under the Curve (AUC)[32] を使用した.

表 3.1: AUC の比較.

MIL	REM	AUC
None	None	0.636
	Test	<u>0.720</u>
	Train	0.672
	Train and test	0.700
Drop rate 25%	None	0.654
	Test	0.702
	Train	0.675
	Train and test	0.694
Drop rate 50%	None	0.574
	Test	0.653
	Train	0.626
	Train and test	0.583
Drop rate 75%	None	0.546
	Test	0.581
	Train	0.588
	Train and test	0.663

3.4.4 結果と考察

表3.1に、各手法での AUC の比較をまとめる. MIL と REM の両方を使用していない設定を Baseline とする (MIL=None,REM=None). すべての設定において最も良い精度であったものにアンダーラインを引き, MIL における各 Drop rate 別に最も良い精度であったものを太字で表示する. REM では, 各適用パターンにおいて, 周辺領域の設定候補すべての中で, 最も良い精度であったものを表示する.

MIL

REM を用いずに MIL のみを適用した場合 (REM=None), Drop rate 25% が Baseline より少量の改善を実現し, MIL による学習データでの信頼できる ROI 選択の有効性を確認した. しかしながら, Drop rate が増加していくにあたって全体の AUC が減少していることから, 学習データの不足による CNN モデルの分類性能の低下が確認できる. また MIL の Drop rate None,25%;50% 内において, REM をテストデータに適用しない設定では, 分類に不要な ROI による分類性能の低下がみられる.

REM

MIL が適用されなかった設定 (MIL=None) では, REM を適用した 3 つの設定において分類性能の向上を確認し, 肝硬変検出のための REM による信頼性評価法の有効性を CNN モデル使用時においても確認した. 特に, AUC が 0.7 を上回った 2 つの設定では, REM がテストデータでの信頼できる ROI の選択に重要であったことが分かる.

MIL と REM の統合

MIL と REM の統合による肝硬変検出の精度向上の目的とは対照的に, MIL を学習に使用せず REM のみをテストデータに適用する設定が最も良い精度であった. MIL 適用による AUC の低下は, 大幅な学習データの棄却による CNN モデルの学習不足が原因であると考えられる. そのような少量の学習データを用いた場合でも, REM は Drop rate 75% の設定において AUC0.663 を達成した. REM をテストデータもしくは学習データのどちらかに適用した場合, すべての実験設定において一定程度の性能改善を達成したが, テストデータと学習データの双方に適用した際には大幅な改善が見られた. このことから, REM はテストデータへの適用に限定されず, Drop rate 75% のような大幅に少ない学習データを用いた際に分類性能の向上に貢献できると考えられる.

3.5 実験

予備実験では, MIL と REM の統合によって学習データの信頼性評価を行った際に, 学習データの不足によってモデルの分類性能が低下することが分かった. この節では, 学習データの信頼性評価を MIL のみで行い, テストデータの信頼性評価を REM で行う. さらに, MIL による学習データの信頼性評価では, クラス別の Drop rate を適用することで, 各クラスに適した ROI の棄却による分類性能の向上を期待する.

3.5.1 実装の詳細

CNN モデルの学習では, クロスエントロピー損失を用い, Adam[64] 最適化手法によってモデルのパラメーターを更新した. 学習の初期エポックではモデルの出力がパラメーター初期値のランダム性に強く依存するため, 10 エポック目までは通常の教師あり学習を行い, それ以降のエポックで MIL による学習を行う. モデル学習率は 0.001 から開始する. 学習中は, アーリーストッピングを取り入れて, 検証データに対するロスを 150 エポック監視する. 150 エポック中に

表 3.2: 比較手法の実験結果.

	ベース ライン	MIL (DR 共通)	MIL (DR クラス別)	MIL 未適用	MIL (DR 共通)	MIL (DR クラス別)
MIL	なし	(0.12, 0.12)	(0.24, 0.06)	なし	(0.24, 0.24)	(0.30, 0.12)
REM	なし	なし	なし	(40, 48)	(48, 48)	(48, 48)
識別率	0.6050	0.6775	0.6825	0.6325	0.6975	0.7200
感度	0.6600	0.7050	0.6050	0.7300	0.7700	0.6600
特異度	0.5500	0.6500	0.7600	0.5350	0.6250	0.7800
F 値	0.6256	0.6769	0.6805	0.6651	0.7179	0.7021

検証データに対するロスが減少しないエポックが 50 エポック連続するごとに学習率を半減させる. 150 エポック連続して検証データに対するロスが減少しなかった場合に, 最後に減少した際のモデルパラメーターを最良のモデルとしてテストデータに対する分類性能の評価を行う. モデルの汎化性能を向上させるために, データ拡張として水平および垂直方向の反転を確率 $p = 0.5$ でそれぞれ行い, コントラストと明るさ調整をランダムに係数 $[0.7; 1.3]$ の範囲で適用した. 学習, 検証, テストの被験者は各クラスそれぞれ 5; 1; 2 名ずつとした. 被験者の選び方をランダムに 100 回行い, 複数 ROI を用いた分類方法では, 各 ROI に対する出力を Voting Rule で統合して被験者単位での予測を行った.

3.5.2 評価指標

評価指標として, 第 1 章で定義した感度 (Sensitivity, または Recall), 特異度 (Specificity), 識別率 (Accuracy), Precision, F 値を使用した.

3.5.3 比較する手法

実験では, まずベースラインとして, MIL を用いない通常の教師あり学習を行い, テストデータのすべての ROI を被験者単位での統合に使用する. 提案手法としては, MIL による学習データの信頼性評価と REM によるテストデータの信頼性評価を用いることによる分類性能の向上を検討する. MIL による学習データの信頼性評価では, 設定する Drop rate を 2 クラス共通の設定とクラス別の設定で実験を行い, クラス別の適切な Drop rate の適用による分類性能向上を期待する. クラス別の Drop rate を用いる MIL と, ROI が分類されたクラス別の周辺領域を設定する REM の組み合わせでは, それぞれのパラメーターで精度向上に有効な組み合わせを調査する.

3.5.4 結果と考察

表3.2に比較手法の実験結果をまとめる．まず，ベースラインと比較して MIL による学習データの信頼性評価を行うことで，テストデータに対する分類性能が向上している．特に，2 クラス共通の Drop rate を用いた際の最高識別率 0:6775 に対して，クラス別 Drop rate (0:24; 0:06) を用いた際の識別率が 0:6825 と向上していることから，各クラスに適切な Drop rate を設定することの有効性がわかる．また，MIL を適用せず通常の教師あり学習を行った場合には，テストデータに対する REM を適用することで，第 2 章の Gabor 特徴と QDA と同様に分類性能の向上に有効であることを確認した．さらに，クラス別の Drop rate を用いた MIL と REM を組み合わせることで，比較手法の中でベースラインから最も良い分類性能の向上を確認した．

表3.3にクラス別の Drop rate を用いた際のテストデータに対する精度をまとめる．クラス別 Drop rate の組み合わせが (cir; nor) = (0:00; 0:48) に近づくにつれ，感度が増加し特異度が減少している．これは，健常者クラスの学習データを MIL によって多く棄却する設定であり，健常者クラスの特徴を適切に学習できなかった結果モデルの出力が肝硬変クラスに偏ったためである．反対に，(cir; nor) = (0:48; 0:00) に近づくにつれ，特異度が増加し感度が減少している．これは，肝硬変クラスの学習データを MIL によって多く棄却する設定であり，肝硬変クラスの特徴を適切に学習できなかった結果モデルの出力が健常者クラスに偏ったためである．感度と特異度の平均である識別率では，2 クラスに最もバランス良く学習ができた (cir; nor) = (0:24; 0:06) が識別率の最大値となった．ここで，Drop rate の値を比較すると，健常者クラスに設定する値が 0:06 に対して肝硬変クラスに設定する値が 0:24 と大きい値であるのは，肝硬変クラスの学習データ内でクラスラベルと一致しない健常な特徴を持つ ROI がノイズとして含まれていたことが原因であり，MIL による棄却を行うことで適切に学習用 ROI の選別が行えたと考える．一方で，Drop rate を高くしすぎると肝硬変クラスの特徴を持つ ROI も学習中に棄却してしまうため，高すぎない値が適切であった．超音波画像上の肝臓領域では肝硬変の領域が狭く存在しており，肝硬変疾患の肝臓にも健常な領域が広くみられると仮定している．そのため，肝硬変クラスに設定する Drop rate は一定量の棄却を行う値が適切であると想定される．肝硬変クラスの学習データとは異なり，健常者クラスの学習データでは肝硬変の特徴を持つような ROI の割合が少量であったため，小さい Drop rate の値が適切であった．

表3.4では，クラス別の Drop rate と REM を適用した際の識別率の良い設定を上位 20 パターン表示している．総じて肝硬変クラスに設定する Drop rate が健常者クラスに設定する Drop rate よりも高い値となっており，肝硬変クラスの学習データから MIL による一定量の ROI の棄却を行う有効性がわかる．さら

表 3.3: クラス別 Drop rate を用いた際のテストデータに対する精度.

(a): 識別率.

	nor									
Drop rate	0.00	0.06	0.12	0.18	0.24	0.30	0.36	0.42	0.48	
cir	0.00	0.6100	0.6250	0.6325	0.6150	0.6175	0.6300	0.6100	0.5875	0.5775
	0.06	0.6350	0.6700	0.6700	0.6600	0.6300	0.6275	0.6200	0.5950	0.5900
	0.12	0.6550	0.6475	0.6775	0.6650	0.6425	0.6375	0.6100	0.6450	0.6150
	0.18	0.6450	0.6500	0.6800	0.6750	0.6675	0.6600	0.6375	0.6250	0.6050
	0.24	0.6450	0.6825	0.6800	0.6450	0.6550	0.6700	0.6475	0.6325	0.6250
	0.30	0.6350	0.6675	0.6775	0.6525	0.6550	0.6575	0.6425	0.6400	0.6300
	0.36	0.6400	0.6450	0.6575	0.6750	0.6700	0.6675	0.6475	0.6375	0.6200
	0.42	0.6225	0.6350	0.6475	0.6575	0.6550	0.6500	0.6575	0.6475	0.6350
	0.48	0.6125	0.6425	0.6275	0.6225	0.6350	0.6425	0.6400	0.6475	0.6325

(b): 感度.

	nor									
Drop rate	0.00	0.06	0.12	0.18	0.24	0.30	0.36	0.42	0.48	
cir	0.00	0.6650	0.6450	0.7050	0.7150	0.7650	0.7950	0.8100	0.8300	0.8600
	0.06	0.6500	0.6600	0.7100	0.7600	0.7450	0.7800	0.8050	0.8350	0.8800
	0.12	0.6250	0.6300	0.7050	0.7100	0.7150	0.7550	0.7900	0.8550	0.8550
	0.18	0.6200	0.5850	0.6550	0.6750	0.7400	0.7550	0.7900	0.8050	0.8250
	0.24	0.5650	0.6050	0.6300	0.6000	0.6800	0.7100	0.7400	0.7900	0.8200
	0.30	0.4850	0.5500	0.5600	0.5700	0.6300	0.6800	0.7000	0.7350	0.7850
	0.36	0.4650	0.4800	0.4950	0.5600	0.5950	0.6350	0.6600	0.7250	0.7400
	0.42	0.4150	0.4150	0.4500	0.4850	0.5100	0.5650	0.6000	0.6650	0.7000
	0.48	0.3350	0.3800	0.3850	0.4000	0.4550	0.4800	0.5350	0.5800	0.6200

(c): 特異度.

		nor								
Drop rate		0.00	0.06	0.12	0.18	0.24	0.30	0.36	0.42	0.48
cir	0.00	0.5550	0.6050	0.5600	0.5150	0.4700	0.4650	0.4100	0.3450	0.2950
	0.06	0.6200	0.6800	0.6300	0.5600	0.5150	0.4750	0.4350	0.3550	0.3000
	0.12	0.6850	0.6650	0.6500	0.6200	0.5700	0.5200	0.4300	0.4350	0.3750
	0.18	0.6700	0.7150	0.7050	0.6750	0.5950	0.5650	0.4850	0.4450	0.3850
	0.24	0.7250	0.7600	0.7300	0.6900	0.6300	0.6300	0.5550	0.4750	0.4300
	0.30	0.7850	0.7850	0.7950	0.7350	0.6800	0.6350	0.5850	0.5450	0.4750
	0.36	0.8150	0.8100	0.8200	0.7900	0.7450	0.7000	0.6350	0.5500	0.5000
	0.42	0.8300	0.8550	0.8450	0.8300	0.8000	0.7350	0.7150	0.6300	0.5700
	0.48	0.8900	0.9050	0.8700	0.8450	0.8150	0.8050	0.7450	0.7150	0.6450

表 3.3: クラス別 Drop rate を用いた際のテストデータに対する精度 (続き).

(d): F 値.

		nor								
	Drop rate	0.00	0.06	0.12	0.18	0.24	0.30	0.36	0.42	0.48
cir	0.00	0.6303	0.6328	0.6577	0.6499	0.6667	0.6827	0.6754	0.6681	0.6708
	0.06	0.6404	0.6661	0.6833	0.6910	0.6683	0.6768	0.6792	0.6742	0.6829
	0.12	0.6442	0.6413	0.6864	0.6794	0.6673	0.6757	0.6702	0.7069	0.6900
	0.18	0.6349	0.6257	0.6716	0.6753	0.6903	0.6899	0.6859	0.6823	0.6764
	0.24	0.6138	0.6557	0.6628	0.6281	0.6640	0.6830	0.6776	0.6826	0.6864
	0.30	0.5705	0.6230	0.6336	0.6209	0.6468	0.6643	0.6624	0.6713	0.6798
	0.36	0.5636	0.5747	0.5911	0.6330	0.6430	0.6574	0.6521	0.6668	0.6612
	0.42	0.5238	0.5321	0.5607	0.5864	0.5967	0.6175	0.6377	0.6533	0.6583
	0.48	0.4637	0.5149	0.5083	0.5145	0.5541	0.5732	0.5992	0.6243	0.6277

表 3.4: 実験結果一覧.

Drop rate (cir)	Drop rate (nor)	周辺領域 (cir)	周辺領域 (nor)	識別率	感度	特異度	F 値
0.30	0.12	48	48	0.7200	0.6600	0.7800	0.7021
0.36	0.12	32	48	0.7100	0.6450	0.7750	0.6898
0.24	0.06	48	48	0.7075	0.6700	0.7450	0.6961
0.18	0.12	48	48	0.7050	0.7250	0.6850	0.7108
0.30	0.06	32	48	0.7050	0.6650	0.7450	0.6927
0.24	0.06	40	48	0.7050	0.6850	0.7250	0.6990
0.24	0.12	32	48	0.7025	0.7500	0.6550	0.7160
0.18	0.12	40	48	0.7025	0.7400	0.6650	0.7133
0.30	0.12	24	48	0.7025	0.7250	0.6800	0.7090
0.24	0.18	48	48	0.7025	0.7150	0.6900	0.7062
0.36	0.12	24	48	0.7025	0.7100	0.6950	0.7047
0.24	0.12	48	48	0.7025	0.7000	0.7050	0.7018
0.24	0.06	32	48	0.7025	0.6900	0.7150	0.6987
0.36	0.24	48	48	0.7025	0.6800	0.7250	0.6957
0.30	0.06	40	48	0.7000	0.6400	0.7600	0.6809
0.30	0.24	48	48	0.7000	0.7300	0.6700	0.7087
0.36	0.18	48	48	0.7000	0.6300	0.7700	0.6774
0.42	0.18	32	48	0.7000	0.6250	0.7750	0.6757
0.24	0.24	48	48	0.6975	0.7700	0.6250	0.7179
0.18	0.12	32	48	0.6975	0.7500	0.6450	0.7126

に、REM の周辺領域では健常者クラスに分類された ROI には常に $dy; dx = 48$ とし、肝硬変クラスに分類された ROI には $24; 32; 40$ の値も設定されている。MIL によって肝硬変クラスの学習データがある程度棄却され、肝硬変クラスとして学習された ROI がテストデータの超音波画像上で一定程度局所的に存在したと考えられる。一方で、健常者クラスの学習データは MIL によって棄却される ROI の割合が肝硬変クラスより少なく、テストデータの超音波画像上でも大域的に存在した結果、周辺領域が広く設定されたと考えられる。

3.5.5 制約

Drop rate 最適化

本章の実験では、MIL における Drop rate と REM における周辺領域パラメーターは恣意的に選択を行い、性能評価を行った。一方で、これらのハイパーパラメーターは実験内で検証データを用意し、自動最適化を行うことが実応用の場面では求められる。その方法としては、第 2 章での周辺領域自動最適化と同様に、ハイパーパラメーターの探索の際に検証データに対するロス指標とし、全パラメーターの中で最もロスを最小化するパラメーターを選択する方法がある。

第 2 章 Gabor 特徴と QDA による肝硬変検出との比較

第 2 章の肝硬変検出で使用した、Gabor 特徴と QDA による分類性能と比較して、本章での分類性能は劣る結果となった。この原因として、畳み込み層による特徴抽出を行うことによる、表現力の過剰がある。CNN の畳み込み層では、毎エポックフィルタの重みを更新し、学習データに対するロスが最小となるように最適化される。また、複雑な画像表現を学習するために多層の畳み込み構造であるため、超音波画像上の肝硬変検出には不要な超音波ノイズに対しても敏感に適合してしまう。一方で、Gabor 特徴による特徴抽出では、あらかじめ設定された周波数などによって抽出される特徴が決定されており、学習データに対して過適合するリスクが小さい。また、特徴抽出を行う畳み込み操作が段階であるため、複雑な特徴を抽出しづらい。加えて、QDA は特徴空間における分類境界を二次曲面に制約することで、データへの過適合を抑制している。第 2 章と第 3 章で使用した超音波画像はノイズによる影響を受けやすく、特に今回のデータセットサイズでは CNN による学習での汎化性能を実現することが困難であった。一方で、さらに大規模な超音波画像データセットを活用できる場面では、本章で提案した手法が有効である可能性がある。

3.6 まとめ

本章では，第 2 章で超音波画像を用いた肝硬変検出に対して有効性が確認された REM の適用を CNN モデルへ拡張し，人手に依存しない分類器での有効性を検討した．また，従来手法として提案されていた MIL フレームワークと REM を統合することで，特徴空間における信頼性評価と，画像空間における信頼性評価の統合による分類性能の向上を検討した．予備実験では，学習データへの MIL と REM の適用を統合する場合に，学習データの不足による分類性能の低下が明らかになった．一方で，実験の節において，学習データに MIL をクラス別の Drop rate を用いて適用し，テストデータには REM を適用することで分類性能の向上を確認した．特に，肝硬変クラスの ROI には健常な領域が含まれていると想定され，Drop rate の値を大きくする一方で，健常者クラスの ROI には肝硬変の領域が含まれる割合が少ないため Drop rate の値を小さくすることが有効であった．また，テストデータで肝硬変クラスに分類された ROI に対しては REM で周辺領域を狭く設定し，健常者クラスに分類された ROI に対しては周辺領域を広く設定すると分類性能が向上した．健常な肝臓表面では健常な領域が広くみられる一方で，肝硬変クラスの肝臓表面では局所的に肝硬変の所見が見られる仮定と一致する結果となり，分類性能の向上による提案手法の有効性を確認した．

第 4 章

複数解像度画像における共変量シフト 適応を用いたコンクリートひび割れ抽 出の高精度化

インフラ構造物のコンクリート表面におけるひび割れ調査自動化のために、深層学習ベースの手法を提案する。一般的に機械学習モデルは現場での運用時に想定されるデータの分布がモデル学習時のデータと同一であることを前提するが、共変量シフト環境では想定される運用時のデータに対する適切な学習を行えず、ひび割れ抽出の精度が低下する。本章では、1 ステップ共変量シフト適応をひび割れ抽出に適応し、低解像度画像に偏った現場での運用時のデータを扱う場合において、通常の教師あり学習と比較してひび割れ抽出精度の向上を示す。

4.1 はじめに

日本を含む先進国では、約 50 年前に建設された構造物が現在その耐用年数を迎えつつある。これらの構造物のコンクリート表面におけるひび割れ調査は、建築物やインフラ設備の構造的安全性を維持する上で極めて重要である。人力による調査が一般的ではあるが、点検対象の数は急速に拡大しており、利用可能な労働力の供給を上回っている。

目視での点検の効率を向上させるために、ひび割れ抽出に関する多くの深層学習 (Deep Learning) ベースの手法が提案されている。これらのモデルは、大規模な学習データと膨大な計算資源を活用できるため、効率・精度の両面で人間によるひび割れ調査のパフォーマンスを上回ることが多い。

一般的にこれらのモデルは、学習データ上の損失を最小化する教師あり学習によって訓練されており、現場での運用時のデータの分布が学習時のデータと

同一であることを前提としている。しかし、学習時のデータと現場での運用時のデータの間にドメイン／分布のずれが存在すると、その性能が低下することが知られている。このようなドメインシフトの一種として頻繁に見られるのが共変量シフト（Covariate Shift）であり、これは条件付きラベル分布が変化しない一方、入力となる特徴量の周辺分布がモデル学習時のデータと現場での運用時のデータ間で変化する状況を指す。

本論文では、画像解像度の違いによって生じる共変量シフト、複数解像度適応（Multi-Resolution Adaptation, MRA）に注目し、このような非定常性の下でモデルのロバスト性を向上させるための共変量シフト適応を適用する。解像度の変化のうち、検証実験では学習時の画像データが比較的高解像度である一方、運用時には低解像度の撮影装置を使用するというシナリオを扱う。

提案するフレームワークは、セグメンテーションモデルと重要度推定モデルから構成される。セグメンテーションモデルは画素単位の予測損失を最小化するように訓練され、テスト画像に対するひび割れ予測に用いられる。重要度推定モデルは、各学習画像がテストドメインに出現する可能性を推定し、それに基づいてセグメンテーションモデルのロスに対して重みを割り当てることで、重要度に基づくセグメンテーションモデルの学習を可能にする。

以下に本章における貢献を述べる。

- ひび割れ抽出における MRA: 本章では、現場での運用時に想定されるデータの入力分布に基づいて学習時の各画像データの重要度を推定できる MRA を提案する。これにより、ひび割れセグメンテーションモデルは、重要度の高い訓練画像を優先して学習することができる。
- ひび割れセグメンテーション精度の向上: 実験において、提案手法は Intersection over Union (IoU) によって測定されるひび割れセグメンテーション性能の向上を示す。特に、低解像度画像が学習データ内に少量しか含まれていない場合に、その改善効果が顕著である。
- CAM による注目領域の解析: Class Activation Mapping (CAM) を用いて、重要度推定モデルが注目した領域を可視化する。可視化結果から、モデルはひび割れそのものよりも背景クラスに注目することで解像度の違いを識別していることを明らかにする。

4.2 先行研究

4.2.1 深層学習によるひび割れ抽出

深層学習を活用した多数の手法が、複雑な画像特徴を自動で学習できる点で優れており、これまでひび割れ抽出に応用されてきた。一般的に、セグメンテーションモデルは予測結果と対応するマスク画像との損失を最小化するように学習する。Li ら [65] はひび割れクラスが背景クラスと比較して少数であるクラス不均衡問題を扱い、クロスエントロピーを改良した損失関数を提案した。Zou ら [66] が提案した DeepCrack は、SegNet[67] をベースとしたセグメンテーションモデルであり、畳み込み層を通じて多様なスケールでひび割れの特徴を抽出できる構造を持つ。Skip-layer Fusion と呼ばれるスキップレイヤーの統合により、中間層でも損失関数を適用して学習を効率的に行っている。他方、画像取得や伝送に起因するノイズについても検討した研究がある [68]。結果では、AlexNet[51] が計算コストの面で最も効率的でありつつ、高精度を維持できるモデルであることを確認した。

4.2.2 ひび割れ抽出におけるドメイン適応

前述の研究のほかに、より現実的な状況を考慮してひび割れセグメンテーションでのドメイン適応に焦点を当てた研究が数多く行われている。これは、学習データとテストデータのドメイン (特徴量/ 分布) 変化に対応する手法である。先行研究 [69] では、ひび割れセグメンテーションに影響を与える様々な要因について検討しており、ラベル付けの解像度 (画素レベル対画像レベル)、ひび割れの種類、明るさ条件の差異などをモデル内の中間ドメインで緩和することで、セグメンテーション精度向上を実現した。近年の研究では、UCrack-DA[70] が階層的かつ敵対的構造でのエントロピー最小化を行うマルチスケール・教師なしドメイン適応手法を提案した。Mix-Transformer エンコーダー、マルチスケール膨張アテンションモジュール、混合畳み込みアテンションデコーダーを組み合わせることで、異なるひび割れデータセット間のドメインシフトを効果的に克服し、UAV (無人航空機) ベースおよびベンチマークデータセットの両方で優れた性能を達成した。さらに、Shan ら [71] は CDCR-ISeg と呼ばれるフレームワークを提案した。これは、UAV をベースとしたひび割れ抽出におけるドメイン間および解像度間の変化に対応するために、超解像とドメイン適応を統合したフレームワークである。著者らは複数解像度データセット (UAV-CrackX4/ X8/ X16) を構築し、ひび割れエッジ認識ベクトルマップと専用の教師信号を組み合わせることで、一般化能力を大幅に向上させ、アノテーションコストを削減し、多様な UAV 撮影条件下でも高精度なひび割れセグメンテーションを実現した。

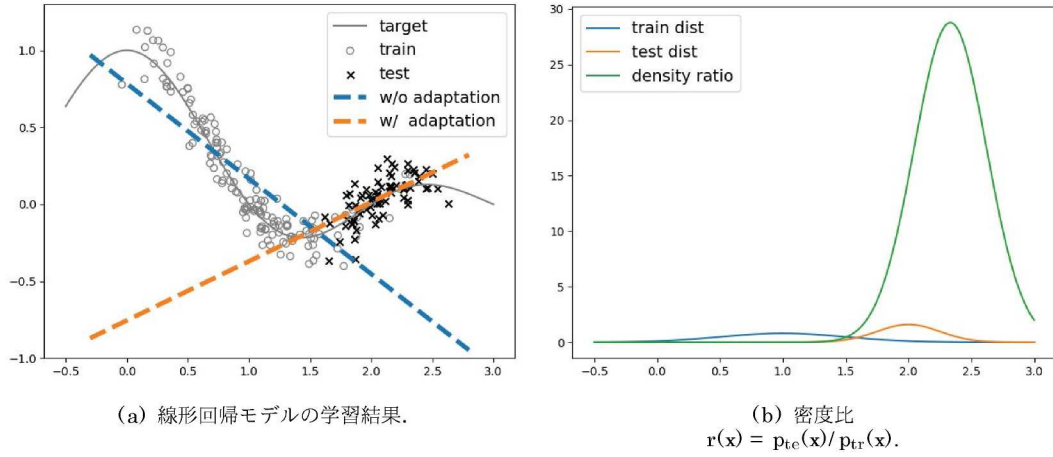


図 4.1: (a) 共変量シフト適応を用いた場合 (オレンジ) と用いない場合 (青) の、線形回帰モデルの学習結果を示す. (b) 学習データとテストデータの入力分布および密度比を示す. 学習データの各サンプルに対する重要度は、密度比 $r(x) = p_{te}(x)/p_{tr}(x)$ で与えられる. (a) におけるオレンジの線形回帰は密度比を用いた重要度重みづけによってロス関数への重みづけを行った結果である. 本予備実験の詳細は第 4 章付録 に記載する.

これらの研究は、複雑なモデル、敵対的学習フレームワーク、ひび割れに特化したデータセットなどを活用することで、高精度なひび割れのセグメンテーションを達成している. しかし、これらの手法を運用・維持するには、一定の機械学習の知識が必要となる. 機械学習技術の導入容易性を考慮すると、より単純な手法でも実用上有用となる可能性がある.

4.2.3 共変量シフト適応のための重要度重みづけ

共変量シフト (Covariate shift) とは、学習データとテストデータの間で、入力データ (特徴量) に対するラベルの条件付き分布が変化しない一方で、入力の周辺分布が異なる状況を指す. すなわち、 $p_{tr}(y|x) = p_{te}(y|x)$ および $p_{tr}(x) \neq p_{te}(x)$ が成り立つ. さらに学習データの分布のサポート (定義域) がテストデータの分布のサポートを包含していること、つまり $\text{supp}(p_{te}(x)) \subseteq \text{supp}(p_{tr}(x))$ を仮定する. これらの条件によって、重みづけに基づく適応が可能となる.

重要度重みづけは、共変量シフト適応のための最もシンプルな手法の一つであり、Shimodaira [72] を代表的な起源とする. 重要度重みづけの基本的な考え方は、学習データのサンプルを使用して得られる損失に、推定された密度比 $r(x) = p_{te}(x)/p_{tr}(x)$ に基づく重みづけを行い、モデルが学習時により重要なサンプルに対する学習を重視できるようにする点にある. 図4.1は、共変量シフト下での回帰問題における重要度重み付けの例を示している. ここでは条件付き分布が $p_{tr}(y|x) = p_{te}(y|x)$ で変化しない一方、周辺分布が異なり $p_{tr}(x) \neq p_{te}(x)$ となる. このとき、密度比 $r(x) = p_{te}(x)/p_{tr}(x)$ によって、学習データ内で注目すべきサンプルにより大きい重みが付与される. 重要度重み付けによる適応を用

いて学習した回帰モデルは、適応を行わない場合に比べて、Target function をより良く近似することができている。

多くの共変量シフト適応手法は 2 ステップのアプローチをとる。1 ステップ目では密度比をモデリングし、2 ステップ目ではモデル化した密度比を重要度重みとして予測モデルの学習に使用する。一般的なロス関数は以下で表される。

$$\begin{aligned} E_{(\mathbf{x}^{\text{te}}; \mathbf{y}^{\text{te}}) \sim p_{\text{te}}} [\ell(f(\mathbf{x}^{\text{te}}; \mathbf{y}^{\text{te}}))] \\ = E_{(\mathbf{x}^{\text{tr}}; \mathbf{y}^{\text{tr}}) \sim p_{\text{tr}}} [\ell(f(\mathbf{x}^{\text{tr}}; \mathbf{y}^{\text{tr}})) r(\mathbf{x}^{\text{tr}})]: \end{aligned} \quad (4.1)$$

式(4.1)より、推定された損失の値は普遍推定量であることが分かる。つまり、重要度重み付けにおける経験的損失を最小化することで、テストデータの分布における真の損失値を、テストデータとは異なる分布を持つ学習データから得られるサンプルを用いて実現できる。

重要度重み付けは共変量シフト適応として非常に有用な手法である一方で、密度比推定において推定結果のばらつきによるバリエーションの高さが推定された重みを用いた予測モデルの学習を不安定にするため課題となる。極端に偏った重みによる学習への影響を抑制するために、重み付けされた損失値を指数関数的に平坦化する手法が Shimodaira [72] で提案されている。その後、山田ら [73] は Relative Density Ratio Estimation を提案し、限られたサンプル数でもより滑らかに安定し重要度推定を実現した。

これらの手法により、予測モデルの学習における密度比推定の推定値変動の影響は抑制された。一方で、2 ステップアプローチでの密度比推定モデルの under fitting(学習不足) は、推定された密度比に基づく学習を行う予測モデルの予測性能の限界を低下させる原因となる。近年、Zhang ら [74] は共変量シフトの 1 ステップアプローチを提案した。提案手法では、密度比推定モデルと予測モデルの学習を以下に示す 1 つの統合したフレームワークで行うことで、2 ステップアプローチでボトルネックとなっていた密度比推定モデルの学習不足による影響を克服し、予測モデルの性能向上を実現した。

$$\begin{aligned} L(f; g; S) = & \left(\frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} \ell(f(\mathbf{x}_i^{\text{tr}}; \mathbf{y}_i^{\text{tr}})) g(\mathbf{x}_i^{\text{tr}}) \right)^2 \\ & + m^2 \left(\frac{1}{n_{\text{tr}}} \sum_{i=1}^{n_{\text{tr}}} g^2(\mathbf{x}_i^{\text{tr}}) - \frac{2}{n_{\text{te}}} \sum_{i=1}^{n_{\text{te}}} g(\mathbf{x}_i^{\text{te}}) \right) \end{aligned} \quad (4.2)$$

ここで、 $S = \{(\mathbf{x}_i^{\text{tr}}; \mathbf{y}_i^{\text{tr}})\}_{i=1}^{n_{\text{tr}}} \cup \{\mathbf{x}_i^{\text{te}}\}_{i=1}^{n_{\text{te}}}$ はラベル付き学習データとラベルなしテスト/検証データの和集合を表す。式4.3を最適化することで、テストデータの入力分布 p_{te} と重なっている入力の学習データ $\mathbf{x}_i^{\text{tr}}$ に対して $g(\mathbf{x}_i^{\text{tr}})$ の出力が大きく

なり，予測モデルの学習を促進させるため，テストデータ $g(x_i^{te})$ に対する汎化性能を向上させることができる．

これまで様々な，複雑で強力な深層学習モデルの手法がひび割れセグメンテーションや 2 ステップ共変量シフト適応に対して提案されてきた．しかしながら，ひび割れ抽出のための 1 ステップ共変量シフト適応，特に画像解像度に注目した手法は研究の余地が大きくある．

4.3 提案手法: 1 ステップ複数解像度適応 (MRA)

本章では，1 ステップ共変量シフト適応をひび割れ抽出に適用し，複数解像度の設定における学習時のデータと現場での運用時のデータの分布変化による予測モデル分類性能の劣化抑制を目的とする．これ以降，提案手法と実験における学習時のひび割れ画像を学習データ，現場での運用時のデータをテストデータという想定のもとで記述する．また，本章では共変量シフトとして以下を仮定する． $p_{tr}(y|x) = p_{te}(y|x)$ ， $p_{tr}(x) \neq p_{te}(x)$ ， $\text{supp}(p_{te}(x)) \subseteq \text{supp}(p_{tr}(x))$ ，ここで $p_{tr}(x)$ と $p_{te}(x)$ は学習データとテストデータの入力画像データ分布を示し， $p_{tr}(y|x)$ と $p_{te}(y|x)$ は各画素のひび割れクラス確率を学習データドメインとテストデータドメインでそれぞれ表す． $p_{tr}(x)$ と $p_{te}(x)$ 間の差が，入力画像の解像度の差異を表す．

さらに，重要度推定モデルの中間層で注目された領域を可視化により分析し，どのような画像特徴が正確な重要度推定に必要であるか確認する．

4.3.1 モデル構造

予測モデル

本章では，U-Net[75] を使用する．U-Net は，スキップ接続を組み込んだ畳み込み型のエンコーダ-デコーダネットワークであり，精密な出力の位置特定を可能にする構造を持つ．学習中の安定化を目的として，各畳み込み層の後に Layer Normalization[76] を適用する．

重要度推定モデル

本章では，軽量の畳み込み深層学習モデルを重要度推定モデルとして用いる．このモデルは 7 つの畳み込み層で構成されており，4 つのステージに分かれている．それぞれのステージは，順に 1 層，2 層，2 層，2 層である．各層の後には，Group Normalization[77] と ReLU 活性化関数 [78] を適用する．最初の畳み込み層のみ，より広い領域から特徴を抽出するためにフィルタサイズ 7×7

を用い、その他の層では計算負荷を抑えるために 3×3 のフィルタを使用する。ダウンサンプリングは第 2 および第 3 ステージでの max pooling によって行われ、これによりコンパクトな特徴表現が得られる。最終ステージでは、Global Average Pooling (GAP) を適用して空間情報をチャンネル単位のベクトルに集約し、これを全結合層に入力して単一のスカラー重要度重みを予測する。予測出力は Sigmoid 関数によって正規化し、さらにあらかじめ定義された最大値係数によってスケールする。また、最後の畳み込みステージの中間特徴マップを Class Activation Map (CAM)[79] として利用することで、重要度推定に貢献する領域の可視化を可能にしている。このアーキテクチャにより、共変量シフト下でも安定し解釈可能な重要度推定が実現される。

4.3.2 学習フレームワーク

Zhang ら 1 ステップ共変量シフト適応 [74] をベースとし、予測モデル f と重要度推定モデル g は以下に示す損失関数を最小化し、交互にパラメーター更新を行う。

$$L(f; g; S) = \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \ell(f(x_i^{tr}); y_i^{tr}) g(x_i^{tr}) + \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} g^2(x_i^{tr}) - \frac{2}{n_{tc}} \sum_{i=1}^{n_{tc}} g(x_i^{te}); \quad (4.3)$$

ここで、 $S = \{(x_i^{tr}; y_i^{tr})\}_{i=1}^{n_{tr}} \cup \{x_i^{te}\}_{i=1}^{n_{te}}$ はラベル付き学習データとラベルなしテスト/検証データの和集合を表す。式(4.1)との違いは、まず $m^2 = 1$ として、学習量の調整を学習率でのみ行う点である。また、第 1 項に付与された 2 乗についても外し、学習を単純化している。損失関数 $\ell(f(x); y)$ はクロスエントロピーを使用する。本章では、予測モデルとして U-Net を使用し、重要度推定モデルとして前述の 7 層畳み込みニューラルネットワークを使用する。最適化手法には確率的勾配降下法 (Stochastic Gradient Descent, SGD) の一つである Adam[64] を使用する。U-Net と重要度推定モデルの学習は 200 エポックとする。U-Net は学習率を 0.01 から始め、50 エポックごとに半減させる。重要度推定モデルは学習率を 0.0001 とし変化させない。これらの学習率は学習のバランスを調整するために、経験的に選択されている。学習において、テストデータもしくは検証データはラベルなしデータとして使用する。そのため、各画像のラベル情報は必要なく、 $f(x)$ と $g(x)$ の学習には画像特徴のみ使用しラベル付け作業は生じない。

4.4 実験

この節では、複数解像度におけるひび割れ抽出のための 1 ステップ共変量シフト適応の有効性を検証する。

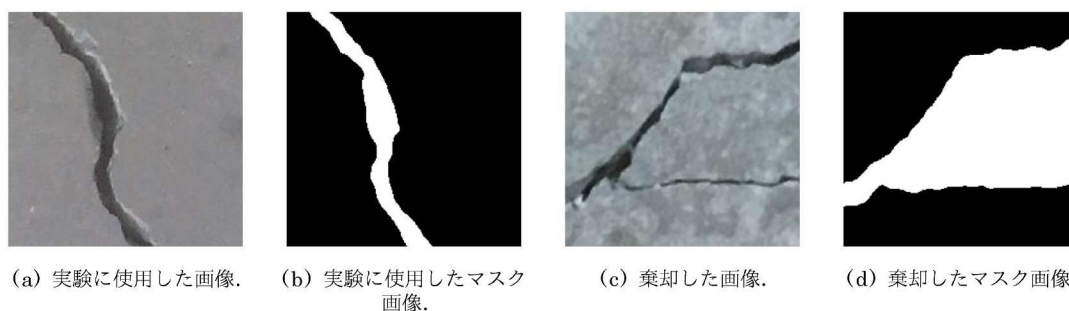


図 4.2: 本章で実験に使用した画像とひび割れ画素数の条件により棄却した画像, およびそれらに対応するマスク画像の例. (a)–(b): 使用した画像と対応するマスク画像, (c)–(d): 棄却した画像と対応するマスク画像.

4.4.1 データセット

実験では, 公開されているひび割れマスク付きのデータセット [80] を使用する. 画像はコンクリート構造物の表面を撮影し, ひび割れ画素を含む. ラベルはひび割れと背景の 2 クラスが用意され, 画素単位でマスク画像が存在する. 本データセットは 458 枚の大きいサイズ (3024×4032 画素 438 枚, 2448×3264 画素 20 枚) の画像で構成される. 用意されているマスク画像は, ラベル付けがあいまいな領域を含むことと, ひび割れクラスの画素数が少ないことから, 単純な学習では学習したモデルの分類性能が低下する. そのような影響を抑制するため, 各画像を 224×224 のパッチ画像に分割し, ひび割れ画素を 500 以上, 全画素の $1/4$ ($(224 \times 224)/4 = 12,544$) 以下の条件で含むパッチ画像のみ実験に使用した. 実験に使用したパッチ画像と棄却したパッチ画像の例を図4.2に示す. このパッチ画像の使用条件は, 経験的に決定しており他のデータセットやモデル使用環境では留意する必要がある. データセットは学習, 検証, テストの 3 つに分割した. まず全体のランダムに 20% をテストデータとし, モデルの性能評価に使用する. 残り 80% のうち, 再度ランダムに 20% を検証データとし, 残りのデータを学習データとする. このデータセット作成の設定により, 学習データセットは 8,080 枚, 検証データセットは 2,020 枚, テストデータセットは 2,525 枚のパッチ画像でそれぞれ構成される.

4.4.2 複数解像度作成方法

図4.3は, 高解像度画像 (元画像) と低解像度画像 (作成された画像) の例を示す. 低解像度画像はダウンスケーリングとアップスケーリングを順に行うことで作成され, 元画像と同じ画像サイズとなる. ダウンスケーリングにおける圧縮倍率は $\{1/1, 1/10, 1/20\}$ から選択され, 作成される低解像度画像の解像度を決定する. ここで, $1/1$ は元の高解像度画像を指し, ダウンスケーリングを行わない. $1/10$ と $1/20$ はそれぞれ, 画像を元画像の $1/10$ と $1/20$ のサイズまで圧

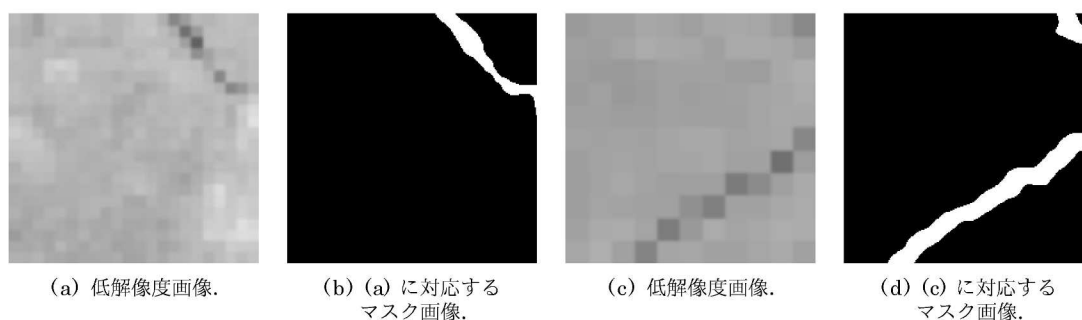


図 4.3: ダウンスケーリングとアップスケーリングにより作成された低解像度画像および、対応するマスク画像の例.

縮する. ダウンスケーリングとアップスケーリングの後, Gaussian blur(フィルタサイズ 7×7 , 標準偏差 1:4) を適用し, アップスケーリングによって生じるエッジを抑制する.

本手法で低解像度画像を作成する場合, 画像撮影時のセンサレベルで発生するノイズが平均化されるため, 実際の低解像度画像の再現には限界がある. 一方で, 問題設定としてセンサレベルでの低解像度画像の作成よりも, 解像度の違いが入力分布に与える変化とモデルの性能低下を疑似的に再現することが目的であり, この方法で問題ないと判断している.

本章での注目するシナリオは, ひび割れ発生の現場で撮影される画像が信頼度高く高解像度の画像を撮影できない場面を想定し, テストデータの多くは低解像度画像である一方で, 学習データでは高品質な高解像度画像が多く使用できる場面を想定する. ここで, 低解像度画像は高解像度画像から画像処理によって疑似的にではあるが簡単に作成可能であることから, テスト時に想定される入力分布を手元の高解像度画像で再現できる.

4.4.3 評価指標

評価指標として, 第 1 章で定義した IoU を使用した.

4.5 結果と考察

4.5.1 IoU 結果

表4.1に 5 回実験を行った際の平均の IoU の値をまとめる. ここで, "Baseline" とは通常の教師あり学習を指し, "Proposed" は MRA による学習を指す. 上 2 段の行は 90% の高解像度画像と, 5% の圧縮比率 $1/10$ と $1/20$ の低解像度画像で学習データが構成される. 下 2 段の行は 98% の高解像度画像と, 1% の圧縮比率 $1/10$ と $1/20$ の低解像度画像で学習データが構成される. これらの学習デ

表 4.1: 複数解像度画像に対するひび割れ抽出 IoU 精度の比較. 学習・テストデータにおける各圧縮比率 (1/1, 1/10, 1/20) の構成割合をパーセンテージで示し, 各設定においてスコアが高かった値を太字で示す.

Train comp. (%)			Test comp. (%)			Method	IoU (%)				
1/1	1/10	1/20	1/1	1/10	1/20		1/1	1/10	1/20	Mean	Weighted
90	5	5	10	30	60	Baseline	83.64	76.89	67.91	72.03	72.18
						Proposed	83.09	77.84	68.49	72.61	72.76
90	5	5	10	60	30	Baseline	83.54	76.67	67.98	74.65	74.76
						Proposed	82.88	77.78	68.29	75.34	75.44
98	1	1	10	30	60	Baseline	83.80	73.19	62.84	67.78	68.05
						Proposed	83.02	75.91	65.62	70.28	70.45
98	1	1	10	60	30	Baseline	83.82	73.13	62.88	70.94	71.13
						Proposed	83.21	75.79	65.63	73.35	73.49

ータの構成において, テストデータでは 2 つのパターン, 10% の高解像度画像と 30% の圧縮比率 1/10 と 60% の 1/20 の低解像度画像, 圧縮比率 1/10 と 1/20 の割合を入れ替えた場合を用意した.

4.5.2 学習データ内の低解像度画像が 5% の場合

表4.1から確認できるように, Baseline と比較して提案手法は高解像度画像に対して IoU スコアを低下させた一方で, 平均 IoU と重み付き平均 IoU の値で微少な改善があった. 低解像度画像に対する改善が限定的だったのは, U-Net のモデル構造がシンプルながらも複雑な画像特徴を学習することができ, 少量の低解像度画像に対しても適切に学習が行えた点がある.

4.5.3 学習データ内の低解像度画像が 1% の場合

学習データ内に低解像度画像が 1% しかない状況において, 提案手法は平均 IoU と重み付き平均 IoU を改善しており, 提案手法の有効性が確認できる.

4.5.4 重要度推定の結果

重要度推定モデルの学習後に出力した重要度推定値を表4.2にまとめる. 全体的に推定された重要度は高解像度と低解像度の画像を区別できており, 学習データ内の重要サンプルに対して高い重要度と重要でないサンプルに対して低い重要度を付与できていることが分かる.

一方で, 推定された重要度は Shimodaira [72] で提案された共変量シフト適応のためのオリジナルの密度比 $r(\mathbf{x}) = p_{te}(\mathbf{x})/p_{tr}(\mathbf{x})$ とは異なることに注意する必要がある. 要因として, Shimodaira [72] で理論対象としたモデルはパラメーターに対して線形であり, パラメーター空間内での大域的最適解を保証できる一

表 4.2: 提案手法における各解像度に対する重要度推定結果 (Estimated importance). 各構成割合における 1 ステップ共変量シフト適応で算出された重み係数を示す.

Train comp. (%)			Test comp. (%)			Estimated importance		
1/ 1	1/ 10	1/ 20	1/ 1	1/ 10	1/ 20	1/ 1	1/ 10	1/ 20
90	5	5	10	30	60	0.369	6.879	8.190
90	5	5	10	60	30	0.341	7.935	7.560
98	1	1	10	30	60	0.124	6.866	6.870
98	1	1	10	60	30	0.127	6.871	6.871

方で、本章のような SGD による深層学習モデルの学習はパラメーター空間での大域的最適解を保証せず局所的最適解に陥る可能性がある。加えて、重要度推定モデルの出力層には Sigmoid 活性化関数を使用して出力の範囲を限定して極端な値を回避することで学習を安定させていた。これによって重要度推定モデル自体とその出力値がロスの係数となる U-Net の 2 つのモデルの学習を安定させることができるが、出力の範囲を限定しているため正確な重要度推定は困難になる。一方で、深層学習モデルは線形モデルと異なりモデルが学習できる表現力の高さから出力値が極端な値に収束することがあるため、深層学習モデルによる複雑な表現を学習した密度比推定と出力値の安定の両立は課題である。

4.5.5 CAM の可視化結果

図4.4に、CAM の可視化結果を示す。各画像に対する重要度推定は解像度に従って正しく行われていることが分かる。さらに、低解像度画像に対して CAM が強く反応しており、ひび割れ領域を除いた背景クラスの領域を参照して重要度推定を行っていることが分かる。このことから、CAM は背景クラスの領域を参照して中間層で活性化しないことで高解像度画像を重要でないサンプルと判断していることが分かる。これらの結果から、ひび割れ領域の有無は重要度推定モデルの推定性能には影響を与えず、どのようなひび割れの形状やデータセット内のクラスバランスでも背景クラスの領域を参照することで高い精度で重要度推定が可能であるといえる。

4.5.6 本章の課題

式(4.3)では、1 ステップ共変量シフト適応はテストデータのラベル情報をモデルの学習に必要としない一方で、学習過程でのテストデータに対しての予測精度や重要度推定の精度を監視して汎化性能や過学習および学習不足を確認するためには、教師あり学習の場合と同様に、ラベル付き検証データを用意する必要がある。実際の実装現場では、少量でもラベル付き検証データがテスト時

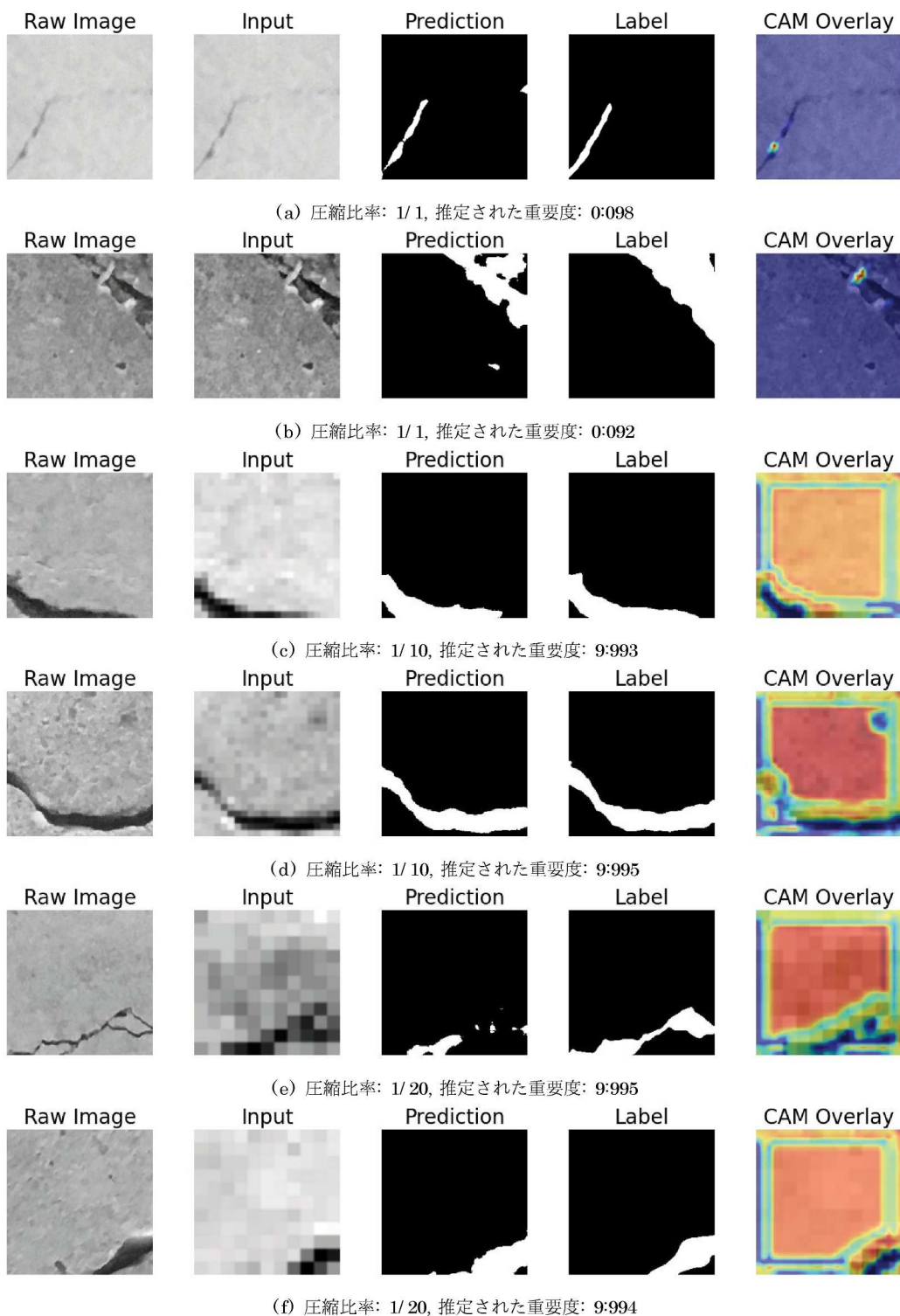


図 4.4: 各圧縮比率 1/1; 1/10; 1/20 における元画像, 入力画像, U-Net 予測出力画像, マスク画像, 重要度推定モデルの CAM 可視化画像の例.

の環境から収集できていることが機械学習を活用したシステム全体の信頼性向上のために非常に重要である。U-Net および重要度推定モデルの学習率の選択も課題となる。基本的に、重要度推定モデルによって推定された重みは U-Net の学習に直接影響を与える。そこで本章では、重要度推定モデルの学習率を微小な値で固定し、U-Net の学習を開始することで重要度推定モデルの学習の不安定性に起因する U-Net の学習の不安定性を抑制した。また、U-Net の画素単位での予測は重要度推定よりも複雑であると考えられるため、学習の中で局所的最適解に陥るのを抑制するために一定の大きさの学習率から学習を始めることが有効である。本章では、式(4.3)において m^2 が第 2 項の学習スケールを調整するが、重要度推定モデルの学習率が同様の役割を担うため、 $m^2 = 1$ とした。

ハイパーパラメータの設定のほか、計算資源の観点では、単一モデルを使用したひび割れ抽出と比較して 1 ステップ共変量シフト適応は学習時に重要度推定モデルを追加することによる計算資源消費量の増加がある。一方で、ひび割れ抽出モデルの学習を行った後の現場での活用では、重要度推定モデルは基本的に必要でないため単一モデルでの学習時と同様の計算資源で活用することができる。

結果については、IoU による定量的な評価で提案手法の有効性を評価しているが、画素単位での評価ではひび割れ形状を考慮した評価は改善余地がある。実応用を考えると、ひび割れの幅や連続性を考慮してペナルティが大きい評価指標を使用することで、コンクリート構造物の維持管理に適した評価を行うことが可能になる。また、予測モデルの学習に使用できる低解像度画像の枚数が一定の場合、高解像度画像を追加で学習に使用することの影響について、追加の検討が必要である。高解像度画像からも低解像度画像のひび割れ抽出に有効な特徴が一部学習可能である場合、運用時に想定される解像度に限定されない学習データの収集によってひび割れ抽出の精度向上の可能性がある。

4.6 まとめ

本章では、ひび割れ調査のための機械学習によるひび割れ抽出モデル学習において、モデル学習時と現場での運用時のデータの分布が変化する問題を共変量シフトとして扱った。特に、現場で使用できる画像撮影用カメラの性能に制限があり低解像度画像に偏った場合を想定した。複数解像度による共変量シフトへの適応として、1 ステップ共変量シフト適応をひび割れ抽出モデルに適用した。提案手法はセグメンテーションモデルと重要度推定モデルのパラメータを交互に更新して学習する。また、CAM を用いて重要度推定に重要である領域の可視化を行った。提案手法の有効性を公開されているひび割れデータセットを用いて人工的に低解像度画像を作成し検証した。本章での実験は単一のデ

ータセットで実装と評価が行われているため、今後は異なる地域やインフラ構造物を対象とした複数のデータセットを用いた提案手法の有効性検証が必要である。また、本章では重要度推定を学習データ内の各画像単位で行っていたが、画素単位での重要度推定を行うことで、画像内の影やコンクリート表面の欠損など画像内の領域を限定した重要度の推定による共変量シフト適応の活用も研究の方向性の一つである。その際は、解像度の変化だけでなく影による画素値の分布変化や、撮影角度の変化によるひび割れの見え方の変化など、様々な共変量シフトへの適応が可能になると期待される。

第 5 章

結論

この章では、本論文の総括と今後の展望についてまとめる。

5.1 総括

本論文では、限られたデータを用いた機械学習の適用法と有効性評価を行った。機械学習におけるデータとは、特徴量 (説明変数) x とラベル y のペアで構成されるサンプルの集合を指し、利用できるデータ数の増加に従って機械学習モデルの予測性能も向上する。一方で、常に高品質なデータを多量に収集・活用できる場面は限定的である。本論文では一貫して、データが限られる環境においても機械学習モデルの学習やテスト時の分類に有効なサンプルの推定や重みづけを提案している。第 2 章および第 3 章ではラベル情報が限られる弱教師あり学習での疾患検出の精度向上に有効なサンプルの推定を提案し、第 4 章では特徴量が限られる共変量シフト環境でのひび割れ抽出の精度向上に有効なサンプルの推定を提案した。各章での貢献を以下にまとめる。

第 2 章では、超音波画像に自動設定される ROI に対して学習済みモデルを用いて得られる事後確率と周辺 ROI との位置情報に基づく信頼性評価法を提案し、肝硬変検出システムとしてその有効性を検証した。従来手法では超音波画像を用いた肝硬変検出において、自動設定された各 ROI の信頼性を特徴空間でのみ評価しており、各 ROI の位置関係を考慮していないことが課題であった。そこでまず仮説として、超音波画像の肝臓領域では、超音波の乱反射によるノイズの影響を受けた領域や影・血管など診断に不要な領域が存在し、信頼できる領域 (ROI) の周辺には似た特徴を持つ領域が存在すると考えた。またこの仮説をもとに、位置情報に基づく信頼性評価法 (Reliability Evaluation Method, REM) を提案した。実験では、識別率を含む定量的な評価指標によって信頼性評価法による信頼できる ROI の選択が分類器の学習とテスト時の分類性能向上に有効であることを確認した。この結果により、仮説の正当性を確認した。結果

では特に、クラス別の周辺領域の広さ（分類結果が肝硬変クラスでは狭く、健康者クラスでは広く）を設定することが分類性能向上に有効であった。各 ROI の分類結果を超音波画像上に可視化した結果、信頼できる ROI の分類結果は各クラスで一定の分離を画像空間上で持ち、すべての ROI を使用した分類器の学習とテスト時の分類を行う場合と比較して、より信頼性の高い出力であることを確認した。今後は、REM における周辺領域パラメーター設定において、肝臓領域の形状など追加の情報を考慮することで被験者単位もしくは画像単位の最適なパラメーター選択と分類性能向上が期待できる。

第 3 章では、従来の人手による画像特徴と分類器の設計を CNN モデルで代替することで、人の知識と経験に依存しない肝硬変検出を検討した。第 2 章での課題として Gabor 特徴と二次識別器の使用による、分類性能への限界があった。また、CNN モデルを用いた学習では手動設定された ROI を使用しており、ROI の品質が分類性能に影響を与える属人性の課題があった。そこでまず仮説として、第 2 章で有効性が確認できた自動設定 ROI を用いた REM が CNN モデルの学習と推論でも有効であると仮定した。CNN モデルの学習では、従来手法で提案されていた MIL フレームワークを使用し、自動設定 ROI を用いた学習でも CNN モデルの学習に有効な ROI の選択を検討した。テスト時は REM の位置情報に基づく信頼性評価によって、テストデータの分類に有効な ROI の選択を検討した。実験では、MIL による少量のデータ棄却や REM によるテストデータの信頼性評価が肝硬変検出精度の向上に有効であることが分かった。実験では REM の周辺領域や MIL での Drop rate などパラメーター総当たりでのテストデータに対する結果を考察している。そのため、実用化に向けては検証データを用いたパラメーター最適化を含めた制度検証が必要である。また、第 2 章で用いた Gabor 特徴と QDA による分類性能よりも CNN モデルによる分類性能が劣っていることから、CNN モデルを超音波画像データセットに用いる際の超音波画像ノイズへの過適合を課題として確認した。

第 4 章では、学習データと現場での運用時のデータで入力分布が変化する共変量シフトに対して重要度重みづけによる重要サンプルの自動推定と、セグメンテーションモデルの学習を提案した。ひび割れ調査における機械学習を用いたひび割れ抽出では、モデル学習と現場での運用時でデータ分布が同一であることを想定している。一方で運用時には画像撮影時に使用できるカメラの性能に制限があり、使用できる画像が低解像度に限定される可能性があり、高解像度画像で学習されたモデルではひび割れ抽出精度が劣化する課題がある。そこで仮説として、機械学習モデルを用いたひび割れ抽出では、共変量シフト適応によって学習データへの重要度重みづけを行い、運用時に想定されるデータに対するひび割れ抽出性能の向上が期待できると考えた。提案では、共変量シフト適応の 1 ステップアプローチにより、重要度推定を行いつつ U-Net の学習

を行い、低解像度に適応した学習を行った。実験では重要なサンプルが訓練データで 1% しかないとき、重要サンプルへの重みづけによってテスト時のひび割れ抽出性能向上を実現した。これにより、提案手法が重要サンプル数の少ない訓練データを用いる際に有効であると明らかにした。また、Class Activation Mapping を用いた注目領域の可視化により、重要度推定において解像度間の違いを判別するために重要となる領域を明らかにした。実験は単一のデータセットを用いて行われているため、異なる地域やインフラ構造物を対象とした複数のデータセットを用いた検証が必要である。また、画素単位で重要度推定を行うことで、画像内の影やコンクリート表面の欠損など画像内の領域を限定した重要度の推定による共変量シフト適応の活用も研究の方向性の一つである。これにより、その際は、解像度の変化だけでなく影による画素値の分布変化や、撮影角度の変化によるひび割れの見え方の変化など、様々な共変量シフトへの適応が可能になると期待される。

5.2 今後の展望

第 2 章および第 3 章の研究の展望として、ChatGPT に代表される Attention 機構を用いた医療画像分類が期待できる [81]。これまでの MIL では主にインスタンス単位のインスタンススコアをバッグ単位に統合するために、特徴次元方向の平均値プーリングや最大値プーリングを行っていたが、選択されたプーリング方法がテスト時の分類性能に影響を与える。そこで Ilse らによって、MIL におけるインスタンスごとの重要度を Attention 機構で自動最適化する Attention-Based MIL[81] が提案され、MIL を用いた医療画像分類のデファクトスタンダードとなっている。以下に、Ilse らの提案である Attention-Pooling を示す。

$$g(\mathbf{X}) = p(\sum_i A_i f(\mathbf{x}_i)) \quad (5.1)$$

ここで、 $\mathbf{x}_i$ はインスタンス（パッチ画像）、 $\mathbf{X} = \{\mathbf{x}_1; \mathbf{x}_2; \dots; \mathbf{x}_N\}$ は N 個のインスタンスで構成されるバッグ、 $f(\mathbf{x}_i)$ は特徴抽出器、 A_i は $\sum_i A_i = 1$ となるアテンションスコア、 $p(\cdot)$ は最終的な分類器を示す。MIL を用いた疾患の検出では被験者単位の分類性能だけでなく、モデルが注目した領域をアテンション機構によって可視化することで解釈性の高い診断支援が可能となる。特に Jang らは医療画像のようなインスタンスが独立な場合と依存関係を持つ場合の双方について、バッグ単位の学習だけでなくインスタンス単位の学習が MIL で可能であることを理論的に解析している [82]。Ilse らによる Attention-Based MIL の提案 (5.1) では、各インスタンスから得られた特徴量をアテンションスコアに応じて統合した後、バッグ単位での分類を行っている。そのため、各インスタンス単

体で正しい分類を可能にする特徴量を出力層まで保持することが保証されていないことから、Jang らによってインスタンス単位の学習可能性は理論的に保証されないと報告されている。一方で、以下に示す **Conjunctive Pooling** では、各インスタンス単位での特徴量を用いた分類結果をアテンションスコアに基づいて統合してバグ単位の分類とする。

$$g(\mathbf{X}) = \sum_i A_i p(f(\mathbf{x}_i)) \quad (5.2)$$

各インスタンスに対する分類結果の重みづけで学習を行うためインスタンス単位での学習が Jang らの定義した条件下では理論的に学習可能であることが示され、テスト時のインスタンス単位での分類が可能となる。これにより、バグ単位での予測の際に重要な領域をインスタンス単位で正確に学習・分類することが可能となり、医師の診断支援において解釈性向上につながる可能性が高い。また、画像単位でのラベル付けのほかに、医師による疾患領域のあいまいな教示を用いた CNN の学習への活用が期待できる。本論文で提案した位置情報に基づく信頼性評価法をあいまいな教示の増強に使用することで、各インスタンス間の関係性を考慮した学習が可能になり、より信頼性の高いアテンション機構の学習を可能にする分類性能の向上が期待できる。

第 4 章の研究の展望として、テスト環境の連続的な変化への対応があげられる。本論文で提案した共変量シフト適応では、適応先としてテスト環境で入力分布がシフトする 1 つの分布を想定していた。一方で、そのような分布シフトを想定するたびに共変量シフトによる適応でモデルの学習をやり直していると、コストが高くなる課題がある。ひび割れの検出・抽出においては、同じ検査現場でも時間経過とともに経年劣化や季節ごとの によって入力画像の分布が変化することが想定される。近年、Zhang ら [83] は、再学習なしで学習済みモデルの中間表現を移動平均統計量に基づいて更新し、非定常なテスト分布に適応させる手法を提案した。本手法では、分類問題におけるテスト環境の非定常性を取り扱っているが、画素単位での適応に拡張することで、ひび割れ抽出にも適応可能であると考えられる。これにより、さらに広い範囲での非定常環境において、ひび割れ検査の自動化・省人化が期待できる。

本論文では一貫して、データが限られる環境においても機械学習モデルの学習やテスト時の分類に有効なサンプルの推定を行い、精度良い予測が可能な機械学習の実現を目指してきた。機械学習による自動化や省人化は人口減少の時代にある日本の重要な研究テーマの一つである。単純に汎化性能の高いモデルや学習法を使用するだけでなく、応用の観点ではモデル出力の解釈性や信頼性の担保が非常に重要である。本論文ではモデルの出力や中間層の特徴量を可視化することで予測根拠のブラックボックス化を抑制した。機械学習を活用するうえで十分なデータ数が確保されることが望ましいが、限られたデータしか活

用できない環境においても，予測精度だけでなく根拠を考察している本論文の手法を活用・改良することで，応用場面での自動化・省人化に広く貢献できると考える．

付録

第 2 章，第 3 章 付録 超音波画像データセット

超音波画像データセット

第 2 章，第 3 章で使用するデータセットは山口大学附属病院で取得された，肝硬変症患者 12 名と健常者 8 名の超音波画像を用いる．各被験者から 5 枚の肝臓超音波画像を用いる．これらの画像は同一の医師によって撮影された．これらの画像は 640×480 ピクセル，256 階調の解像度を有する．撮影時の超音波周波数は腹部・深部観察向けの低～中周波数，4MHz とし，分解能よりも超音波減衰の低減を考慮している．画像化には表示深度を 160mm (= 16.0cm) とする．図1に示すように，超音波画像の操作野（台形のグレースケール部分）は，画像の X 軸中心での深さ（高さ）方向がおおよそ 351 画素，Y 軸座標 200 での横方向がおおよそ 314 画素（参照する高さで変化）である．

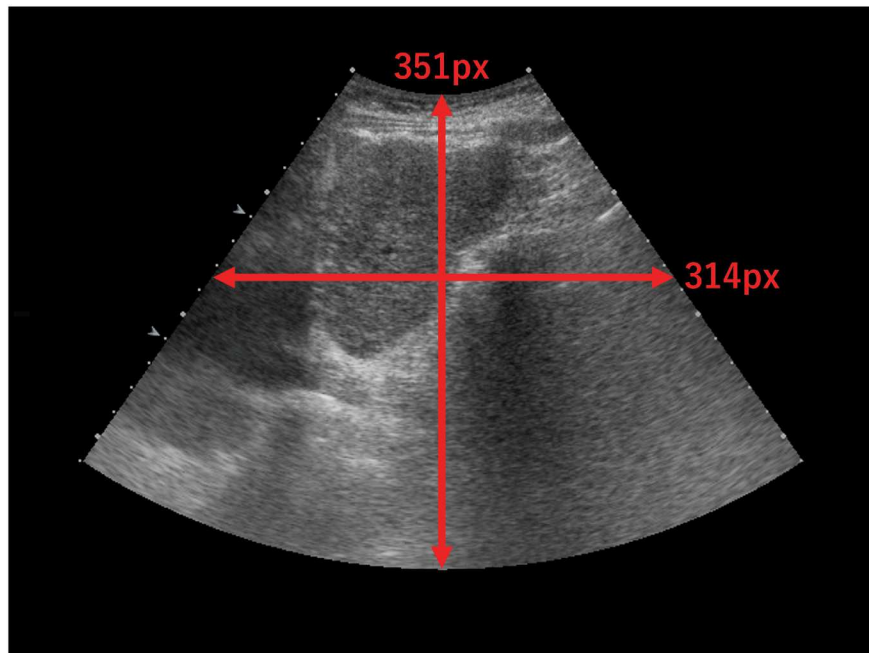


図 1: 超音波画像の操作野（観察領域）の大きさ．

画素あたりの実寸スケール

表示される超音波画像の 1 画素あたりの実寸スケールは、深さ方向が以下の値と推定される。

$$\text{深さ方向の画素スケール} = \frac{\text{表示深度}}{\text{深さ方向の表示画素}} \quad (1)$$

$$= \frac{160 \text{ (mm)}}{351 \text{ (px)}} \quad (2)$$

$$\approx 0.4558 \text{ (mm/px)} \quad (3)$$

横方向は視野角に依存して変化するため、一意に推定できない。一般的なコンベックスの視野角 θ を 60° もしくは 70° と仮定すると、横方向の画素スケールは以下の値と推定される。

$$\text{深さ } D \text{ における半幅} = D \tan\left(\frac{\theta}{2}\right) \quad (4)$$

$$\text{全横方向の画素スケール} = \frac{\text{深さ } D \text{ における全幅}}{\text{横方向の表示画素}} \quad (5)$$

$$= 2 \cdot D \tan\left(\frac{\theta}{2}\right) \cdot \frac{1}{\text{横方向の表示画素}} \quad (6)$$

$$= 2 \cdot 160 \tan 30^\circ \cdot \frac{1}{314} \approx 0.5884 \text{ (mm/px)} \quad (\theta = 60^\circ) \quad (7)$$

$$= 2 \cdot 160 \tan 35^\circ \cdot \frac{1}{314} \approx 0.7136 \text{ (mm/px)} \quad (\theta = 70^\circ) \quad (8)$$

ここで、画素スケールは分解能の実効値とは異なることに注意する必要がある。

実寸での分解能について

周波数 f を 4MHz としたとき超音波の波長 λ は、以下の値と推定される。

$$\lambda = \frac{c}{f} \quad (9)$$

$$= \frac{1540 \text{ (m/s)}}{4 \times 10^6 \text{ (m)}} \quad (10)$$

$$= 0.385 \text{ (mm)} \quad (11)$$

ここで、 c は人体軟部組織中の音速を指し、一般的に $1540(\text{m/s})$ とされる [84]. 深さ方向の分解能は、深さの異なる 2 つの反射体からの反射エコーがプローブ受信時に時間的に重ならないことが条件となり、パルス長の半分が分解能とされる. つまり、1 パルスに含まれる正弦波の数を $n = 3$ と仮定すると、

$$\Delta z = \frac{\text{SPL}}{2} \quad (12)$$

$$= \frac{n \cdot \lambda}{2} \quad (13)$$

$$= \frac{3 \cdot 0.385(\text{mm})}{2} \quad (14)$$

$$= 0.5775(\text{mm}) \quad (15)$$

で与えられる. $n = 2$ では、 $\Delta z \approx 0.385(\text{mm})$, $n = 4$ では、 $\Delta z \approx 0.770(\text{mm})$ であり、正弦波の数に依存するがおよそ $\Delta z \approx 0.39 \sim 0.77(\text{mm})$ であると推定される.

超音波画像の深さ方向の分解能が $0.5775(\text{mm})$ と仮定すると、画像上では

$$\frac{0.5775(\text{mm})}{0.4558(\text{mm/px})} \approx 1.2670(\text{px}) \quad (16)$$

のスケールとなる. 線維化によって生じる結節性を $3(\text{mm})$ とすると、

$$\frac{3(\text{mm})}{0.4558(\text{mm/px})} \approx 6.5818(\text{px}) \quad (17)$$

の大きさで表示される. そのため、 $4(\text{MHz})$ の周波数で $160(\text{mm})$ の表示深度とした際、目視診断での線維化の観察に十分な分解能で超音波画像が撮影されたと考えられる. 横方向の分解能の計算には、超音波ビーム幅 (超音波エネルギーが集中している横方向の広がり) が必要なため推定が困難である.

第 2 章で使用する Gabor 特徴について

先行研究については第 2 章で述べるが、ここでは画像所見に対して Gabor 特徴がとらえる特徴スケールを計算により明確にする. Gabor 特徴のガウス包絡は $\sigma \approx 7.07(\text{px})$ のため、深さ方向と横方向 (視野角に依存) はそれぞれ

$$\text{深さ方向} : 7.07(\text{px}) \times 0.4558(\text{mm/px}) \approx 3.2225(\text{mm}) \quad (18)$$

$$\text{横方向} : 7.07(\text{px}) \times 0.5884(\text{mm/px}) \approx 4.1600(\text{mm}) \quad (\theta = 60^\circ) \quad (19)$$

$$7.07(\text{px}) \times 0.7136(\text{mm/px}) \approx 5.0452(\text{mm}) \quad (\theta = 70^\circ) \quad (20)$$

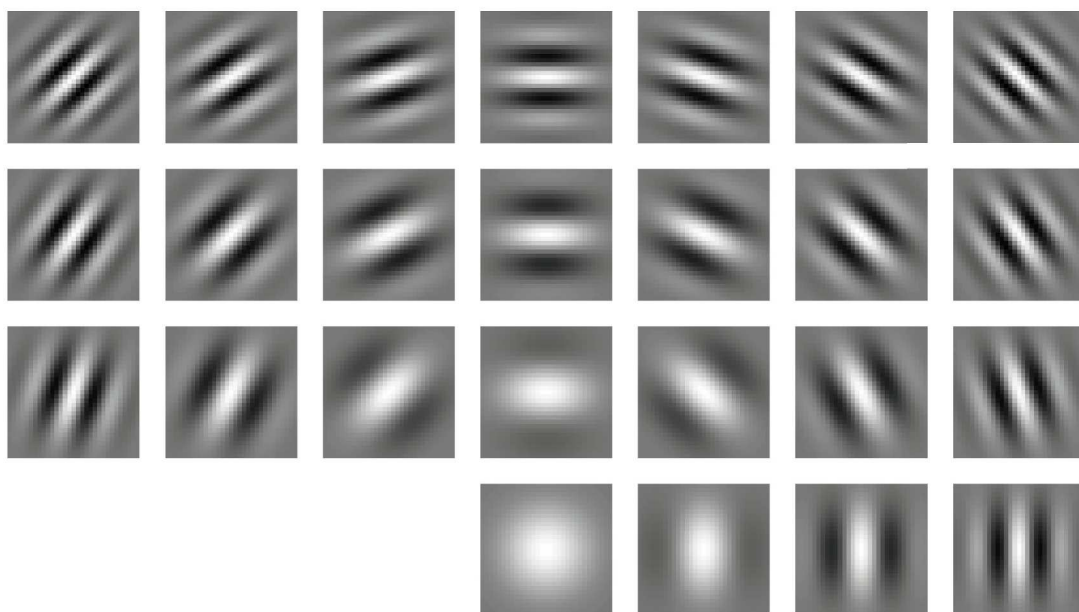


図 2: Gabor フィルタ 25 次元の実部.

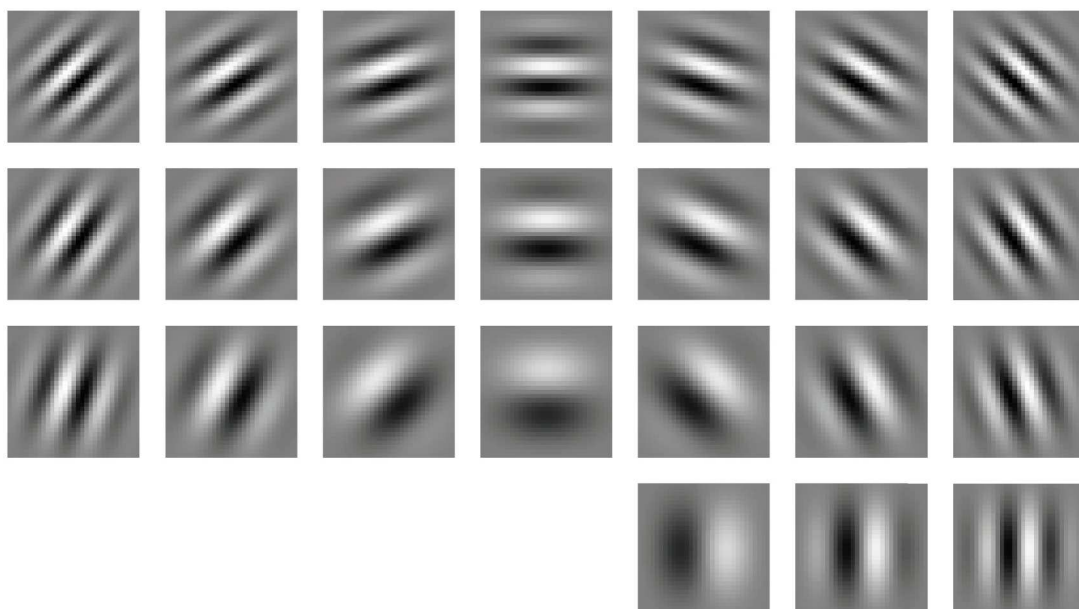


図 3: Gabor フィルタ 25 次元の虚部.

図2,3に 25 次元 Gabor フィルタの一覧を示す．Gabor 特徴の正弦波，余弦波の周波数は各方向を決定するパラメーターとして $k_y \in \{0; 1; 2; 3\}$; $k_x \in \{-3; -2; -1; 0; 1; 2; 3\}$ ($k_y = 0$ では， $k_x \in \{0; 1; 2; 3\}$) で， $f = (k_x/31; k_y/31)$ (フィルタサイズで正規化される) として決定される．2次元の縞 (等位相線) に直行する方向の代表波長は，

$$\lambda = \frac{1}{|f|} \quad (21)$$

$$= \frac{31}{\sqrt{k_x^2 + k_y^2}} \quad (22)$$

最大波長： $\sqrt{k_x^2 + k_y^2} = 1$ のとき

$$\lambda = \frac{31}{1} \quad (23)$$

$$= 31(\text{px}) \quad (24)$$

最小波長： $\sqrt{k_x^2 + k_y^2} = \sqrt{18} \approx 4.2426$ ($k_y = 3; k_x = 3$) のとき

$$\lambda = \frac{31}{\sqrt{18}} \quad (25)$$

$$\approx 7.3068(\text{px}) \quad (26)$$

となり，約 7.3 ～ 31(px) の範囲で明暗の周期構造を主に抽出する構造となる．7.3(px) の波長は，

$$\text{深さ方向：} 7.3(\text{px}) \times 0.4558(\text{mm/px}) \approx 3.3273(\text{mm}) \quad (27)$$

$$\text{横方向：} 7.3(\text{px}) \times 0.5884(\text{mm/px}) \approx 4.2953(\text{mm}) \quad (\theta = 60^\circ) \quad (28)$$

$$\text{横方向：} 7.3(\text{px}) \times 0.7136(\text{mm/px}) \approx 5.2032(\text{mm}) \quad (\theta = 70^\circ) \quad (29)$$

31(px) の波長は，

$$\text{深さ方向：} 31(\text{px}) \times 0.4558(\text{mm/px}) \approx 14.1298(\text{mm}) \quad (30)$$

$$\text{横方向：} 31(\text{px}) \times 0.5884(\text{mm/px}) \approx 18.2404(\text{mm}) \quad (\theta = 60^\circ) \quad (31)$$

$$\text{横方向：} 31(\text{px}) \times 0.7136(\text{mm/px}) \approx 22.1216(\text{mm}) \quad (\theta = 70^\circ) \quad (32)$$

のスケール (横方向は視野角に依存) となり，この Gabor 特徴が強く反応するスケールは，おおむね 3 ～ 18(mm) 程度と見積られる．そのため，結節性など肝硬変の所見を特徴として超音波画像上で捉えるには十分な設定で実験が行われている．

実験を行うにあたって，超音波画像から肝臓領域を手動で抽出した．抽出した肝臓領域は合計 100 枚であり，画像空間上で肝臓領域の縦幅は 159:1188 (平

均), 53:1913 (標準偏差), 横幅は 223:0792 (平均), 71:8754 (標準偏差) であった.

第 4 章 付録 共変量シフトの例

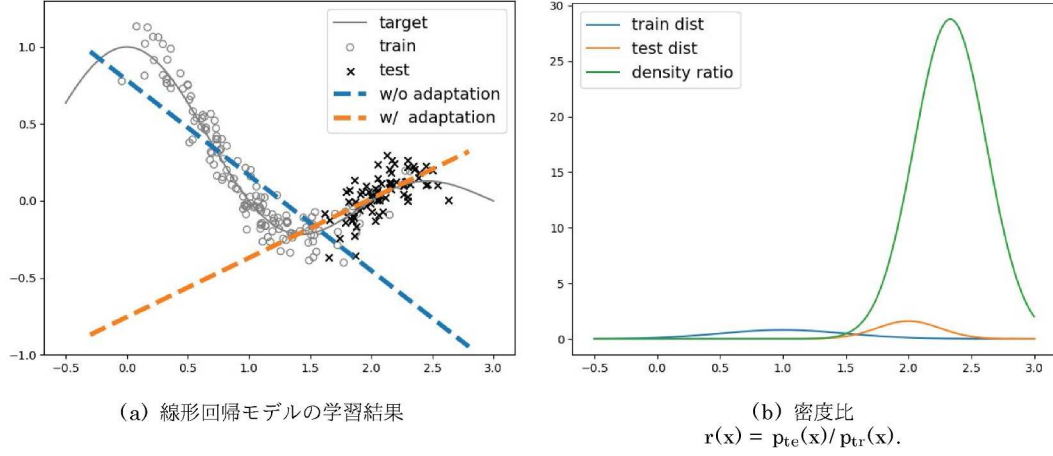


図 4: (a) 共変量シフト適応を用いた場合 (オレンジ) と用いない場合 (青) の、線形回帰モデルの学習結果を示す. (b) 学習データとテストデータの入力分布および密度比を示す. 学習データの各サンプルに対する重要度は、密度比 $r(x) = p_{te}(x)/p_{tr}(x)$ で与えられる. (a) におけるオレンジの線形回帰は密度比を用いた重要度重みづけによってロス関数への重みづけを行った結果である.

図4は 1 次元特徴を用いた回帰問題における共変量シフトのための重要度重みづけの有効性を示している. 実験の詳細な条件は以下の通りである. 学習データとテストデータの入力分布は $p_{tr}(x) = N(x; 1; (0.5)^2)$; $p_{te}(x) = N(x; 2; (0.25)^2)$ とする. ここで, $N(x; \mu, \sigma^2)$ は平均 μ , 分散 σ^2 の正規分布を示す. 予測目標の関数を $f(x) = \sin x/x$ とする. 学習データとテストデータのサンプルは $y = f(x) + c$, $c \sim N(0; (0.1)^2)$ からサンプリングする. 学習データとテストデータのサンプル数は $n_{tr} = 150$ と $n_{te} = 75$ とする. 線形回帰モデルのパラメータを SGD により学習率 0.001 で 3,000 回更新する. 密度比 $r(x) = p_{te}(x)/p_{tr}(x)$ は学習の前に、あらかじめ既知であり、密度比に基づく線形回帰モデルの学習に使用することとする. 密度比に基づく共変量シフトを使用した場合と学習データを用いた単純な学習を行った場合での学習結果は、図4.1で青とオレンジで示されている.

謝辞

本研究を進めるにあたり，学部 4 年次の研究室配属日から博士後期課程修了に至るまで，日頃よりご指導いただきました山口大学大学院創成科学研究科 藤田悠介准教授に心から感謝するとともに，厚く御礼申し上げます．特に研究室配属当初は実験方法や論文作成方法など，至らぬ点が多々ありましたが，丁寧かつ継続的に添削を行っていただく過程で研究者としての自立的な思考を育むことができました．また学術論文投稿にあたっては，複数回のリジェクトや査読コメント回答について根気よくご指導いただき，研究者としての技術力だけでなく忍耐力を培う貴重な機会を頂きました．これらの期間を通じて研究活動に専念し，日ごろの研究室ゼミや学会発表，論文投稿，また研究室での実験を行う環境を多忙な中でも整えていただき，博士課程を修了することができることに對して深謝申し上げます．

本論文を審査していただきました山口大学大学院創成科学研究科 多田村克己教授，中村秀明教授，田村慶信教授，間普真吾教授，藤田悠介准教授には論文の細部までご精査いただき，適切な助言を受け賜りました．心よりお礼申し上げます．

山口大学医学部附属病院の皆様には，第 2 章および第 3 章で使用した超音波画像データセットを提供していただき，心より感謝申し上げます．第 2 章および第 3 章の研究成果は，研究室における長年の実績と知見の積み重ねの上に成り立つものであり，その基盤となる本データセットが果たした貢献は非常に大きいものです．今後もこのような多くの方々のご貢献の上に学びや活動が成り立っていることを自覚し，感謝の気持ちを忘れずに歩んでまいります．

また，学部 4 年生次の研究室配属以来，日ごろお世話になりました研究室の先輩方，同期配属学生，後輩の皆さんには研究活動をはじめとする様々な活動を通じて実りある時間を共に過ごすことができましたことを大変感謝いたします．研究活動においては忖度なくお互いの気になる点を議論し，分からない点がありながらもより良い研究成果を求めて時間を費やした日々は，大学という環境ならではの貴重な経験でした．年代を超えて研究活動に限定せず交流を深めることで，研究面だけでなく人間的にも多くの学びを得ることができました．

山口大学大学院創成科学研究科 兵動正幸特命教授には，研究活動だけでなく

進路についても真摯にご指導いただき、心より感謝の意を表します。長年の研究活動に基づく豊富なご経験をもとに、アカデミアと産業界の双方でご活躍を拝見し、研究者および指導者として分野を超えて貢献される姿勢を学ばせていただきました。

最後に、これまでの生活を支えてくれた両親に感謝します。

本論文に関連する発表論文

- K. Nakata, Y. Fujita, Y. Mitani, Y. Hamamoto, M. Segawa, S. Terai, and I. Sakaida:
Automatic ROI Selection with Reliability Evaluation Method for Cirrhosis Detection Using Ultrasound Images,
IEEEJ Transactions on Electrical and Electronic Engineering, 19, 7, 1234-1244, 2024.
- K. Nakata, Y. Fujita, Y. Mitani, Y. Hamamoto, M. Segawa, and S. Terai:
ROI Selection to Improve Prediction Performance of CNN for Cirrhosis Detection Using Ultrasound Images,
in Proc. 2024 IEEE 13th Global Conf. on Consumer Electronics (GCCE), pp. 636-640, Oct. 2024.
- K. Nakata, and Y. Fujita:
Crack Segmentation with Covariate Shift Adaptation for Multi-Resolution Images,
in Proc. 10th Int. Conf. on Intelligent Informatics and BioMedical Sciences (ICIIBMS 2025), pp. 458-463, Online, Dec. 2025.

参考文献

- [1] Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar. Foundations of machine learning. MIT press, 2018.
- [2] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu and Dario Amodei. Scaling Laws for Neural Language Models. 2020. arXiv: 2001. 08361 [cs. LG]. url: <https://arxiv.org/abs/2001.08361>.
- [3] Oded Maron and Tomás Lozano-Pérez. “A framework for multiple-instance learning”. in *Advances in neural information processing systems*: 10 (1997).
- [4] Marthinus C Du Plessis, Gang Niu and Masashi Sugiyama. “Analysis of learning from positive and unlabeled data”. in *Advances in neural information processing systems*: 27 (2014).
- [5] Gang Niu, Marthinus Christoffel Du Plessis, Tomoya Sakai, Yao Ma and Masashi Sugiyama. “Theoretical comparisons of positive-unlabeled learning against positive-negative learning”. in *Advances in neural information processing systems*: 29 (2016).
- [6] Ryuichi Kiryo, Gang Niu, Marthinus C Du Plessis and Masashi Sugiyama. “Positive-unlabeled learning with non-negative risk estimator”. in *Advances in neural information processing systems*: 30 (2017).
- [7] Fabio Garcea, Alessio Serra, Fabrizio Lamberti and Lia Morra. “Data augmentation for medical imaging: A systematic literature review”. in *Computers in Biology and Medicine*: 152 (2023), page 106391. issn: 0010-4825. doi: <https://doi.org/10.1016/j.compbiomed.2022.106391>. url: <https://www.sciencedirect.com/science/article/pii/S001048252201099X>
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations”. in *International conference on machine learning*: PmLR. 2020, pages 1597–1607.

- [9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo and others. “Segment anything”. in *Proceedings of the IEEE/ CVF international conference on computer vision: 2023*, pages 4015–4026.
- [10] Yichi Zhang, Zhenrong Shen and Rushi Jiao. “Segment anything model for medical image segmentation: Current applications and future directions”. in *Computers in Biology and Medicine: 171* (2024), page 108238. issn: 0010-4825. doi: <https://doi.org/10.1016/j.compbiomed.2024.108238>. url: <https://www.sciencedirect.com/science/article/pii/S0010482524003226>.
- [11] Detlef Schuppan and Nezam H Afdhal. “Liver cirrhosis”. in *The Lancet: 371.9615* (2008), pages 838–851.
- [12] N Goyal, N Jain, V Rachapalli, DL Cochlin and M Robinson. “Non-invasive evaluation of liver cirrhosis using ultrasound”. in *Clinical radiology: 64.11* (2009), pages 1056–1066.
- [13] Sebastian Mueller and Laurent Sandrin. “Liver stiffness: a novel parameter for the diagnosis of liver disease”. in *Hepatic medicine: evidence and research: (2010)*, pages 49–67.
- [14] Arturo A Bravo, Sunil G Sheth and Sanjiv Chopra. “Liver biopsy”. in *New England Journal of Medicine: 344.7* (2001), pages 495–500.
- [15] Fikri M Abu-Zidan, Ashraf F Hefny and Peter Corr. “Clinical ultrasound physics”. in *Journal of emergencies, trauma, and shock: 4.4* (2011), pages 501–503.
- [16] Liv E Hetland, Thit M Kronborg, Mira Thing, Mikkel P Werge, Anders E Junker, Elias B Rashu, Malene B O’ Connell, Beth H Olsen, Anne-Sofie H Jensen, Nicolai J Wewer Albrechtsen and others. “Suboptimal diagnostic accuracy of ultrasound and CT for compensated cirrhosis: Evidence from prospective cohort studies”. in *Hepatology communications: 7.9* (2023), e0231.
- [17] Adrian Huber, Lukas Ebner, Johannes T Heverhagen and Andreas Christe. “State-of-the-art imaging of liver fibrosis and cirrhosis: A comprehensive review of current applications and future perspectives”. in *European journal of radiology open: 2* (2015), pages 90–100.
- [18] Chih-Ching Choong, Sudhakar K Venkatesh and Edwin PY Siew. “Accuracy of routine clinical ultrasound for staging of liver fibrosis”. in *Journal of clinical imaging science: 2* (2012), page 58.

- [19] A Di Lelio, C Cestari, A Lomazzi and L Beretta. "Cirrhosis: diagnosis with sonographic study of the liver surface." in *Radiology*: 172.2 (1989), pages 389–392.
- [20] Agostino Colli, Mirella Fraquelli, Marco Andreoletti, Barbara Marino, Enrico Zuccoli and Dario Conte. "Severe liver fibrosis or cirrhosis: accuracy of US for detection—analysis of 300 cases". in *Radiology*: 227.1 (2003), pages 89–94.
- [21] Roy A Filly, Shireesha G Reddy, Allen B Nalbandian, Ying Lu and Peter W Callen. "Sonographic evaluation of liver nodularity: Inspection of deep versus superficial surfaces of the liver". in *Journal of clinical ultrasound*: 30.7 (2002), pages 399–407.
- [22] Lei Shen, Ji-Qiang Li, Min-De Zeng, Lun-Gen Lu, Si-Tao Fan and Han Bao. "Correlation between ultrasonographic and pathologic diagnosis of liver fibrosis due to chronic virus hepatitis". in *World Journal of Gastroenterology: WJG*: 12.8 (2006), page 1292.
- [23] Hidenori Toyoda, Takashi Kumada, Naohisa Kamiyama, Katsuya Shiraki, Kojiro Takase, Tadashi Yamaguchi and Hiroyuki Hachiya. "B-mode ultrasound with algorithm based on statistical analysis of signals: evaluation of liver fibrosis in patients with chronic hepatitis C". in *American Journal of Roentgenology*: 193.4 (2009), pages 1037–1043.
- [24] Yu Igarashi, Hiroshi Ezuka, Tadashi Yamaguchi and Hiroyuki Hachiya. "Quantitative estimation method for liver fibrosis based on combination of Rayleigh distributions". in *Japanese Journal of Applied Physics*: 49.7S (2010), 07HF06.
- [25] JF Gerstenmaier and RN Gibson. "Ultrasound in chronic liver disease". in *Insights into imaging*: 5.4 (2014), pages 441–455.
- [26] TZ Ong and HJ Tan. "Ultrasonography is not reliable in diagnosing liver cirrhosis in clinical practice". in *Singapore medical journal*: 44.6 (2003), pages 293–295.
- [27] Paul Kriss, Haruka Miki-Imoto, Hiroshi Nishimaki and Takashi Riku. Evolution of Quality Infrastructure Investment in Japan. Report. Prepared by the Tokyo Development Learning Center (TDLC) with the support of the Government of Japan. Washington, DC: World Bank, 2021. doi: 10.1596/35457. url: <https://hdl.handle.net/10986/35457>.

- [28] Shinji Yoshioka and Hirofumi Kawasaki. Japan's High-Growth Postwar Period: The Role of Economic Plans. ESRI Research Note 27. The views expressed are those of the authors and do not necessarily reflect those of the ESRI or the Cabinet Office. Tokyo, Japan: Economic and Social Research Institute (ESRI), Cabinet Office, august 2016. url: https://www.esri.cao.go.jp/jp/esri/archive/e_rnote/e_rnote030/e_rnote027.pdf.
- [29] 国土交通省. インフラメンテナンスに関する調査・研究. https://www.nilit.go.jp/sogoseisaku/maintenance/02research/02_01.html. 最終閲覧日: 2026-01-04. 2024.
- [30] Qianwei Xu, Jinli Xie, Feng Zhou and Zhuohua Tang. “Numerical Simulation and Analysis of the Causes and Distribution of Secondary Lining Cracks in Overlapping Railway Tunnels”. in *Applied Sciences*: 13.11 (2023). issn: 2076-3417. doi: 10.3390/app13116436. url: <https://www.mdpi.com/2076-3417/13/11/6436>.
- [31] 国土技術政策総合研究所. PRC 道路橋の性能照査に関する研究. 第 8 章 ひび割れ幅が鋼材腐食に与える影響. 国総研資料第 620 号. 技術資料, PRC 道路橋性能照査に関する研究, 8 章: ひび割れ幅と鋼材腐食の影響. 独立行政法人土木研究所・国土技術政策総合研究所, january 2011. url: <https://www.nilim.go.jp/lab/bcg/siryou/tmn/tmn0620.htm>
- [32] Tom Fawcett. “An introduction to ROC analysis”. in *Pattern Recognition Letters*: 27.8 (2006). ROC Analysis in Pattern Recognition, pages 861–874. issn: 0167-8655. doi: <https://doi.org/10.1016/j.patrec.2005.10.010>. url: <https://www.sciencedirect.com/science/article/pii/S016786550500303X>
- [33] Qinghua Huang, Fan Zhang and Xuelong Li. “Machine learning in ultrasound computer-aided diagnostic systems: a survey”. in *BioMed research international*: 2018.1 (2018), page 5137904.
- [34] Ahmed Gaber, Hassan A Youness, Alaa Hamdy, Hammam M Abdelaal and Ammar M Hassan. “Automatic classification of fatty liver disease based on supervised learning and genetic algorithm”. in *Applied Sciences*: 12.1 (2022), page 521.
- [35] Benot Schmauch, Paul Herent, Paul Jehanno, Olivier Dehaene, Charlie Sailard, Christophe Aubé, Alain Luciani, Nathalie Lassau and Simon Jégou. “Diagnosis of focal liver lesions from ultrasound using deep learning”. in *Diagnostic and interventional imaging*: 100.4 (2019), pages 227–233.

- [36] Xiangfei Feng, Xin Chen, Changfeng Dong, Yingxia Liu, Zhong Liu, Ruixin Ding and Qinghua Huang. “Multi-scale information with attention integration for classification of liver fibrosis in B-mode US image”. in *Computer Methods and Programs in Biomedicine*: 215 (2022), page 106598.
- [37] Amit Das, Mary Connell and Shailesh Khetarpal. “Digital image analysis of ultrasound images using machine learning to diagnose pediatric nonalcoholic fatty liver disease”. in *Clinical imaging*: 77 (2021), pages 62–68.
- [38] Cristian Vicas, Monica Lupsor, Mihai Socaciu, Sergiu Nedevschi and Radu Badea. “Influence of Expert-Dependent Variability over the Performance of Noninvasive Fibrosis Assessment in Patients with Chronic Hepatitis C by Means of Texture Analysis”. in *Computational and Mathematical Methods in Medicine*: 2012.1 (2012), page 346713.
- [39] Yu-Hsiang Wu, Jhu-Yun Huang, Shyi-Chyi Cheng, Chen-Kuei Yang and Chih-Lang Lin. “Evolutionary feature construction for ultrasound image processing and its application to automatic liver disease diagnosis”. in *2011 International Conference on Complex, Intelligent, and Software Intensive Systems*: IEEE, 2011, pages 565–570.
- [40] U Rajendra Acharya, U Raghavendra, Hamido Fujita, Yuki Hagiwara, Joel EW Koh, Tan Jen Hong, Vidya K Sudarshan, Anushya Vijayananthan, Chai Hong Yeong, Anjan Gudigar and others. “Automated characterization of fatty liver disease and cirrhosis using curvelet transform and entropy features extracted from ultrasound images”. in *Computers in biology and medicine*: 79 (2016), pages 250–258.
- [41] Pengfei Zhang, Hong Huang, Qiuju Xiong, Xinlu He and Yong Liu. “Feature analysis and automatic classification of B-mode ultrasound images of fatty liver”. in *Biomedical Signal Processing and Control*: 79 (2023), page 104073.
- [42] Fayyaz ul Amir Afsar Minhas, Durre Sabih and Mutawarra Hussain. “Automated classification of liver disorders using ultrasound images”. in *Journal of medical systems*: 36.5 (2012), pages 3163–3172.
- [43] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. 2 edition. San Diego: Academic Press, 1990. isbn: 978-0-12-269851-4.
- [44] Y. Fujita, Y. Hamamoto, M. Segawa, S. Terai and I. Sakaida. “An improved method for cirrhosis detection using liver’s ultrasound images”. in *2010 20th International Conference on Pattern Recognition (ICPR)*: 2010, pages 2294–2297. doi: 10.1109/ICPR.2010.561.

- [45] Y. Fujita, T. Goto, Y. Mitani, Y. Hamamoto, M. Segawa, S. Terai and I. Sakaida. “A liver cirrhosis detection method using probabilistic ROI combination”. in *11th International Conference on Quality Control by Artificial Vision (QCAV 2013)*: 2013, pages 81–85.
- [46] Josef Kittler, Mohamad Hatef, Robert P. W. Duin and Jiri Matas. “On Combining Classifiers”. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*: 20.3 (1998), pages 226–239. doi: 10.1109/34.667881.
- [47] Paul Viola, John C. Platt and Cha Zhang. “Multiple Instance Boosting for Object Detection”. in *Advances in Neural Information Processing Systems 18 (NeurIPS 2005)*: 2006, pages 1417–1424.
- [48] Y. Fujita, Y. Mitani, Y. Hamamoto, M. Segawa, S. Terai and I. Sakaida. “Training ROI Selection Based on MILBoost for Liver Cirrhosis Classification Using Ultrasound Images”. in *Trends in Applied Knowledge-Based Systems and Data Science (IEA/AIE 2016)*: volume 9799. Lecture Notes in Computer Science. Springer, 2016, pages 451–459.
- [49] Y. Fujita, Y. Mitani, Y. Hamamoto, M. Segawa, S. Terai and I. Sakaida. “Training sample selection based on self-training for liver cirrhosis classification using ultrasound images”. in *Proc. SPIE 10338, 13th International Conference on Quality Control by Artificial Vision (QCAV 2017)*: 2017.
- [50] 藤田 悠介, 三谷 芳弘 and 浜本 義彦. “肝硬変診断のための超音波画像解析”. in *日本信頼性学会誌信頼性*: 45.6 (2023), pages 312–318. doi: 10.1134/8/reaishinrai.45.6_312.
- [51] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. in *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*: 2012, pages 1097–1105.
- [52] scikit-learn developers. *Classifier comparison* —scikit-learn 1.7.2 documentation. https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html. アクセス: 2025 年 11 月 4 日. (一部コード改変). 2025.
- [53] Donglai Wei, Bolei Zhou, Antonio Torralba and William T. Freeman. mNeuron: A Matlab Plugin to Visualize Neurons from Deep Models. <https://donglai.github.io/proj/mneuron/index.html>. Accessed: 2025-11-04. 2017.

- [54] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei. “ImageNet: A large-scale hierarchical image database”. in *2009 IEEE Conference on Computer Vision and Pattern Recognition: 2009*, pages 248–255. doi: 10.1109/CVPR.2009.5206848.
- [55] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. “Deep residual learning for image recognition”. in *Proceedings of the IEEE conference on computer vision and pattern recognition: 2016*, pages 770–778.
- [56] Alex Krizhevsky, Geoffrey Hinton and others. *Learning multiple layers of features from tiny images*. Toronto, ON, Canada, 2009.
- [57] Zeynettin Akkus, Jason Cai, Arunnit Boonrod, Atefeh Zeinoddini, Alexander D. Weston, Kenneth A. Philbrick and Bradley J. Erickson. “A Survey of Deep-Learning Applications in Ultrasound: Artificial Intelligence-Powered Ultrasound for Improving Clinical Workflow”. in *Journal of the American College of Radiology: 16.9 Part B (2019)*, pages 1318–1328. doi: 10.1016/j.jacr.2019.06.004. url: <https://doi.org/10.1016/j.jacr.2019.06.004>.
- [58] Kai Wang, Xiaohua Lu, Hong Zhou, Yaqiong Gao, Jiong Zheng, Ming Tong, Chuan Wu, Chao Liu, Lujun Huang, Tao Jiang, Fangfang Meng, Yifan Lu, Huiling Ai, Xiaoyan Xie, Liping Yin, Ping Liang, Jie Tian and Rongxin Zheng. “Deep learning radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis B: a prospective multicentre study”. in *Gut: 68.4 (april 2019)*. Epub 2018 May 5, pages 729–741. doi: 10.1136/gutjnl-2018-316204. url: <https://pubmed.ncbi.nlm.nih.gov/29730602/>.
- [59] Dan Meng, Libo Zhang, Guitao Cao, Wenming Cao, Guixu Zhang and Bing Hu. “Liver fibrosis classification based on transfer learning and FCNet for ultrasound images”. in *Ieee Access: 5 (2017)*, pages 5804–5810.
- [60] Micha Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, ukasz Michaowski, Rafa Paluszkiewicz, Bogna Ziarkiewicz-Wróblewska, Krzysztof Zieniewicz, Piotr Sobieraj and Andrzej Nowicki. “Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images”. in *International journal of computer assisted radiology and surgery: 13.12 (2018)*, pages 1895–1903.
- [61] Mainak Biswas, Venkatanaresbhabu Kuppili, Damodar Reddy Edla, Harman S Suri, Luca Saba, Rui Tato Marinho, J Miguel Sanches and Jasjit S Suri. “Symtosis: A liver ultrasound tissue characterization and risk stratification

- in optimized deep learning paradigm”. in *Computer methods and programs in biomedicine*: 155 (2018), pages 165–177.
- [62] Xiang Liu, Jia Lin Song, Shuo Hong Wang, Jing Wen Zhao and Yan Qiu Chen. “Learning to diagnose cirrhosis with liver capsule guided ultrasound image classification”. in *Sensors*: 17.1 (2017), page 149.
 - [63] Yusuke Fujita, Kaito Ishihara, Kazuma Nakata, Yoshihiko Hamamoto, Makoto Segawa, Isao Sakaida, Yoshihiro Mitani and Shuji Terai. “Weakly supervised multiple instance learning for liver cirrhosis classification using ultrasound images”. in *2022 7th International Conference on Intelligent Informatics and Biomedical Science (ICIIBMS)*: volume 7. IEEE. 2022, pages 225–229.
 - [64] Diederik P Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. in *arXiv preprint arXiv:1412.6980*: (2014).
 - [65] Kai Li, Bo Wang, Yingjie Tian and Zhiquan Qi. “Fast and Accurate Road Crack Detection Based on Adaptive Cost-Sensitive Loss Function”. in *IEEE Transactions on Cybernetics*: 53.2 (2023), pages 1051–1062. doi: 10.1109/TCYB.2021.3103885.
 - [66] Qin Zou, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang and Song Wang. “DeepCrack: Learning Hierarchical Convolutional Features for Crack Detection”. in *IEEE Transactions on Image Processing*: 28.3 (2019), pages 1498–1512. doi: 10.1109/TIP.2018.2878966.
 - [67] Vijay Badrinarayanan, Alex Kendall and Roberto Cipolla. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”. in *IEEE Transactions on Pattern Analysis and Machine Intelligence*: 39.12 (2017), pages 2481–2495. doi: 10.1109/TPAM.2016.2644615.
 - [68] Riccardo Chianese, Andy Nguyen, Vahidreza Gharehbaghi, Thiru Aravinthan and Mohammad Noori. “Influence of image noise on crack detection performance of deep convolutional neural networks”. in *arXiv preprint arXiv:2111.02079*: (2021).
 - [69] Huijun Liu, Chunhua Yang, Ao Li, Sheng Huang, Xin Feng, Zhimin Ruan and Yongxin Ge. “Deep Domain Adaptation for Pavement Crack Detection”. in *IEEE Transactions on Intelligent Transportation Systems*: 24.2 (2023), pages 1669–1681. doi: 10.1109/ITITS.2022.3225212.

- [70] Fei Deng, Shaohui Yang, Bin Wang, Xiujun Dong and Siyuan Tian. “UCrack-DA: A Multi-Scale Unsupervised Domain Adaptation Method for Surface Crack Segmentation”. in *Remote Sensing*: 17.12 (2025). issn: 2072-4292. url: <https://www.mdpi.com/2072-4292/17/12/2101>.
- [71] Jinhuan Shan, Wei Jiang and Xiao Feng. “Bridging cross-domain and cross-resolution gaps for UAV-based pavement crack segmentation”. in *Automation in Construction*: 174 (2025), page 106141. issn: 0926-5805. doi: <https://doi.org/10.1016/j.autcon.2025.106141>. url: <https://www.sciencedirect.com/science/article/pii/S0926580525001815>.
- [72] Hidetoshi Shimodaira. “Improving predictive inference under covariate shift by weighting the log-likelihood function”. in *Journal of Statistical Planning and Inference*: 90.2 (2000), pages 227–244. issn: 0378-3758. doi: [https://doi.org/10.1016/S0378-3758\(00\)00115-4](https://doi.org/10.1016/S0378-3758(00)00115-4). url: <https://www.sciencedirect.com/science/article/pii/S0378375800001154>.
- [73] Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya and Masashi Sugiyama. “Relative density-ratio estimation for robust distribution comparison”. in *Proceedings of the 25th International Conference on Neural Information Processing Systems: NIPS'11*. Granada, Spain: Curran Associates Inc., 2011, pages 594–602. isbn: 9781618395993.
- [74] Tianyi Zhang, Ikko Yamane, Nan Lu and Masashi Sugiyama. “A One-Step Approach to Covariate Shift Adaptation”. in *SN Comput. Sci.*: 2.4 (june 2021). doi: 10.1007/s42979-021-00678-6. url: <https://doi.org/10.1007/s42979-021-00678-6>.
- [75] Olaf Ronneberger, Philipp Fischer and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. in *International Conference on Medical image computing and computer-assisted intervention*: Springer. 2015, pages 234–241.
- [76] Jimmy Lei Ba, Jamie Ryan Kiros and Geoffrey E Hinton. “Layer normalization”. in *arXiv preprint arXiv:1607.06450*: (2016).
- [77] Yuxin Wu and Kaiming He. “Group normalization”. in *Proceedings of the European conference on computer vision (ECCV)*: 2018, pages 3–19.
- [78] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines”. in *Proceedings of the 27th international conference on machine learning (ICML-10)*: 2010, pages 807–814.

- [79] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva and Antonio Torralba. “Learning deep features for discriminative localization”. in *Proceedings of the IEEE conference on computer vision and pattern recognition*: 2016, pages 2921–2929.
- [80] Çalar Frat Özgenel. Concrete Crack Segmentation Dataset. Mendeley Data, version 1. CC BY 4.0 license. 2019. doi: 10.17632/jwsn7tfbrp.1. url: <https://data.mendeley.com/datasets/jwsn7tfbrp/1>.
- [81] Maximilian Ilse, Jakub Tomczak and Max Welling. “Attention-based deep multiple instance learning”. in *International conference on machine learning*: PMLR. 2018, pages 2127–2136.
- [82] Jaeseok Jang and HYUK-YOON KWON. “Are Multiple Instance Learning Algorithms Learnable for Instances?” in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*: 2024. url: <https://openreview.net/forum?id=cUvlgkQxP>.
- [83] Zhen-Yu Zhang, Zhiyu Xie, Huaxiu Yao and Masashi Sugiyama. “Test-time adaptation in non-stationary environments via adaptive representation alignment”. in *Advances in Neural Information Processing Systems*: 37 (2024), pages 94607–94632.
- [84] Alex Benjamin, Rebecca E. Zubajlo, Manish Dhyani, Anthony E. Samir, Kai E. Thomenius, Joseph R. Grajo and Brian W. Anthony. “A Novel Approach to the Quantification of the Longitudinal Speed of Sound and its Potential for Tissue Characterization (Part I)”. in *Ultrasound in Medicine & Biology*: 44.8 (2018), pages 1592–1604. issn: 0301-5629. doi: 10.1016/j.ultrasmedbio.2018.03.023.