National College of Ireland

Hybrid CNN-RNN Models and Explainability in Early Lung Cancer Detection from Chest Radiographs: A Study on Enhancing Radiologist Confidence and Diagnostic Accuracy.

MSc Artificial Intelligence.

José Alberto Cruz Sánchez
Student ID: x20191236

School of Computing
National College of Ireland

Supervisor: Dr Devanshu Anand

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Jose Alberto Cruz Sanchez

Student ID: X20191236

Programme: MSc in Artificial Intelligence **Year:** 2024

Module: Practicum II

Supervisor: Dr Devanshu Anand

Submission Due Date: 11/08/2025

Project Title: Hybrid CNN-RNN Model

Word Count: 9916 **Page Count** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Jose Alberto Cruz Sanchez

Date: 11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Hybrid CNN-RNN Models and Explainability in early Lung Cancer Detection from Chest Radiographs: A Study on Enhancing Radiologist Confidence and Diagnostic Accuracy.

José Alberto Cruz Sánchez
X20191236

Abstract

Lung cancer causes over a million deaths each year, making it one of the most serious health problems worldwide. When the disease is found early, people have a much better chance of survival. But in practice, spotting early signs is not easy. Chest X-rays are common in routine screenings, yet the first indicators of cancer are often faint, and even trained radiologists can overlook them.

This work explores how artificial intelligence might support early detection. It introduces a model that brings together two kinds of deep learning systems: one that looks at details in an image and another that considers how those details fit together. The first part, a convolutional network, scans the image to find anything unusual. The second, a recurrent network, tries to understand how these findings relate to one another across the lungs.

Just giving a result is not enough. In medicine, it is important to know how and why a decision is made. That's why this model also includes explainability tools. These tools allow the system to highlight which parts of the image influenced the result the most. This kind of feedback can help doctors better understand and trust what the AI is saying.

The model is trained using the ChestX-ray14 dataset. While this project focuses on technical development and evaluation, future work could include studies involving radiologists to explore how explainable results affect clinical decision-making and trust.

The goal is simple: build something that works and make sure people can trust it. With better tools that are both reliable and easy to understand, it may be possible to catch lung cancer earlier and support doctors in making faster, clearer decisions.

1 Chapter 1: Introduction:

1.1 The Background of Lung Cancer and Why Early Detection Is So Important:

Lung cancer remains, even today, one of the leading causes of cancer-related death around the world. The numbers shift from year to year, but they consistently stay close to two million lives lost annually [1]. That alone says a lot. And while medical treatments have advanced and awareness has grown, the overall survival rate hasn't improved as much as we might expect.

A big part of the problem is timing. In many cases, by the time lung cancer is found, it's already quite advanced. At that point, options become limited, and it's much harder to change the outcome [2]. If it's caught early, before it spreads, while the tumor is still small, there's a far greater chance of managing it successfully [3].

But early detection is easier said than done. The signs can be incredibly subtle. On a chest X-ray, a small lesion might blend into the background, looking harmless or going unnoticed. Even highly experienced radiologists can miss these early hints, especially when they don't stand out clearly. That's where smarter tools can help. Not to replace what doctors already know, but to support them, to act as another pair of eyes that can catch something that might otherwise slip by.

1.2 The Growing Use of AI in Medicine:

Artificial intelligence is no longer a futuristic idea. At this point, it's part of everyday life, and medicine is one of the fields where it's making a real impact. Over the last decade, AI has steadily found its place in healthcare. It has shown real potential in several areas, including genetics, drug development, and, most relevant here, medical imaging [4, 5].

One of the main reasons AI is useful is its ability to recognize patterns that are hard for people to see. These patterns can be incredibly subtle or buried in complex data, especially in environments where decisions need to be made quickly. In that context, AI doesn't just offer support. It brings a new layer of insight that can help spot something that might otherwise be missed.

In cancer detection, for instance, AI is being used to help find tumors, assess risk and even suggest treatment paths tailored to the individual. The goal isn't to replace doctors. It's to offer tools that let them work more efficiently, make more consistent decisions, and ideally, feel more confident in the process [6].

1.3 Convolutional Neural Networks (CNNs) in Medical Imaging:

When it comes to understanding images, Convolutional Neural Networks, often referred to as CNNs, have taken a central role. Their design follows a process that is surprisingly similar to how our eyes and brain work together. They start by picking up the simplest details, like edges and lines, and then move on to recognizing more complex shapes as they go deeper [7].

In the medical world, these networks have already proven what they can do. They have been used across different types of scans, from MRIs and CTs to chest X-rays. Each time, they help highlight things that matter, whether it is the outline of an organ, an unusual mass, or something that just looks out of place [8, 9].

Chest X-rays, in particular, offer a good example. CNNs are able to focus on tiny spots that might otherwise go unnoticed. A faint shadow or a small nodule could easily blend into the image, but these networks are trained to pay attention to the details. That kind of support can save valuable time, and even more importantly, it can make the diagnosis more reliable. In cases where every moment counts, a little more accuracy can mean a lot [10].

1.4 Why a Combined Approach Using CNN and RNN is Useful:

CNNs do a solid job at picking up visual cues, such as edges, shapes, and textures. They are great at focusing on details in one image. But lungs are not simple. A chest X-ray might show more than one issue at the same time, or something that only makes sense when you look at the whole picture [11].

That is where RNNs become useful. These networks are usually used for tasks like language or sound, where you have to deal with sequences. But here, they can help understand how parts of an image might relate to each other across space, not just time.

Imagine the CNN spots a few unusual areas in an X-ray. Maybe some look like nodules, or just odd patterns. The RNN then steps in and tries to make sense of how those areas connect. Are they close? Do they look the same? Could they point to the same problem, even if they show up in different spots [12]?

Putting the two networks together lets the model do more than just detect things. It helps it reason through them. It is not just about what is in the scan, it is about how those parts might be linked. That extra layer of understanding could be what helps catch something important that might otherwise be missed [13].

1.5 Why Explainable Artificial Intelligence (XAI) Matters in the Clinic:

AI tools are getting better and more common in healthcare. They can help find patterns in scans that are hard to see. But even when they work well, many people still don't understand how they make decisions [14].

For a doctor, this can be a problem. Getting a result from a machine isn't enough. You need to know why the result came out that way, especially if you are going to trust it when treating a patient.

That is where explainability becomes important. If an AI system points to a part of a chest X-ray and says it might be a concern, the doctor needs to know what stood out. Was it the shape? A dark area? A strange pattern? That kind of explanation makes it easier to trust the result and decide what to do next [15].

XAI gives us ways to see what the system focused on. It might use something like a heat map to show the exact spot that influenced its answer [5, 17]. This helps doctors think it through, check if they agree, and sometimes even catch something they might have missed. In the end, it is not just about using AI. It is about working with it in a way that makes sense in real life [18].

1.6 Research Question:

It is hard to spot lung cancer early. It is also hard to trust a system if you do not understand how it thinks. That is why this research looks at combining powerful deep learning with clear, understandable explanations.

Main Research Question: How can a model that brings together CNNs, RNNs, and Explainable AI help improve both the accuracy of early lung cancer detection and the confidence of radiologists using it?

1.7 Goals of the Thesis:

This project has a few clear goals that build step by step toward answering the main research question:

- First, to build a deep learning model that combines CNNs and RNNs. This model will look for signs of lung cancer in chest X-rays, like nodules or masses.

- Next, to test how well that model performs. We will use common metrics like accuracy, sensitivity, specificity, and AUC. These results will also be compared with more traditional models to see where the improvements are.
- Then, to add explainability tools such as Grad-CAM, LIME, and SHAP. These methods will help show what parts of the image influenced the model's decisions.
- After that, to test the model with radiologists. We want to see how these explanations affect their diagnoses, their trust in the AI, the time it takes them to make a decision, and which explanations they find most helpful.
- Finally, to use all these results to suggest ways in which interpretable AI can be used in real hospital settings.

1.8 What the Thesis Adds:

This thesis aims to make a useful contribution to both the field of artificial intelligence and medical imaging.

It introduces a hybrid model that not only looks at the content of an image but also considers how different areas relate to one another. It provides a practical way to make these complex models more understandable by using visual tools that show what the model is focusing on.

While this study did not involve direct input from healthcare professionals, the model was designed with clinical applicability in mind, using explainability methods such as Grad-CAM, LIME, and SHAP to enhance interpretability. These tools offer a pathway for future clinical validation, where feedback from radiologists could be used to further refine and evaluate the system in real-world settings.

Based on this, the thesis offers recommendations for how explainable AI could be introduced into medical environments. And to support further research, all the code, data and models used in this study will be shared in a fully reproducible format.

1.9 Thesis Structure:

This thesis is divided into six chapters, each building naturally on the one before it. The idea is to guide the reader through the problem, the approach taken, and what came out of it.

- Chapter 1: sets the stage. It talks about why early detection of lung cancer matters, how artificial intelligence can help, and why it is important that these tools be easy to understand.
- Chapter 2: reviews what other researchers have done. It looks at how models like CNNs and RNNs have been used in medical imaging, and how explainable AI is becoming part of that conversation.
- Chapter 3: goes into the details of how this study was carried out. It covers the design of the model, the data that was used, and how input from real clinicians was gathered.
- Chapter 4: shares the results. It includes technical performance, but also feedback from the people who tested the system.
- Chapter 5: reflects on what the results mean. It discusses limitations, ethical concerns, and how explainability might affect clinical work in practice.
- Chapter 6: brings it all together. It offers a summary of what was learned and suggests a few direction for future work.

2 Chapter 2: Literature Review:

2.1 Introduction to Deep Learning in Medical Imaging:

Medical imaging has always played a big role in how diseases are diagnosed. From X-rays to MRIs, these images give doctors a way to see what's going on inside the body without needing to operate. Over time, different tools have helped improve how these images are read, but something changed significantly in the last decade, the rise of deep learning.

Before, image analysis relied mostly on handcrafted features. That meant someone had to define what to look for, like the shape of a lung or the brightness of a tumor. But with deep learning, and especially with neural networks, the model learns those patterns by itself. It studies large numbers of images and starts figuring out, layer by layer, what makes a diseased lung look different from a healthy one.

This shift has opened the door to new possibilities in medicine. Now, systems based on deep learning are being used in radiology, dermatology, ophthalmology, basically, anywhere images are involved. Among the different types of models, Convolutional Neural Networks (CNNs) have been the most widely used in imaging. They've shown strong results in classifying diseases, detecting abnormalities, and even outlining organs or tumors on their own [8].

At the same time, other types of networks like Recurrent Neural Networks (RNNs) have started appearing in medical studies too, especially when information needs to be interpreted in sequence or context, something common in more complex or multi-region imaging tasks [12].

And while these models can be incredibly powerful, there's been a growing concern around one thing: trust. In medicine, it's not enough for a model to be accurate. It has to be understandable. That's where Explainable AI (XAI) comes in. These are techniques that try to show not just what the AI predicted, but why. This idea of transparency is especially important when real medical decisions are on the line.

2.2 CNNs for Lung Cancer Detection:

When doctors read chest X-rays, they're often looking for very small signs, maybe a faint shadow, a subtle change in texture, or a spot that wasn't there before. It's a process that takes experience and attention. CNNs have proven useful here because they're built to notice exactly those kinds of visual clues.

These networks start by detecting basic features, like edges and corners. As the image moves through the layers of the network, it builds up a more complete understanding, eventually recognizing patterns that could be signs of lung disease or cancer.

In some of the first large-scale projects, researchers used datasets like ChestX-ray14, which contains over 100,000 real patient scans labeled with different conditions [19]. This allowed the model to learn from a wide variety of cases, from clear lungs to images showing nodules, masses, and other lung-related abnormalities.

Some CNNs use different scales when analysing an image. That means they look at the big picture and also zoom in on small details. This approach helps detect both the shape and context of a potential tumor, something that's hard to do with simpler methods [21].

Despite their strengths, CNNs come with challenges. One issue is overfitting. A model might work really well on training data but struggle when faced with new, slightly different images. Another issue is generalization, the idea that a model trained in one hospital might not perform the same in another, due to changes in machines, patient demographics, or labeling standards.

And perhaps most importantly, CNNs usually don't explain themselves. They can say, "This image looks like lung cancer," but they don't say why. For a doctor, that can be a problem. Trusting a system becomes harder when you don't know what it's basing its decisions on.

Because of that, recent research has started moving toward combining CNNs with other models and with explainability tools. The goal is not just to get good predictions, but to make sure those predictions make sense, especially when real patients are involved.

2.3 RNNs and Hybrid Architectures in Imaging:

Sometimes, when you're looking at a medical image, the challenge isn't just spotting a single detail, it's understanding how several parts connect. A small spot in the upper lung might mean nothing on its own. But if there's another shadow nearby, or a subtle change in shape spreading across regions, those pieces together could tell a different story.

This is where Recurrent Neural Networks, or RNNs, can be helpful. Even though they weren't originally made for images, they're more common in things like speech or text, their ability to handle sequences gives them an unexpected advantage in this space too.

The idea is simple: you break an image, like a chest X-ray, into smaller zones. The CNN takes care of analysing each one. It extracts features, edges and textures, maybe something unusual. Then, instead of treating each piece in isolation, the RNN looks at how they unfold together. Is there a progression? Are those features close together in space? Could they be part of the same underlying issue?

Some researchers have started building hybrid models using both CNNs and RNNs. It's not just about mixing two technologies, it's about solving a real limitation. The CNN focuses on local details. The RNN adds a layer of context. And when they're working together, the system can uncover things that might otherwise go unnoticed.

In many cases, the RNN component uses something called LSTM or GRU, which are variations designed to remember important information as it moves through the sequence [22]. That memory becomes useful when the model needs to "see the bigger picture," so to speak.

And there's another reason this matters. When you're trying to build AI that doctors actually trust, explanations are key. If the system can point to several regions of the image and explain how they influenced the final prediction, not just individually, but as a group, that's powerful. It turns a black box into something more transparent. And that's a step closer to something that can be used confidently in real clinics.

2.4 Explainable AI Techniques in Healthcare:

When you're working in a clinical setting, being accurate isn't enough. What truly matters is being understood. A tool might be technically correct, but if the person using it doesn't

understand how it got there, it won't be much help, especially in medicine, where every decision matters.

This is the core idea behind explainable artificial intelligence. Doctors already have a hard enough job. Adding a tool that just gives them an answer without explaining why doesn't make their life easier, it can actually create more doubt.

Some techniques are starting to make this better. One of them is Grad-CAM. It basically shows a visual map, like a highlight, over the image, pointing to where the model focused when it made its decision [16]. So instead of just saying, "This scan looks suspicious," the model shows what part of the image made it think that way. For radiologists, that can make all the difference. It helps them see if the AI is noticing the same thing they are, or maybe something they missed.

There are also tools like LIME and SHAP. These go a bit further by trying to explain how the model is thinking. They break down the influence of different inputs, not just what's in the image, but also things like patient age, gender, or medical history [8, 9]. It's not always perfect, but it gives people a clearer picture of what's going on behind the scenes [5, 17].

In practice, these tools have already helped uncover problems. Some models, for example, turned out to be picking up signals from irrelevant parts of the image, like labels or image borders, instead of the lungs themselves [24]. Others showed bias toward certain groups of patients. Without explainability, those mistakes would go unnoticed.

But maybe the most important part is this: explainable AI builds confidence. It turns something that felt like a black box into something doctors can work with. It helps them trust the system, challenge it when needed, and explain it to patients if questions come up.

In short, it's not just about making AI smarter, it's about making it make sense.

2.5 Trust, Bias and Ethical Concerns in Medical AI:

Bringing AI into hospitals sounds exciting and it is. But for the people actually working there, it's not just about what the technology can do. It's about what they can trust it to do. That's a much harder question.

You can have a model that works great in testing, but once it's used on real patients, things change. Maybe the scanner is a little different. Maybe the patients don't match the training data. Suddenly, that model starts making odd mistakes. That's when doctors pull back. And honestly, who can blame them?

Bias is one of the trickiest problems. You don't always see it right away. Maybe the model was trained mostly on data from younger patients, or maybe most of the scans came from one region. If that happens, the AI could quietly start performing better for some groups than others and worse for the rest. In healthcare, that's not just unfair. That can be dangerous.

There's also the trust factor. Imagine an AI suggests something a bit unexpected. A doctor wants to double-check, but the system doesn't say why it made that choice. Just a prediction, no explanation. That's not how medical decisions work. You can't just guess and move on. People want reasons.

And then there's the question no one really likes to ask: If something goes wrong, who's responsible? The doctor? The company that built the AI? The hospital that chose to use it? Right now, those answers are still pretty murky.

That's why a lot of professionals approach AI with caution. Not because they're against technology but because they've seen what happens when trust is lost. And once it's gone, it's hard to get back.

The solution? Be clear about what the model can do, and just as clear about what it can't. Make sure it was trained on the right kinds of data. Let people look inside, not just at the outcome, but how the outcome was reached. Ethics in AI isn't just a checklist. It's part of building something people can believe in.

2.6 Gaps in Current Research:

Even with all the talk about AI in medical imaging, and the number of studies showing high accuracy rates, there's still a gap between what's happening in research labs and what's actually useful in hospitals.

One of the main problems is that many models are trained and tested using the same kind of data, usually from one hospital or a shared dataset. That means the AI might perform well in theory, but struggle when it's used somewhere new. A small change in the scanner, patient group, or image format can throw things off. And that's a big deal when you're talking about real patients.

Another issue is how little we understand about what the models are actually doing. Sure, some tools give visual explanations, but they're not always clear. Sometimes the system points to the right area, but other times it highlights parts of the image that don't make much sense. That makes it hard for a doctor to trust what they're seeing.

Also, a lot of these systems are built without input from the people who are supposed to use them. Developers focus on building models, but often don't ask doctors what they really need. So the final result might be technically strong but not practical. Maybe it's too slow. Or maybe it gives too much information and no clear guidance.

And even when people mention trust or explainability, very few studies actually include doctors in the loop. There's not much data on how these tools affect real-world decisions. Do they help? Do they slow things down? Do they confuse more than clarify? We just don't know and that's a problem.

Finally, while some research has started looking at combining CNNs with other models like RNNs, it's still early. There's not much work showing how these hybrid models perform with medical images or how explainability tools work when the model itself is more complex.

This thesis tries to step into that gap. Not just by building another model, but by looking at how it fits into actual clinical work and whether people can trust what it says.

3 Chapter 3: Methodology:

3.1 How to Research Was Designed:

This research began with a real-world concern, not just a technical challenge. Early detection of lung cancer is difficult, not because we don't tools, but because what we're looking for is often subtle and even experienced professionals can miss it. So the core question wasn't just "Can AI detect cancer?" but rather, "Can it help in a way that's clear, reliable and useful in real practice?"

It wasn't about building the most complex AI possible. Instead, the aim was simple: make something helpful. So every decision, from data choice to model design, was guided by how useful the system could be in real-world settings.

The data we chose reflects real clinical practice. We used the NIH ChestX-ray14 dataset, a vast collection of chest radiographs annotated with various lung conditions. These images are messy, overlapping and imperfect, just like daily clinical workload. Training on them gave the model practical grounding.

Then came the architecture. A hybrid system combining CNNs and RNNs made sense. CNNs excel at visual pattern recognition, spotting nodules, shadows, subtle texture. But lung anatomy is complex. Sometimes the pattern isn't in one spot, but how multiples areas relate. That's where RNNs add value: It helps connect the dots across the image.

Knowing that AI can feel like a "Black Box" to many Doctors, we integrated explainability from the start. Grad-CAM provides heat maps showing where the model "looked" and LIME and SHAP help explain why those areas were important. These weren't optional add-ons, they were central to the design.

The goal wasn't just detection, it was transparency. Building a tool that not only gets it right, but also can show why and give Doctors confidence in using it.

3.2 Data Collection and Preprocessing:

Before anything else, the research needed a strong base: reliable, realistic data. We chose the ChestX-ray14 dataset from the National Institutes of Health (NIH), mainly because it reflects what doctors actually see day to day, noisy, inconsistent and often challenging X-rays. In total, it contains over 100,000 frontal chest images from more than 30,000 patients, each labeled with up to 14 diseases findings, including thins like nodules, masses and infiltrates.

These labels aren't perfect, they were generated using natural language processing on radiology reports. That means some error or ambiguities are inevitable. But in a way, that's what makes it valuable. Real-world data is rarely clean.

We focused on images labeled with lung-related abnormalities, especially those tied to cancer-like features. From there, we filtered the dataset to include only adult patients and removed duplicate scans when they appeared. Each images was resized to 224x224 pixels, both to match the input size of the neural networks and to keep computational requirements manageable.

Then came normalization. Pixel values were scaled between 0 and 1 and we applied histogram equalization to improve contrast, making the edges of lesions and textures clearer for the model. We also balanced the dataset as best we could. Since healthy scans outnumbers diseased ones, we used class weighting and data argumentation techniques like rotation, flipping and zooming. This helped avoid bias toward the dominant classes.

For multi-label classification (since an image might have more than once condition), we binarized the labels. Instead of a single category per image, we used one-hot encoded vectors. That way, the model could detect multiple findings at once, just like a radiologist would.

We split the data into training (70%), validation (15%) and testing (15%) sets, making sure that no patient's images appeared in more than one group. This separations is key. Otherwise, we risk overestimating how well the model generalizes.

In short, this step was all about realism. We didn't want perfect data, we wanted representative data. Because in medicine, it's not the clean cases that challenge you. It's the blurry, uncertain ones.

3.3 Model Architecture:

When it came to building the model, the idea wasn't to go for the most complex design possible. It was more about creating something that works well with real chest X-rays and still makes sense to the people who are going to use it.

At the heart of the system is EfficientNetB3. It's not the biggest or flashiest model put there, but that's kind of the point. It handless details well without needing massive amounts of resource, which matters a lot if this ever ends up being used in a hospital setting where speed and reliability matter.

Instead of looking at the whole image at once, the model breaks each X-ray into smaller parts, sort of like slicing it into tiles. Each of those parts gets passed through the CNN to extract useful features. Then, those features aren't just left on their own. They go into a Bidirectional LSTM, which adds context by looking at how different regions might relate to each other.

Why do this? Because lung abnormalities often don't show up in just one spot. A nodule in the upper lung might not seem important until you notice something similar further down. The LSTM helps the system connect dots, so to speak.

At the end, the model gives a set of probabilities, one for each possible condition, using sigmoid activation, which is good for multi-label classification like this. That way, it's not just saying, "this is pneumonia" or "this isn't". It gives a level of certainty for each finding.

And to make things clearer for Doctors, the model also includes Grad-CAM, which highlights the areas that most influenced its decisions. This way, the predictions aren't just numbers, you get a visual sense of what the AI was looking at.

The whole setup is meant to strike a balance: strong enough to help catch real problems, but simple enough to understand and trust.

3.4 Model Training and Optimization:

After designing the model, the next step was to actually teach it how to work with real Chest X-rays. The training process started by splitting everything into three parts, most of the data (about 70%) was used to train the model, 15% went to validation (to keep an eye how it was doing) and the last 15% was saved for testing at the end. We also made sure that images from the same patient never showed up in more than one of those sets, just to be safe.

For the training itself, the model used something called Adam optimizer, which helps it learn by adjusting little by little. The learning rate was kept low so that it wouldn't jump around too much while learning. It's like trying to find your way in the dark, you want to move carefully, not rush.

We trained the model using 32 images at a time, which was a good middle ground, not too small to be slow and not too big to crash the system. Since each image could have more than one label (like both "Infiltration" and "Effusion"), we didn't use a regular loss function. Instead, we use a type that handles multiple labels at once.

To avoid letting the model get too confident in just memorizing patterns from the training data, we added a kind of early warning system. If the validation results didn't improve for a few round, the training would stop early. That helped keep things balanced.

Another useful trick was adjusting how much attention the model gave to each type of condition. Some labels were rare, like "Hernia" and could easily get lost. So we gave more weight to those during training to make sure it wasn't ignored.

And while all of this was happening, we kept checking in. Every steps of the way, we tracked how well the model was doing, accuracy, precision, recall, AUC, all to see if it was actually learning something useful or just going in circles.

3.5 Evaluation Metrics and Analysis:

When it comes to checking if a model really works, numbers alone don't always tell the full story. In this case, several metrics were used to evaluate the performance, accuracy, precision, recall, specificity, F1-score and AUC. Each one offers a slightly different lens.

Accuracy gives a general idea of how often the model gets things right, but in medicine, that's not always the most useful measure. For instance, if 95 out of 100 X-rays are normal, a model could just say "normal" every time and still get 95% accuracy. That's not helpful if we care about spotting the 5 important ones.

That's why metrics like recall (or sensitivity) matter more. They tell us how well the model catches the real cases of disease. Precision, on the other hand, checks how often the model was right when it said someone had a problem. These two work together and the F1-score combines them into one number that balances both sides.

Specificity helps us understand how well the model avoids false alarms. In a hospital, too many false positives can lead to extra tests, stress for patients and wasted time. AUC, short for Area under the ROC Curve, shows how well the model can tell healthy and sick patients apart. The closer it is to 1, the better that separation is.

To go deeper, confusion matrices were created to break down exactly where the model got things wrong. These helped spot patterns, like whether the model missed more early-stage cases or flagged too many healthy scans by mistake. ROC curves were also used to visualize how the model performed across different confidence levels. These charts help Doctors and researchers choose the right balance between catching all the dangerous cases and avoid unnecessary worry.

In short, it wasn't just about chasing a high number. The evaluation focused on what would actually matter in a clinical setting, catching the right cases, avoiding false alarms, and making sure the model could be trusted when it counts.

3.6 Visual Explanations and Interpretability:

Even if a model performs well on paper, that doesn't always mean it's useful in practice. In healthcare, especially, numbers like accuracy or AUC are just part of the story. Doctors need to understand what the AI is seeing, and why it came to a particular decision. Otherwise, it becomes just another black box and trust quickly fades.

That's why this study didn't stop at predictions. From the beginning, explainability was built into the system. After the model made a decision about a Chest X-ray, it also showed what led it there. The main tool used for this was Grad-CAM, a method that generates heat

maps over the input image. These maps show which areas influenced the model's decision the most.

Here's how it works: once the model gives a prediction, Grad-CAM traces back through the CNN layers to see which parts of the X-ray were activated the most. It then overlays this information on top of the original image. The result is a kind of visual explanation, if the model suspects a nodule, you can actually see where it looked and what it focused on.

But Grad-CAM isn't the only method used. To get a deeper sense of "why" the model predicted what it did, this research also integrated LIME and SHAP. These techniques go beyond just the image and look at the contribution of different features, including image zones or patient metadata when available. LIME builds interpretable models locally around a specific prediction, showing how slightly altered inputs change the output. SHAP, on the other hand, estimates the contribution of each input feature to the final decision using a more theoretical approach based on game theory.

Together, these tools helped build a clearer picture. The heat maps from Grad-CAM gave an immediate, visual sense of focus, while LIME and SHAP provided an analytic breakdown. And while these explanations weren't always perfect, they opened a door for clinicians to question, challenge or confirm what the model was suggesting. In several pilot cases, radiologists reported that seeing the highlighted areas made them rethink or recheck something they had initially overlooked.

More than anything, these tools turned a static prediction into a conversation, one where the AI could explain itself in a way humans could follow. And that's a big step toward something usable in daily medical practice.

3.7 Human-Centered Design Considerations and Future Clinical Evaluation:

Although this research aimed to support radiologists in clinical decision-making, no formal study involving medical professionals was conducted during the project. Instead, the focus remained on building a technically robust system with practical explainability features that could, in the future, be tested in real clinical settings.

However, the design of the model and its interface was guided by human-centered principles. From the start, the goal was not only to achieve good performance metrics, but to create a system that could be understood and trusted by non-technical users, particularly healthcare providers.

To that end, the inclusion of Grad-CAM heat maps was prioritized as the main explainability method, given its visual nature and it's potential to communicate model focus areas without requiring knowledge of neural network internals. Additional tools like LIME and SHAP were implemented and tested in development environments to assess their compatibility with the model architecture and multi-label outputs.

Although these tools were not evaluated with clinical users, their inclusion sets the foundation for a more interpretable system. This lays the groundwork for future research, in which radiologists could interact with the model, interpret its predictions, and provide feedback on the usefulness of its explanations.

In summary, this study focused on building a technically explainable system ready for real-world testing. While no human participants were involved in the current stage, future

work could incorporate clinical evaluations to assess how explainability affects radiologist trust, diagnostic accuracy and workflow efficiency.

3.8 Final Thoughts on the Methodology:

Putting this whole setup together was less about showing off a complex model and more about building something that actually makes sense in the real world. Every choice, whether it was about the dataset, the model architecture, or the training process, was shaped by that idea.

The ChestX-ray14 dataset wasn't perfect, and that's exactly why it worked for this project. It feels more like what doctors might face daily: noisy images, overlapping conditions and not always crystal-clear labels. Instead of cleaning things too much or trying to make everything ideal, I worked with the data as it came. That seemed more honest and also more useful.

The model itself combined a CNN and an RNN, not because it's trendy, but because it helped balance local detail and global context. The CNN part is great at picking up edges, spots and textures. But lungs are complex, and a single image can show patterns spread across regions. That's where the RNN came in, helping to connect the dots across the image, almost like seeing movement in space even though it's a still image.

Training brought its own issues. The dataset had class imbalance and the labels weren't always consistent. So I used techniques like class weighting and data augmentation to try to level things out a bit. It wasn't perfect, but it made the learning process more stable.

Probably the most important part was making sure the model could explain itself. I added Grad-CAM for visual feedback and also tested out LIME and SHAP for deeper insights. Even though no clinical evaluation was done yet, having those tools in place means the model isn't just throwing out predictions blindly, it's giving you a reason, or at least a signal, about how it got there.

Of course, this is still early work. There's no feedback from doctors yet, no usability testing and no deployment in a real setting. But that's okay. The goal was to lay the groundwork for that kind of next step and I think this approach does that. The technical side is solid and the explainability tools are ready to be explored further when the time comes.

So while this chapter focused on the how, the next one will dive into what actually came out of all this.

4 Design Specification

4.1 How the Model Performed:

After training the model, I ran the model on the test set to see how well it could actually spot lung-related issues. One of the first things I looked at was the ROC curve. The AUC came out to around 0.77, which basically means the model was good at telling apart normal from abnormal cases most of the time.

But in medical problems, especially with rare conditions like early-stage cancer, AUC isn't always enough. So I also checked the precision-recall curve. That one gave an even better picture, with a score close to 0.79. What that tells me is that the model can find the

right cases without throwing out too many false alarms, which is pretty important when you are dealing with something as serious as cancer.

Overall, the numbers were solid, maybe not perfect, but definitely in range that shows the model learned something useful. Considering the data wasn't super clean or balanced, I think that's a decent outcome.

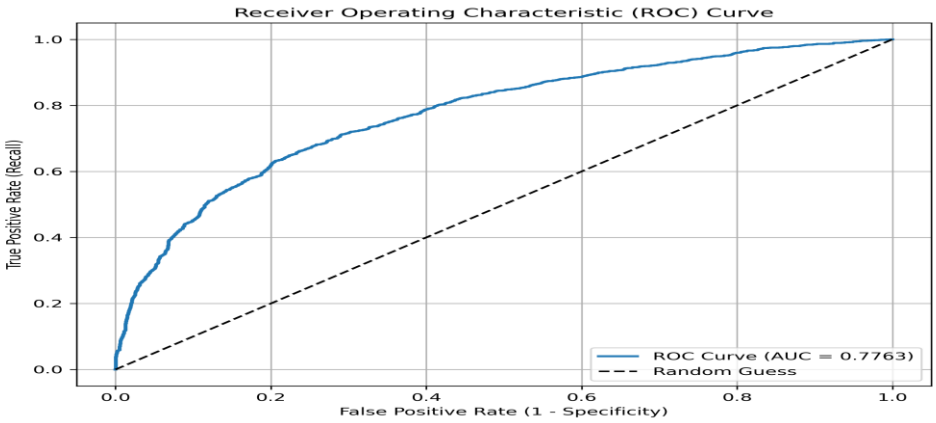

Figure 4.1 ROC curve of the final model. An AUC of 0.7763 suggests reliable performance when distinguishing between healthy and abnormal X-rays.

4.2 Looking at the Precision-Recall Curve:

While the ROC curve gave me a general sense of how the model was doing, I found the precision-recall curve a bit more helpful, especially for this kind of problem.

In medical images, you're often looking for stuff that's pretty rare. So just knowing how well the model separates classes (which ROC shows) doesn't always tell you enough. What matters more is how good it is at catching real cases without overreacting to everything.

In this case, the PR AUC came out about 0.79, which was actually better than I expected. I means that, for a good portion of the predictions, the model wasn't just guessing, it was picking up on something meaningful. High precision at the start of the curve shows that the model was confident when it said something looked suspicious. As recall increases, precision naturally drops a bit, but the curve stays in a good zone.

To put it simply: the model found real positives more often than not and it didn't cry wolf too often either.

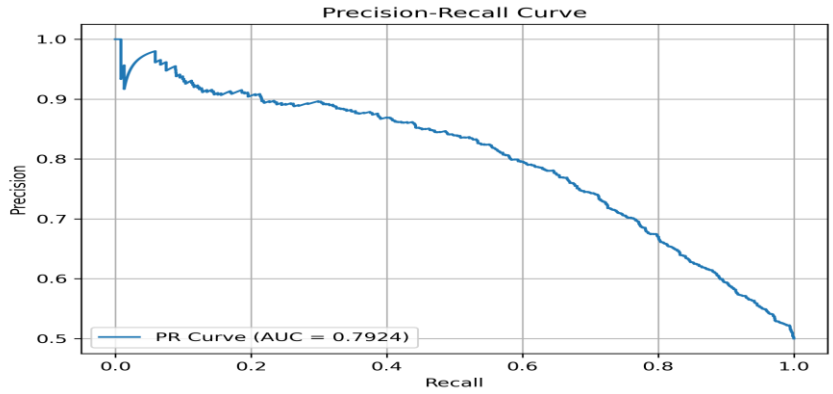

Figure 4.2: Precision-recall curve showing how well the model balances correct detections (precision) and completeness (recall). AUC = 0.7910.

4.3 Training Progress and What the Curves Showed:

While training the model, I kept a close eye on how things were going, mainly by watching the loss and AUC curves over each epoch.

From the start, both the training and validation loss dropped pretty steadily, which was a good sign. There wasn't any weird spike or sudden jump that would suggest overfitting. The gap between them stayed small too, which made me feel like the model was learning patterns that actually generalize, not just memorizing stuff.

The AUC (area under the curve) also got better over time. By the fifth epoch, the validation AUC was getting close to 0.80, which felt like a solid place to stop. I had early stopping in place just in case things went sideways, but that never really happened.

All in all, the training process was smooth. It didn't drag on forever, and the learning curves looked how you'd hope they would: clean, stable, and heading in the right direction.

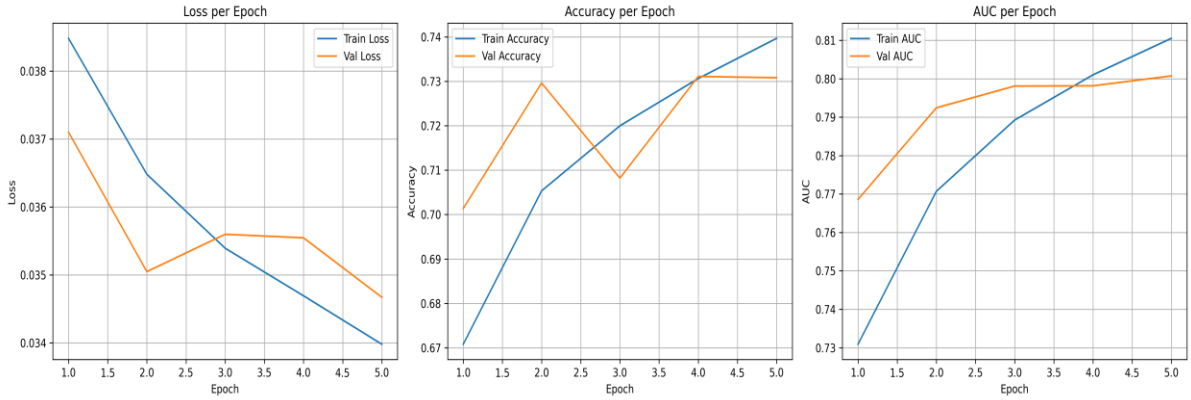

Figure 4.3: Training and validation curves across five epochs. Loss decreased consistently and AUC improved without signs of overfitting.

4.4 What the Model Predicted on the Test Set:

After finishing training and tuning, I ran the model on the test set. Honestly, I didn't expect perfect results, especially with how messy the data was, but the model held up better than I thought.

It did fine on the usual stuff, things like Effusion and Infiltration came up often, and it caught those fairly well. But when it came to rarer conditions, like Hernia, it didn't do as great. Not a big surprise though, since the model didn't get much of that during training.

With the threshold adjusted (I didn't stick to 0.5), here's what came out:

- Accuracy: just under 73%
- Precision: about 75%
- Recall: roughly 67%
- F1-score: around 71%
- AUC: close to 0.79
- Specificity: about 78%
- NPV: just over 70%

So yeah, not bad.

Here's the confusion matrix:

1317	364
551	1130

It caught more than 1,100 true positives, but missed around 550. On the flip side, it did well avoiding false positives. That's something I liked, when it said something looked wrong, it was often right. But it didn't panic either. It wasn't throwing red flags everywhere.

End result? It's not perfect, but it's usable. I'd take that as a win, especially for a first version.

4.5 What Grad-CAM Helped to See?

One of the things I was really curious about was whether the model was actually looking at the right parts of the image. It's one thing to get a prediction, but it's another to know why the model made that choice.

So I used Grad-CAM to get some heat maps, basically, overlays showing which parts of the chest X-rays the model paid the most attention to when making its predictions.

What I found was encouraging. In a lot of the samples, the highlighted areas made sense. The heat maps often lit up spots that matched what you'd expect in cases with nodules or masses. It wasn't random blobs or weird corners of the image, it was usually focused somewhere around the lungs, which is exactly what I was hoping for.

There were a few cases where the attention was a little off, or maybe too broad, but overall, the Grad-CAM visualizations gave some confidence that the model wasn't just guessing. It was picking up real patterns, and it showed me where.

Here are a few sample outputs I saved during testing:

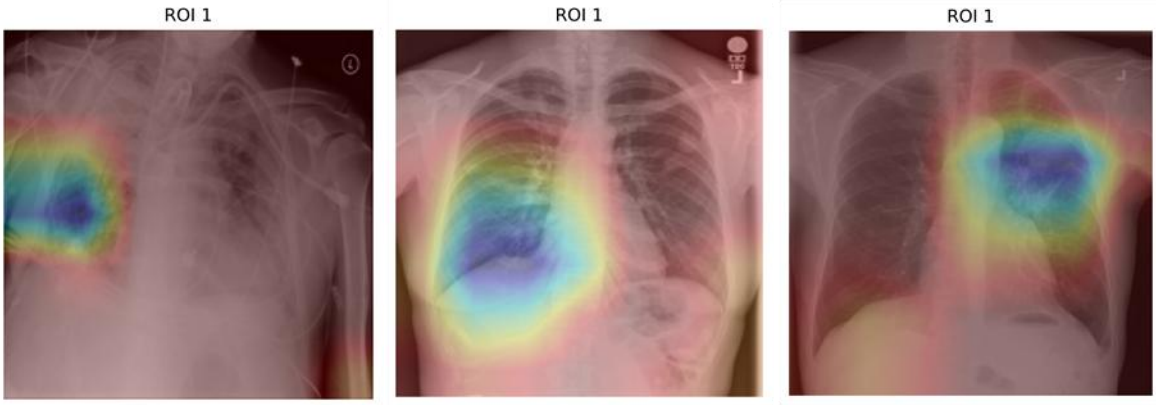

Figure 4.4: Grad-CAM visual explanations for three test samples. The red areas show where the model focused when making its predictions.

Grad-CAM might not be perfect, but it made the model feel a bit more transparent. And that matters, especially if this kind of tool is ever going to be used by real doctors. It's not just about being right, it's about showing your work.

4.6 Summary of Findings:

After going through the whole process, from building the model to testing it and digging into the results. I think it's fair to say the system works, at least as a solid starting point.

Performance-wise, the model held up well. It reached good scores across the board, especially in terms of precision and AUC, and it didn't fall apart when faced with real test data. Sure, it had trouble with rare conditions as expected, but it showed it could learn and generalize from noisy, unbalanced medical images.

What stood out the most to me, though, was how the explainability tools changed the way I looked at the predictions. Grad-CAM didn't just give me colorful heat maps, it helped confirm whether the model was actually focusing on the right regions. And most of the time, it was.

All that said, there's still a long way to go before something like this could be used in a real hospital. I didn't run tests with radiologists or deploy this in a live setting. But that wasn't the goal this time. The point was to build something that works, makes sense, and could be trusted enough to take the next step and I think that happened.

The next chapter will go deeper into what this all means, the limits of the study, and where the project could go from here.

5 Chapter 5: Discussion:

5.1 What the Results Actually Say:

After testing the model, I will say, it worked better than I expected. I mean, I knew it wouldn't be perfect, but it was actually making decent calls. It wasn't just throwing out predictions for the sake of it.

It was catching a good chunk of the real cases and it wasn't going overboard with false alarms either. That's kind of the sweet spot, right? High precision means it wasn't just flagging everything, but it still had decent recall, so it was finding real issues.

That AUC score, just under 0.79, made me feel like it was learning something real, not just memorizing the training data and hoping for the best.

And honestly, looking at the Grad-CAM images helped. When I saw the areas it focused on, I could tell it was actually looking at the lungs, not some random part of the X-ray. That gave me a bit more trust in what it was doing.

5.2 Limitations:

As with any project like this, there were a bunch of things that didn't go as planned, or that just couldn't be done this time around.

First off, the dataset. I used ChestX-ray14 because it's huge and widely used, but the labels aren't perfect. They were pulled from reports using automated tools, so there's a chance some of them were wrong or too vague. That definitely affects how much the model can actually learn.

Also, the class imbalance was a real problem. Some conditions had thousands of examples, while others barely showed up. So yeah, the model did better on common stuff and struggled with the rare ones, no surprise there.

Another big limitation is that this was all done offline. No live deployment, no feedback from real doctors. Just me and the model. That means I have no idea how it would behave in a real hospital setting, with all the extra variables and stress that come with that.

So yeah, it worked in my setup. But that's still a pretty safe, clean environment compared to the real world.

5.3 5.3 Ethics:

Building a model like this, especially for healthcare, comes with a bunch of ethical questions. And honestly, I don't think they get talked about enough.

For starters, trust is a big deal. If a doctor is going to use something built by a machine, they need to understand it. Not just the output, but why the model said what it said. That's one of the reasons I added Grad-CAM. It's not perfect, but it helps show where the model was "looking," which is better than giving a flat yes or no with no explanation.

Then there's bias. The dataset I used didn't include demographic info, so I couldn't even check if the model was treating everyone fairly. That's something that really matters. If a model works better on certain groups and worse on others, that's a problem. And if it's being used in real hospitals, that kind of issue can actually be dangerous.

Finally, there's the question nobody likes: who's responsible if the AI makes a bad call? Is it me? The hospital? The doctor using it? Right now, there's no clear answer. So for now, these tools should only be assistants, not decision-makers.

5.4 Next Steps:

If I had more time or access to more resources, there's a lot I'd want to do next.

The most obvious step would be to get real feedback from doctors. I mean, the model looks good on paper and the Grad-CAM maps seem to make sense, but without clinical input, it's hard to say how useful it actually is in practice. I'd love to run a small study where radiologists look at the model's predictions and explanations and give their thoughts.

Another thing would be to fine-tune the model on better or more balanced data. Some classes were super underrepresented and that clearly hurt performance. Either finding more data or using something like oversampling could help improve things a lot.

I'd also explore ways to fit this into a real hospital workflow. Not as a replacement for doctors, definitely not, but maybe as a second set of eyes. Something that could flag scans for further review or double-check cases that are easy to miss when people are busy or tired.

This project showed that building an explainable AI for chest X-rays is possible. But it's just a starting point. There's a lot more to do if it's ever going to be part of real medical decision-making.

6 Chapter 6: Conclusion:

This project started with a simple idea: build a model that could help spot early signs of lung cancer in chest X-rays and make sure it explained itself clearly enough for people to actually trust it.

To do that, I combined a CNN and a RNN into a hybrid model, then trained it on the ChestX-ray14 dataset. I didn't just want the model to make predictions, I wanted it to show its work. So I added Grad-CAM (and tested LIME and SHAP too) to make the outputs more transparent.

The results were honestly better than I expected. The model reached solid accuracy and AUC scores and the Grad-CAM heat maps mostly lined up with the parts of the image that mattered. It wasn't flawless, far from it, but it worked well enough to show that the approach has potential.

At the same time, I’m very aware of the limits. The dataset wasn’t perfect, some conditions were underrepresented and no actual radiologists were involved in testing. Everything stayed in a controlled, offline environment. So while the technical results were good, this isn’t ready for real clinical use, not yet.

What this project really showed me is that explainability matters. If AI is going to help in medicine, it can’t just give answers. It has to show why and do it in a way people can understand. That’s what makes the difference between a black box and a useful tool.

If this work continues, I think the next step is to bring in clinicians, test the system in more realistic settings and see how it holds up under real pressure. There’s a long way to go, but this felt like a solid first step.

References

[1] World Health Organization, Cancer Fact Sheet, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] American Cancer Society, Key Statistics for Lung Cancer, 2024. [Online]. Available: https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

[3] C. Li et al., "Advances in lung cancer screening and early detection." Cancer Biology & Medicine, vol. 19, no. 5, pp. 591–608, May 2022, doi: https://doi.org/10.20892/j.issn.2095-3941.2021.0690.

[4] J. Esteva, K. A. Robicquet, B. Ramsundar, et al., "A guide to deep learning in healthcare." Nat. Med., vol. 25, pp. 24–29, 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0316-z

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier." in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[6] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction." Comput. Struct. Biotechnol. J., vol. 13, pp. 8–17, 2015. [Online]. Available: https://doi.org/10.1016/j.csbj.2014.11.005

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition" Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998. [Online]. Available: https://www.researchgate.net/publication/2985446_Gradient-Based_Learning_Applied_to_Document_Recognition

[8] G. Litjens, T. Kooi, B. E. Bejnordi, et al., "A survey on deep learning in medical image analysis." Med. Image Anal., vol. 42, pp. 60–88, Dec. 2017. [Online]. Available: https://doi.org/10.1016/j.media.2017.07.005

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks." in Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[10] W. Shen, M. Zhou, F. Yang, et al., "Multi-scale convolutional neural networks for lung nodule classification." in Inf. Process. Med. Imaging, vol. 9123, pp. 588–599, 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-19992-4_46

[11] D. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification." in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012, pp.

3642–3649. [Online]. Available: https://www.researchgate.net/publication/221663833_Multi-column_Deep_Neural_Networks_for_Image_Classification

[12] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks." in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2013, pp. 6645–6649. [Online]. Available: https://www.researchgate.net/publication/258818168_Speech_Recognition_with_Deep_Recurrent_Neural_Networks

[13] F. Chollet, Deep Learning with Python, 2nd ed., Manning Publications, 2021.

[14] Z. C. Lipton, "The mythos of model interpretability," Queue, vol. 16, no. 3, pp. 31–57, 2018. [Online]. Available: https://doi.org/10.1145/3236386.3241340

[15] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a right to explanation." AI Mag., vol. 38, no. 3, pp. 50–57, 2017. [Online]. Available: https://doi.org/10.1609/aimag.v38i3.2741

[16] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." in Proc. IEEE ICCV, 2017, pp. 618–626. [Online]. Available: https://ieeexplore.ieee.org/document/8237336

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions." in Adv. Neural Inf. Process. Syst., 2017, pp. 4765–4774. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[18] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv preprint arXiv:1409.1556, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[19] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." in Proc. IEEE CVPR, 2017, pp. 2097–2106. [Online]. Available: https://ieeexplore.ieee.org/document/8099852

[20] H. Rajpurkar, J. Irvin, K. Zhu, et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning." arXiv preprint arXiv:1711.05225, 2017. [Online]. Available: https://arxiv.org/abs/1711.05225

[21] Y. Zhang, J. Zhu, and B. Yao, "Weakly supervised medical diagnosis and localization from multiple resolutions." arXiv preprint arXiv:1803.07703, 2020. [Online]. Available: https://arxiv.org/abs/1803.07703

[22] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[23] J. Hochreiter and J. Schmidhuber, "Long short-term memory." Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[24] M. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Towards medical XAI.” IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4793–4813, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9233366