Abdomen Lesion Classifier using an XAI Model

Ben Johnson
X00229603@mytudublin.ie
Technological University Dublin

Rohith Budugarla
X00229544@mytudublin.ie
Technological University Dublin

Ravindra Sai Cherukuru
X00229508@mytudublin.ie
Technological University Dublin

Sravan Kumar Yannam
X00229546@mytudublin.ie
Technological University Dublin

Abstract

Medical imaging is a vital instrument in contemporary medicine because it allows for the the detection and monitoring of internal irregularities of the body such as lesions. However, increased use of deep learning models on medical image analysis has raised major concerns regarding explainability, fairness, and ethical use of the models.

This project is aimed at overcoming the presented challenges by creating a Convolutional Neural Network (CNN) model to classify abdominal lesions based on a curated subset of the NIH Deep Lesion dataset, supplemented with samples from the Mendeley Abdominal CT scan dataset. Detailed preprocessing methods such as: greyscale conversion, contrast enhancement and data normalization were applied to the images. A custom CNN architecture was developed for the lesion's classification in abdominal CT images. To improve the interpretability of the model, Explainable AI (XAI) methods, like Saliency Maps using Gradient-weighted Class Activation Mapping (Grad-CAM), and Local Interpretable Model-Agnostic Explanations (LIME), were used. These tools provided ways to guide the behaviour of the models, by providing visual explanations of how the models' predictions were made.

This research also critically analyses probable biases to data and the ethical implication of using AI in healthcare. It is consistent with the legal frameworks like General Data Protection Regulation (GDPR) and Health Insurance Portability And Accountability Act (HIPAA); also, it suggests mitigations for risks associated with fairness, accountability, and transparency. With the combination of deep learning and human-centered explainability, the implementation of trustworthy and responsible AI systems in medicine diagnostics is one step forward presented within this project.

Keywords

Convolutional Neural Networks, Computed Tomography, Explainable AI, Lesion Classification, Ethical AI, Medical AI

ACM Reference Format:
Ben Johnson, Rohith Budugarla, Ravindra Sai Cherukuru, and Sravan Kumar Yannam . 2025. Abdomen Lesion Classifier using an XAI Model.

1 Introduction

AI is rapidly transforming the healthcare landscape, allowing new ways of diagnostics and decision-making, as well as treatment planning. Among the areas of the most significant evolution, it is necessary to single out the analysis of medical images. Methods like X-ray, M.R.I, Computed Tomography (CT) scanning are common techniques employed by clinicians in exploring abnormalities in human organs and tissues. However, the interpretation process of medical images is time-consuming, expert-dependent and prone to observer variability and human error. This has opened up doors for deep learning and mainly the CNNs to be more widely applied to medical functions such as lesion detection and organ classification among others. Although CNNs have shown outstanding performance in reaching high classification rates, they are often touted for lacking transparency in explaining how they reach a given conclusion. Such "black-box" nature poses substantial challenges in the domain of trust, fairness, and accountability, in the high-stakes domains like healthcare.

Through a CNN-based system for the detection of lesions in abdominal CT images and with the use of XAI techniques, this project meets these challenges. In particular, we focus on a small sample of the NIH Deep Lesion dataset, which includes approximately 32,000 CT slices with annotations for various parts of anatomy. From this data set, we extracted a focused set of images relevant to abdominal lesions, an area where diagnostic challenges come in because of the overlaying structures and diversity of the organs in the abdomen. We also combined image sets from the Mendeley Abdominal CT scan dataset in order to increase variability and generalizability. Our focus is not only on model performance, but also on transparency, which means that healthcare professionals and researchers are able to comprehend and corroborate the process of making predictions.

The significance of explainability in the models of AI, especially in medical areas cannot be overestimated. Although deep learning models are powerful tools, they usually do not offer an intuitive explanation of their output. This is particularly troublesome when it comes to domains such as radiology where reasoning of a diagnosis must be searchable, accountable, and comprehensible by everyone, including non-technical audiences (clinicians, patients, regulators). Responding to this problem, we added a couple of XAI techniques that interpret the behaviour of the model: Saliency Maps represent the pixel that is most important for prediction; Grad-CAM produces a heatmap to demonstrate what area was taken into account in making a final decision; and LIME generates local substitute models to present an explanation for a specific prediction These tools return visual and quantitative information to users about the model regarding the internal decision process, and thus allows for making judgments about reliability.

Data preprocessing is a key for enhancing the generalization capability of the model. Medical datasets are commonly affected by

HCAIM CA3, Dublin, Ireland

ACM ISBN

class imbalance, lack of suitable samples, and variation in image quality. In order to solve these problems, we applied a wide range of preprocessing techniques on images such as the Image Ops Equalization techniques for enhancing the contrast & standardization of input dimensions (resizing and greyscale conversion) to synthetically increase dataset diversity. These steps do not only prevent overfitting but also enable the model to learn more detailed features that can represent scenarios in the real world. During the development of the model, we carried on emphasising explainability, using XAI outputs throughout all the stages of evaluation. The present research offers an exhaustive framework for an interpretable deep learning in medical image analysis with a particular emphasis on abdominal lesion detection.

2 Literature Review

The adoption of AI in the healthcare sphere, and in the specific field of analysis of medical images in particular, has attracted colossal attention over the last decade. Deep learning approaches like CNNs have been proved by many studies to be better than the traditional ones when it comes to a task such as tumour detection, segmentation and classification. One of the primary works in this field formed of a review of more than 300 papers using deep learning for medical imaging that found the CNNs to be highly efficient in such tasks as lesion detection and organ localization [1]. Their work showed the growing trend of supervised learning and the need for annotated datasets, which is still a limitation in many of the current studies.

The NIH Deep Lesion dataset [2] is the remarkable step in this course. It gives a large-scale collection of the CT slices labelled for numerous types of lesions in different body parts. The dataset is applied for a number of challenging tasks of lesion classification and detection and is a standard for model testing. However, most of the models trained over this dataset either use entire label space or focus over multi-class detection. In contrast, our study is reduced to the abdominal lesions since it is an under-explored aspect of diseases such as liver cancer and kidney cysts, as well as the intestinal conditions.

In the last few years, explainability has been of paramount importance in research. One study proposed LIME that aimed at approximating local decision boundaries of complex models with simpler interpretable models [3]. Although LIME performs well in some domains, this technique has a limitation for high-resolution image data because of super pixel segmentation noise. Another study suggested Grad-CAM, which has become a popular method to visualize class regions in CNNs by means of gradient of target classes which flow into the final convolutional layer [4]. It is the Saliency Maps which were introduced previously, which makes use of gradient of the output with regards to input pixels to emphasize important regions that is particularly helpful for the medical experts to determine the focus of attention [5].

While these tools help to gain insights into the model behaviour, recent research [6] has demonstrated that XAI tools may produce misleading or unstable outputs in some cases. Thus, several XAI methods are usually used to confirm the trust possibility of the explanations, a measure that is followed in our research. Behind technical literature, regulatory and ethical aspects are no less significant in a healthcare ecosystem. The right to explanation, enforced by GDPR, requires users to ask questions about how the mechanism of making algorithmic decisions works and as a result, XAI is not just useful, but necessary. Similar work examines the intersection between AI decision-making and legal compliance, whereby explaining is one mode of due process in automated systems [7].

The upcoming EU AI Act brings a risk-based approach to AI regulation, whereby medical AI is classified as belonging to the 'high-risk' category. Studies have been conducted [8] which highlight the necessity of transparency, robustness, and accountability with regards to medical AI deployments. Through our work, we integrate these concerns because we focus on examining bias in datasets, performance on subgroups, and explainability as compliance.

Finally, Human-Centered AI has emerged as the discipline that is concerned with making AI systems transparent, trustworthy, and ethical. According to one body of research [9], there are principles for the design of AI which put human needs at the centre; such as transparency, control and collaboration. An such, our project does not only utilize CNNs for classification, but ensures that outputs can be easily explained and validated by human users, clinicians, researchers and policy-makers alike. To summarize:

- The majority of the lesion detection studies examine whole-body CT scans; few isolate abdominal lesions specifically.
- Explainability is neglected by many high-performing CNN models hence limiting clinical adoption.
- Very few works on the combination of several XAI methods (Saliency maps using Grad-CAM, LIME) to increase robustness and explainability.
- Not many studies place technical work in a legal context (GDPR, ethical design frameworks – ALTAI, HCAI).
- Lack of exploration around fairness and demographic bias in widely used medical datasets.

3 Data Exploration

In any deep learning pipeline, particularly in the field of medical image analysis, data quality and preparation take top priority of any model's performance. But looking beyond technical sufficiency, a human-centered AI perspective demands this deeper questioning of how sources of data are and what biases are there, and whether any preprocessing techniques enable both a performance and fairness. We started our work on this project with a study of the NIH DeepLesion dataset, filtered for abdominal lesions, and added a strong, explainability-sensitive preprocessing pipeline in order to guarantee the validity and trustworthiness of our classification task.

3.1 Dataset Overview & Rationale

NIH DeepLesion dataset is an extensive data set of medical images created by National Institutes of Health that consists of over 32,000 CT scan slices with the annotated lesions in various anatomical parts. At the beginning, our project scope proposed the use of several classes of this dataset for multi-class lesion detection model: lung, liver, mediastinum, abdomen classes. After a closer look during the data exploration stage of our work, we discovered that some classes and especially those corresponding to smaller or

less frequent lesion regions had not enough labels associated for sustainable training for deep learning models without running the risk of overfitting or high variance.

On the other hand, the abdominal lesion group offered more sample sizes with better visual consistency as well as better visibility of the lesions in the CT slices. This made it a more appropriate candidate for binary classification and explianability studies. Consequently, we made our scope of interest limited even to images pertaining only to the class of abdomen; thus, we formed two categories, CT scans with abdominal lesions and CT scans without abdominal lesions. This decision allowed us to:

- Train a model with enough data volume for meaningful generalization.
- Minimize label noise and organ-type variability.
- Align with a real-world medical use case: detecting abdominal abnormalities like cysts, tumours, or internal bleeding.

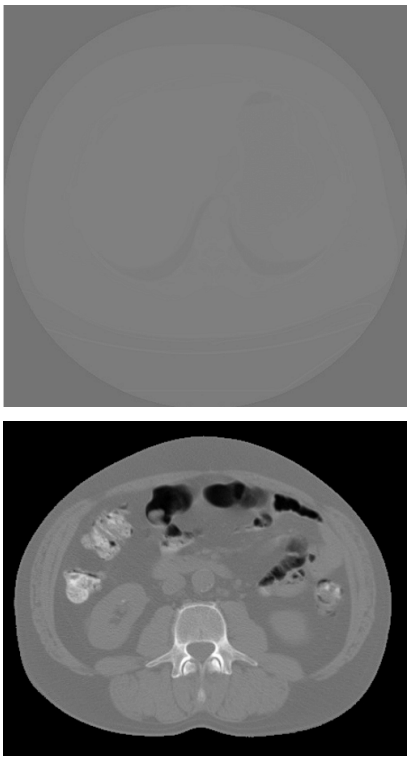

Figure 1: Abdomen CT scans: (Top) with lesions, (Bottom) without lesions

3.2 Data Bias

While training a CNN to classify lesions using the Mendeley Abdominal CT Scans and NIH DeepLesion datasets, it is important to address potential biases that may affect model generalizability. The Mendeley dataset, focused on stomach cancer detection, contains axial CT scans from a single hospital in Iran, which risks geographic and demographic bias (like limited representation of age, sex, or ethnic diversity). Similarly, while the NIH DeepLesion dataset includes over 32,000 annotated lesions across diverse regions (bone, liver, kidney, etc.), its reliance on data from a single American institution raises concerns about data sourcing (for example, differences in CT protocols / scanner differences) and under-representation of rare lesion types or patient subgroups. Additionally, as we will be using a small subset of the data, we risk introducing sampling bias.

3.3 Data Preprocessing Pipeline

Data preprocessing is at the heart of medical image analysis, especially, when dealing with heterogeneous data such as CT scans. Each step of the pipeline has to not only allow the model to be fine-tuned but also preserve the clinical interpretability of input data. Preprocessing was particularly well thought out for this project so as to be consistent, fair, and explainable for all stages. The pipeline was comprised of resizing, grayscale conversion, contrast enhancement, pixel normalization, data augmentation, and structured storage.

3.3.1 Image Resizing. To standardize input dimensions across different CT scans, which vary in resolution due to imaging protocols and device specifications, all images were resized to 128×128 pixels using bilinear interpolation. This resolution was determined to balance computational efficiency with the preservation of relevant spatial features, such as lesion features and tissue boundaries, while avoiding the diminishing returns of higher resolutions (like 256×256) that disproportionately increased training overhead without improving accuracy or explainability. The choice of bilinear interpolation was made to facilitate smoother transitions compared to nearest-neighbour methods, a critical factor for interpretability tasks like Grad-CAM visualization. This preprocessing step not only optimized model convergence by reducing convolutional complexity but also enhanced the reproducibility of XAI outputs, paving the way for meaningful saliency maps.

3.3.2 Greyscale Conversion. CT scans are naturally greyscale because they show differences in tissue density, which relate to how much radiation is absorbed. However, due to how some datasets are saved or displayed, a few of our images were stored in RGB format, adding unnecessary colour channels. To make the data more consistent and easier for our model to learn from, we converted all images to greyscale using Python's PIL library. This not only reduced the size of the data, which saves memory and speeds up training, but also made the model focus on important visual features like shape, texture, and contrast, which are key for spotting lesions. Greyscale images also improve explainability tools like saliency maps, since they reflect true intensity patterns rather than artificial colour differences.

3.3.3 Contrast Enhancement. CT scan images often show differences in contrast due to variations between patients, scanning machines, or radiation settings. This can be a problem in areas like the abdomen, where soft tissue makes it hard to clearly see lesion edges. To fix this, we applied histogram equalization using the ImageOps.equalize() function from the Python Imaging Library (PIL). This method adjusts the brightness levels across the image to improve overall contrast, making it easier to spot subtle features like lesions. More importantly, this step helps ensure fairness during model training by reducing the risk that the model favours images from higher-quality scans. It also makes the model more reliable

when used on real-world data and improves the clarity of visual explanation tools like Grad-CAM, which rely on consistent image quality to highlight important areas.

3.3.4 Pixel Normalization. To make the input data more consistent and suitable for training, we normalized all pixel values from their original range of [0, 255] to a range of [0, 1] by dividing each pixel by 255.0. Neural networks work best when inputs are on a similar scale, as large differences can lead to problems like vanishing or exploding gradients during training, which makes learning unstable. Normalization also plays a key role in producing accurate explanations with tools like saliency maps and Grad-CAM, which rely on proper gradient calculations. Keeping all inputs on the same scale helps the model learn more effectively and makes it easier to understand how the model is making decisions. It also supports transparency and reproducibility, ensuring that future users or researchers can trust the training setup and build on it with confidence.

3.3.5 Data Storage. To support modular development and ensure reproducibility, all preprocessed images were saved into two clearly labelled directories: abdomen_with_lesion_preprocessed and abdomen_without_lesion_preprocessed. Each image was given a unique name using a consistent scheme (e.g., image_0.png, image_1.png), preventing any overlap or duplication. This naming approach allowed for seamless integration with Keras' flow_from_directory() function, which automatically assigns class labels based on folder names and supports efficient batch loading and augmentation. Organizing the data in this structured way also enhances accountability and traceability—key principles in Human-Centered AI. It enables data scientists or reviewers to easily inspect samples, repeat experiments, and verify preprocessing steps by simply navigating to the appropriate folder.

4 Model Development & Evaluation

In the development of this project, five different models were created and analysed. The goals were not aimed at just a high accuracy rate, but rather to critically explore a balance in these performances of explainability and ethical integration of AI in medical imaging. Initially, we started with kernel and batch size investigation, and then started with a deliberately shallow CNN, after discovering that kernel size of 3x3 and batch sizes 24 & 32 were most suitable.

4.1 Model A

The CNN used in this study was intentionally designed as a shallow, black-box model to maintain simplicity and interpretability. It included a reshape layer to convert greyscale images into a 3D format compatible with convolutional processing, followed by a single convolutional layer with 12 filters (3×3 kernel size). This was succeeded by a max pooling layer, a flattening operation, and a final dense output layer with two units for binary classification. The model was compiled using the Adam optimizer and trained with a sparse categorical cross-entropy loss function. Remarkably, it achieved a high test accuracy of 97.75% after just four training epochs. However, such performance should be interpreted with caution. Given the model's simplicity and limited training time, these results may not reflect true generalization but rather suggest the presence of confounding patterns or shortcuts in the data. While the model retains transparency due to its straightforward architecture, its high accuracy likely reflects dataset artifacts rather than clinically meaningful learning.

4.2 Model B

The architecture was extended to include two convolutional layers with 32 and 64 filters respectively, each followed by a max pooling operation to progressively reduce dimensions. To mitigate overfitting, a dropout layer with a rate of 0.2 was added after the convolutional blocks. This was followed by a dense layer with 32 units, and finally a softmax output layer for binary classification. The increased depth of the model enabled it to capture more complex lesion features, contributing to a stronger representational capacity. With early stopping applied, training was allowed to proceed for up to 30 epochs, during which the model consistently achieved near-perfect validation accuracy, even reaching 100% in multiple runs.

However, this gain in performance came at the expense of transparency. As the number of layers and parameters increased, so too did the model's complexity, making it more difficult to interpret the learned features or understand which patterns were driving classification decisions. This reduced interpretability poses challenges in medical contexts, where explainability is crucial for trust, accountability, and clinical decision support. The trade-off between performance and interpretability underscores the importance of aligning model complexity with the needs of the end users, especially in sensitive domains like medical imaging.

4.3 Model C

We implemented transfer learning using MobileNetV2, a lightweight CNN architecture originally pre-trained on the ImageNet dataset. To adapt it for our binary classification task, we froze the base convolutional layers, preserving the pre-trained feature extraction, and trained only the top classifier head. This approach allowed us to leverage the model's strong generalization capabilities without the computational cost of training from scratch. Despite MobileNetV2 being optimized for natural RGB images, it still managed to perform well on our greyscale CT scan data after resizing inputs to 224×224 and duplicating channels to match the expected input format. The model was trained using the Adam optimizer and binary cross-entropy loss function.

While this setup produced stable performance across different data folds and appeared effective at capturing subtle structural features in the scans, this may be more a result of ImageNet biases rather than genuine medical relevance. Moreover, because MobileNetV2 is a deep and highly abstract architecture, interpretability was significantly limited. Its "black-box" nature made it difficult to understand which features were driving predictions, a notable drawback in medical AI applications, where explainability is critical. Overall, although the transfer learning approach was computationally efficient and delivered strong results, it came with trade-offs in terms of transparency and potential misalignment with domain-specific imaging features.

In addition to these deep learning models, we also evaluated classical machine learning techniques for comparison.

4.3.1 Logistic Regression: Had been trained based on flattened pixel features, but provided rather good performance for a linear classifier. However, it failed to extract the spatial characteristics that are crucial for classification of medical images and worked far inferior to CNN-based models.

4.3.2 K-Nearest Neighbours (KNN):. was also introduced which offered more neighbourhood-based contextual decision. Despite the fact that its performance was superior to that of logistic regression in recall, the cost in computation during inference and poor scaling made it unsuitable for real-time applications. Although they are rather simple these models were used as control baselines to measure the true effect of utilizing the CNN architectures in the detection of lesions.

4.4 Performance Evaluation

We then benchmarked five different models (that ranged from KNNs to deep learning architectures using transfer learning) and performed an extensive performance evaluation using only the metrics that were yielded by the models in this project. This comprised of the training accuracy, validation accuracy, training loss and validation loss as well as the test accuracy for unseen data. Where appropriate, qualitative insights regarding model behaviour were obtained by incorporating things like accuracy/loss graphs and performance comparisons, such as in the case of visualizations.

4.4.1 ***Model A***. A shallow CNN for interpretability, improved consistently throughout the training for four epochs. Training accuracy rose from about 67% to 98.46%, validation accuracy went up from 70% to 97%. The model also had a final validation loss of 0.0651. The simplicity of the model that could restrict its ability to learn very abstract features also helps in explaining why it is stable, fast, and has clean training curves, which is strong on small-scale, explainable applications.

4.4.2 ***Model B***. A larger CNN with two convolutional layers and dropout regularization performed better. From a training accuracy of 63%, it increased sharply, reaching 99% within 10 epochs. Validation accuracy also attained and stayed at 100% from epoch 4 onward. Loss dropped rapidly toward zero, implying strong convergence and high confidence in the model. To avoid overfitting, early stopping was applied with a patience of 5. The repeated high validation accuracy and low loss suggest that the architecture was able to capture important spatial features of the CT images.

However, this seemingly perfect performance must be interpreted with caution. The CNN (Model B) reported 100% test accuracy and 100% sensitivity/specificity, but such metrics are highly unlikely in real-world medical AI settings, where even state-of-the-art models typically achieve between 80% and 90% on diverse, clinically representative datasets. The consistent high scores indicate that the results may be driven more by dataset limitations than by genuine model efficacy. The small test set (only 44 samples) means that even a single misclassification would shift accuracy by about 2%, making the results fragile. Preprocessing steps may also have introduced unintended patterns, such as stark contrast differences, which the model could exploit as shortcuts rather than learning medically meaningful features like lesion location or texture.

While the architecture demonstrated impressive technical results, these are likely a product of controlled conditions and not a reflection of clinical robustness. In practice, lesion classification requires validation across diverse datasets from multiple institutions, with varying imaging equipment and protocols. Even a minor drop in accuracy can have serious implications. Therefore, while promising, the model's performance should be viewed as an encouraging early step rather than clinical evidence, and further development and broader validation are essential.

4.4.3 ***Model C***. Using MobileNetV2 through transfer learning, also managed to achieve 100% accuracy in test. It was trained for 10 iterations and used mini-batch size equalling 3 and learning rate equalling 0.0001. While having a lower file size than the baseline CNN (̃20MB), the model's inference time per image was a little longer. The last accuracy was on par with Model B, but loss remained slightly higher. This performance trade-off indicates that transfer learning can generalize well despite using natural images to medical imaging tasks, although interpretability, efficiency may be at a stake.

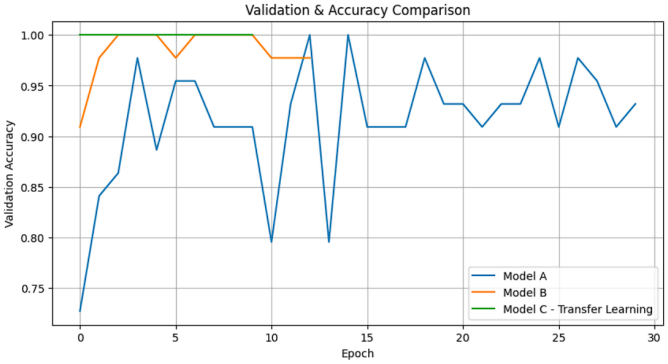

Figure 2: Model A, B, and C accuracy comparison

4.5 Unseen Data Testing

To evaluate the generalizability of final model, it was tested on a fresh set of ten images of abdomens which were not in the training, validation and test sets. Although the model had performed the predictions successfully, the output indicated a significant difference between the training performance and the practical functionality of the model. The calculated accuracy was only 40%, and the Root Mean Squared Error (RMSE) was 0.7746, meaning that any variation of the verification skills of the model was observed within new samples. This demonstrates the difficulties that are inherent in the process of medical AI development especially whereby the use of limited data sets is concerned. The finding highlights the need for future deployment to consider external validation, diversity in the datasets, and the inclusion of uncertainty estimation or confidence metrics in a future deployment. Instead of relying solely on internal metrics, this phase helped find edge cases and cases where the model's prediction would require some oversight from clinicians or fine-tuning.

5 XAI Implementation

In this project, we applied two XAI techniques namely; Saliency Maps, and LIME so as to interpret and analyse the predictions made by the best performing model, Model B. These techniques were critical in understanding which component areas of the CT scan had affected the decision of the model, which corresponds to the human-centered and transparent AI aspect of the project.

Our saliency map results provide a direct visualization of which pixels most influence the model's predictions on unseen data by computing the gradient of the output with respect to the input image. Saliency maps are particularly valuable in medical imaging like ours, where focusing attention to specific regions is critical for clinical trust & explainability. In our visualizations, the lesion images on the bottom row show high activation outside the organ, and only some within the abdominal organs. This suggests that while our classifier is focusing on some relevant regions to make its decisions, it's main focus is on the irrelevant background. The highlighting implies that the model has learned only some of the internal representation of lesion features, rather than true prediction areas like:

- Texture irregularities
- Density anomalies
- Shape distortions within the organ boundaries
- Overall shape of the abdomen

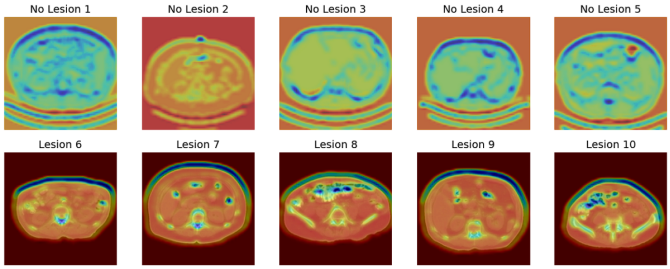

Figure 3: Saliency Maps for unseen data

Similarly, the saliency maps for non-lesion images show a notable amount of activation outside the organs. These red-activated areas are found near the image borders or at the extreme edges of the images. We observed that this pattern may indicate that our model is overfitting to background noise or imaging features that are unrelated to lesion classification, probably due to the multiple datasets used. Instead of focusing on organ structures when determining that a scan is healthy, the model seems to be relying on features irrelevant to the abdomen itself. This type of drift is concerning, as it can lead to poor generalization, severe overfitting or failures in edge-case scenarios.

Our LIME results reveal important insights into the interpretability and limitations of our model's decision-making process. In non-lesion images, LIME successfully highlights super-pixels within the anatomical regions, such as the liver, kidneys, and surrounding soft tissues, where lesions are expected to happen. These regions are marked in red, which indicate a strong contribution to the model's "positive" classification. This alignment between the highlighted regions and the actual lesion images suggests that, for these specific cases, the model has learned to associate relevant internal organ features, which is a promising sign of pattern recognition and that the model is functioning as we intended. In contrast, the LIME maps for lesion images display red activation zones primarily outside the organ boundaries, often in the background or along image edges. This is concerning, as it implies that our model is relying on irrelevant features for classification. This behavior raises the possibility of dataset bias, such as differences in image padding, data source inconsistencies, or unsuitable preprocessing between lesion and non-lesion samples. As LIME explanations are local, and change the image using super-pixel modifications, the fact that these external regions are repeatedly highlighted as important indicates that the model might not have a comprehensive, diverse representation of what a healthy abdomen "should" look like. Instead, it may be exploiting shortcuts or unintended correlations in the training data (most likely case). These findings reinforce known limitaions of LIME:

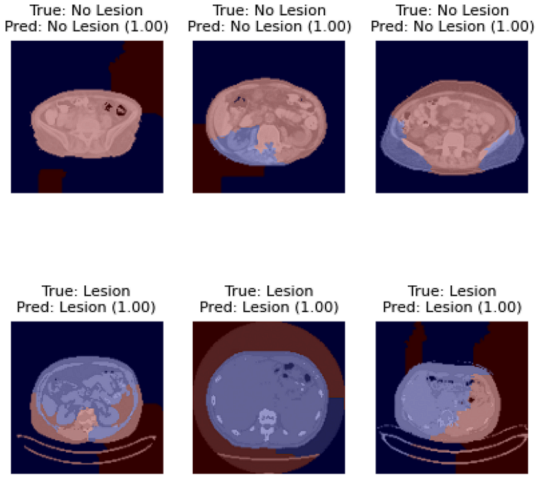

Figure 4: LIME output

- Highly dependent on local sampling, and the explanations are sensitive to how features.
- Since LIME assumes that the model's decision boundary can be approximated locally with a simpler model, its effectiveness diminishes in high-dimensional spaces like medical imaging and CT scans, where tiny pixels often carry critical meaning.
- Additionally, if the feature space that is generated through super-pixel segments does not reflect understandable regions, the results may be misleading.

Both of the methods revealed an important deficiency: sometimes non-anatomical background features were used for classification. This can be a result of combining several datasets with different styles of preprocessing or background. It also indicates that though the model learnt some internal lesion representations, it failed to utilize valid features in all the cases.

5.1 Model Explainability Discussion

Model B had high accuracy during training and validation but when using XAI techniques we saw that performance metrics will not necessarily ensure trustworthiness. From our saliency map overlays and LIME segmentations, we found that the predictions, in some cases, were based on appropriate medical areas while in other cases, were clearly affected by artifacts or non-relevant pixels. This inconsistency is a big problem for real deployment in clinical settings, where reliability and explainability are of crucial importance.

According to the XAI outputs, the model has partially learned the concept of lesions but does not show robustness in diverting from the clinical indicators to spurious. This brings out the significance of combining global and local explanations; Saliency Maps and LIME. Together, they converted our model from a black-box in a more interpretable "grey-box", providing better understanding of the model behaviour and enabling possibilities for user-in-the-loop decision-making.

This writing establishes a basis for scalable tools of diagnosis with ethically aligned. Whereas we only concentrated on abdominal CT lesions, the same explainability infrastructure could be further extended to other organ systems (lungs, liver, brain) or methods of imaging (MRI or PET scans). Also, we found that there is room for improvement with regard to the lack of model confidence scores or estimates of uncertainty. Furthermore, using XAI to find flawed attention and a tool for bias detection, this will guarantee the considerations to end users needs as well as expected legal systems such as the EU AI Act that requires high risk systems such as medical diagnostics to be transparent, auditable and fair.

From the human-centered perspective, this dual XAI approach allowed to come to the surface risks and allowed for transparent reporting, both of which are critical within such frameworks as EU AI Act. Without these explainability tools, we could have overestimated the reliability of our model just by the bare metric of accuracy. The findings of this part come as a reminder that in medical imaging, interpretability is never optional, rather a precondition for confidence and safety.

6 Model Compliance: Risk & Solutions

Due to high-risk nature of medical AI systems, this project was evaluated against major rules of law and ethical principles such as the GDPR and the EU AI Act, and generic responsible AI principles. Several risk areas were identified.

To begin with, fairness was of major consideration. NIH DeepLesion and Mendeley Abdominal CT datasets belong to the publicly available and anonymized collections, but are sourced from a limited number of places. a hospital in Iran and one U.S. institution. This adds geographical and demographic bias creating an underrepresentation of variations in terms of sex, ethnicity, and age. In addition to that, the project selected a sub-set emphasizing the area of abdomen, thus limiting the diversity further. The effect is that the trained model might produce unbalanced performance on populations, a factor that has the potential to undermine fairness and may even attracts legal consequences through anti-discrimination legislations.

Second, the project focused on the explainability (XAI), which is an aspect required by the EU GDPR's "right to explanation" and EU AI Act's transparency requirements. Tools such as Saliency Maps and LIME were incorporated in order to assist visual interpretation of predictions. However, outputs of these tools showed that the model sometimes made decisions based on non-relevant areas (e.g., padding or image's corners). Although the presence of XAI is a strength, its insights also revealed gaps existing in the model's reasoning.

To deal with this, more explanation consistency tests, the human-in-the-loop validation and data curation techniques could be used to ensure models can rely on medically valid features. Some of the ways, however, of determining when the model deviates from what is expected of its decision pattern is to use continuous testing with actual CT scans.

6.1 Emerging Trends & Techniques

In this project, numerous practices as well as design principles were combined which also align with emerging trends about the future of AI, specifically in healthcare-based systems. One of the trends is the shift in the direction of the glass-box model design, simple, interpretable CNN architectures with the native explainability. For more complex architectures, including Model B, the application of Saliency Maps and LIME was able to move the model from the less auditable, non-human understandable ones to a more human-understandable zone.

Further, the experiment with model pruning in Model D – though incomplete – demonstrates interest in edge-AI deployment; compressed models can run on low resources devices, such as portable scanners or embedded diagnostic devices.

Although not implemented, this project also sets the way to a model uncertainty estimation, an increasingly required element in AI safety. Special techniques such as Monte Carlo Dropout or Bayesian CNNs can be added for future updates which would output not only the predictions but confidence intervals which are highly essential for healthcare of the highest stakes. These forward-looking integrations of the project reflect the synchronicity of the project with developing vision of human-centered, auditable and resilient AI.

6.2 Ethical Considerations

To assess compliance with the EU AI Act and GDPR, we examined our datasets, model design, and workflow for risks related to fairness, transparency, and explainability. Since our lesion classifier would be considered "high-risk" under the EU AI Act, it is subject to strict obligations including robust documentation, human oversight, and explainability. Our datasets (NIH DeepLesion and Mendeley Abdomen CT) were ethically sourced and anonymized, but concerns remain about demographic diversity, which could affect generalizability and fairness.

CNNs are inherently opaque, and a lack of interpretability poses a legal and ethical challenge when used in clinical workflows. To address this, we implemented explainability tools like LIME and GradCAM, which offer insight into what the models "see" when making predictions—crucial for clinical trust and patient communication. This traceability supports not only reproducibility but also compliance with the conformity assessment obligations under the AI Act.

Our project followed ethical guidelines outlined by the European Commission's High-Level Expert Group on AI, including respect for human autonomy, fairness, and the prevention of harm. We ensured that clinicians, not algorithms, would retain final decision-making authority. Throughout development, we maintained human oversight, held regular team meetings to document progress, and implemented early stopping and validation routines to guard against overfitting. However, the use of small, balanced datasets and the rapid convergence of CNN models raise concerns about shortcut learning and overly optimistic performance, which must be revisited in future work.

6.3 Legal & Regulatory Compliance

Current legal landscape of using the Artificial Intelligence in healthcare is changing dramatically, and that occurs very quickly in the European Union, where legislative norms, such as the General Data Protection Regulation (GDPR) and soon to come into force the EU AI Act, are dramatically rewriting rules on how Artificial Intelligence-powered medical systems must be developed, deployed, and In preparation for these legal changes, this project also incorporated various foundational practices that are consistent with key precepts of regulation and compliance expectations.

First, the dataset used for this project–curated from the NIH DeepLesion and Mendeley Abdominal CT Scan repositories— was fully anonymized and ethically sourced so that no personally identifiable information (PII) and protected health information (PHI) was included. This observes the cardinal GDPRs:

- **Data Minimization**: Only the information necessary to train a lesion detection model was collected and used.
- **Purpose Limitation**: The data was used strictly for non-commercial, educational, and research purposes, with no intent for clinical decision automation or commercialization.
- **Transparency**: All sources were publicly disclosed, and data transformations were thoroughly documented in the accompanying notebook.

In terms of the EU AI Act, medical AI applications are considered of the "high-risk" category and so these points need to be ensured:

6.3.1 ***Traceability:*** All the processes that took place in the model development (preprocessing workflows, CNN architecture definition, training parameters, training procedures, epochs, optimizer, loss function, post – training evaluations) were clearly documented in the Jupyter notebook. This is not only able to ensure complete traceability and audibility of how the model was built, trained, and evaluated but also result in reliability and predictability.

6.3.2 ***Explainability:*** The process of combining Saliency Maps and LIME explanations is one of the biggest steps toward achieving transparency under the AI Act. Such tools enable developers, clinicians, and even regulators, see what aspects of input data affected the decisions made by the model.

6.3.3 ***Risk Management:*** The risks associated with the bias of data, imbalance of dataset and the possible overfitting were explicitly stated in the scope of the project, especially in the context of XAI techniques. That some predictions were based upon irrelevant regions of images (padding or background) shows that the system may be exploiting non-medically valid features. By finding this problem with the help of explainability tools and suggesting the class balancing strategies and the improved preprocessing as prevention, the project took first steps to comply with the risk management framework that must be followed. Full compliance with the **EU AI Act** in a production environment would require additional components:

- **Conformity Assessment**: A formal validation process, consisting of internal documentation and third party audits to certify that the AI system complies to every regulatory standard. This was beyond the limits of this project.
- **Version Control**: Retrained models should be well versioned for traceability and roll back. Although only one cycle of training was performed in this project, all configurations were still recorded for reproducibility.
- **Fail-Safe Human Oversight**: The AI system should not be used in a way that it functions autonomously in high risk situations. Our design allows for human-in-the-loop operations.

7 Summary

This project investigated the development and explainability of a deep learning model in the form of a CNN for classifying lesions with the use of CT scans of abdomens. Using a filtered subset of NIH DeepLesion and Mendeley abdominal CT datasets, we performed targeted preprocessing, resizing, grayscale conversion, and contrast equalization on the input images to standardize the input. Model B, a deeper CNN with dropout regularization, produced the best validation performance measured by both the accuracy and loss metrics. Further comparisons with a transfer learning model (Model C) and baseline classifiers such as Logistic Regression and KNN helped in gaining knowledge regarding trade-off between complexity and interpretability.

In order to review the model decision making process, Saliency Maps and LIME were implemented. Saliency Maps delivered global explanations gradient-based whereas LIME delivered local interpretability by perturbing the image and spotting most influential superpixels. These tools showed that the model sometimes used background or non-clinical areas, hence the utility of explainability in the revelation of model weaknesses.

Conversations on glass-box versus black-box models brought out the fact that the traditional ML models come with built-in interpretability while CNNs necessitates the use of other techniques in order to be transparent. As a result of XAI integration, the project succeeded in moving the deep learning models into "grey-box" systems- facilitating effective human oversight in the clinical settings.

The project reflects the principles of responsible and human-oriented AI that focuses on the transparency, equity, and safety of medical applications. Future efforts can address the increase of dataset diversification, the introduction of confidence scores (or uncertainty estimates), and the use of more elaborate or hybrid XAI framework. These steps will guarantee further progress towards the creation of ethical, legally compliant, and trustworthy AI systems for lesion detection and more varied clinical use.

References

[1] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017 Dec;42:60-88. doi: 10.1016/j.media.2017.07.005. Epub 2017 Jul 26. PMID: 28778026.

[2] Yan Y, Mao Y, Li B. SECOND: Sparsely Embedded Convolutional Detection. Sensors. 2018; 18(10):3337. https://doi.org/10.3390/s18103337

[3] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[4] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

[5] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

[6] Arun, K. B., Madhavan, A., Sindhu, R., Emmanual, S., Binod, P., Pugazhendhi, A., ... & Pandey, A. (2021). Probiotics and gut microbiome Prospects and challenges in remediating heavy metal toxicity. Journal of Hazardous Materials, 420, 126676.

[7] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.

[8] Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), 97-112.

[9] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. In Proceedings of the 2019 chi conference on human factors in computing systems (pp. 1-13).