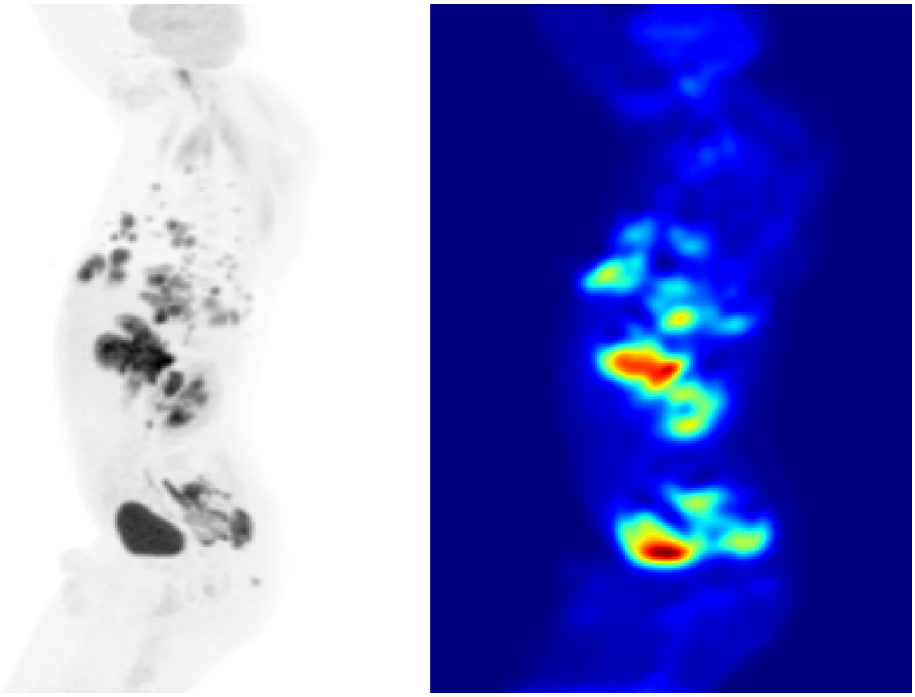
CHALMERS
UNIVERSITY OF TECHNOLOGY



Explainable AI in Healthcare

Physicians Perspectives and Technical Evaluation of AI-Based Decision Support and Explainability Methods

Master's thesis in Biomedical Engineering

MOA TOMASSON & SAGA WESTERKULL

Department of Electrical Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026

Master's thesis 2026

Explainable AI in Healthcare

Physicians Perspectives and Technical Evaluation of
AI-Based Decision Support and Explainability Methods

MOA TOMASSON & SAGA WESTERKULL

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2026

Explainable AI in Healthcare
Physicians Perspectives and Technical Evaluation of
AI-Based Decision Support and Explainability Methods
MOA TOMASSON & SAGA WESTERKULL

Supervisor: Victor Wåhlstrand, Department of Electrical Engineering
Examiner: Ida Häggström, Department of Electrical Engineering

Master's Thesis 2026
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: PET image and Grad-CAM attribution map generated from the CNN-based survival prediction model presented in this thesis.

Gothenburg, Sweden 2026

Abstract

Artificial intelligence (AI) is increasingly integrated into healthcare, particularly in clinical decision support and predictive modeling. However, the limited interpretability of many machine learning models remains a major challenge for clinical implementation, motivating growing interest in explainable artificial intelligence (XAI).

This thesis investigates XAI in healthcare from both technical and clinical perspectives. The clinical perspective was explored through qualitative interviews with physicians focusing on AI-based decision support systems and the role of explainable AI in clinical practice. The findings revealed a cautiously optimistic view of AI-supported decision-making, while emphasizing that clinically useful explanations should be concise, intuitive, and seamlessly integrated into existing workflows.

The technical part of the study investigated XAI methods for survival prediction in lymphoma patients using both tabular clinical data and medical imaging data. Multiple survival modelling approaches, including Cox regression, DeepSurv, and convolutional neural network models, were implemented and evaluated using several post-hoc explainability methods across the different data modalities. While both modalities demonstrated strong predictive performance, the tabular models achieved slightly stronger results with more stable, interpretable explanations. Furthermore, different XAI approaches highlighted complementary but inconsistent patterns, illustrating challenges related to the robustness and reliability of post-hoc explanations.

Overall, the findings demonstrated that successful clinical integration of AI depends as much on providing reliable, clinically meaningful explanations as it does on achieving strong predictive performance.

Keywords: Explainable AI, survival analysis, SHAP, Grad-CAM, Integrated Gradients, Occlusion Sensitivity, clinical decision support, deep learning, medical imaging.

Acknowledgements

We would like to express our sincere gratitude to our supervisor, Victor Wåhlstrand, for the valuable guidance, support, and feedback provided throughout this thesis project. We also wish to thank our examiner, Ida Häggström, for her flexibility, support, and encouragement throughout the project. Her willingness to help and accommodate challenges during the early stages of the thesis was greatly appreciated and meant a lot to us.

We are especially grateful to all physicians who participated in the interview study and generously shared their time, experiences, and perspectives. Their contributions were essential to this thesis.

In addition, we would like to thank Anna Bakidou for valuable guidance related to the interview process.

Finally, we would like to thank everyone who, in different ways, contributed to and supported this work throughout the project.

Moa Tomasson and Saga Westerkull, Gothenburg, 2026-05-25

Contents

List of Figures

1

Introduction

Recent advances in machine learning and deep learning have increased the use of artificial intelligence (AI) in healthcare, particularly in areas such as clinical decision support, risk prediction, and medical data analysis [1]. These machine learning approaches can identify complex patterns in large clinical datasets and have shown promising potential in both structured clinical data analysis and medical imaging applications [1], [2]. At the same time, traditional statistical approaches such as the Cox proportional hazards model remain widely used in clinical survival research due to their interpretability and established role in medical practice [3].

Despite their predictive capabilities, many advanced AI models remain difficult to interpret and are therefore often described as black boxes [4]. In clinical settings, this lack of transparency may limit trust and adoption, particularly when AI predictions influence high-stakes medical decisions.

To address this challenge, explainable artificial intelligence (XAI) has emerged as an important research field focused on improving the interpretability and transparency of machine learning models [4]. XAI methods provide insight into how features or image regions influence model predictions, which is particularly important in healthcare where clinicians must be able to assess whether AI-generated recommendations are clinically reasonable and reliable [5].

However, explainability in healthcare involves challenges beyond technical interpretation alone [4]. Even when explanations can be generated, it remains unclear how they should be presented and interpreted in clinical practice. Different explanation methods may provide different perspectives on model behaviour, and explanations that are technically meaningful may not necessarily be perceived as useful or trustworthy by clinicians. As a result, there is a growing need to investigate explainability from both technical and clinical perspectives.

This thesis combines technical analyses of survival prediction models with qualitative interviews exploring physicians' perspectives on AI-based clinical decision support systems and explainable AI. By integrating technical and clinical perspectives, the study addresses both the methodological challenges of explainability and the practical requirements for applying AI systems in healthcare.

1.1 Study Aims

The overall aim of this thesis is to investigate explainable artificial intelligence in healthcare from both technical and clinical perspectives. By combining technical evaluations of explainability methods with qualitative interviews exploring physicians' perspectives on explainable AI, the study examines how explainable AI can be interpreted, evaluated, and applied within clinical practice.

The aim of the interview study is to explore physicians' perceptions of AI-based decision support systems and explainable AI in healthcare. It further aims to investigate how physicians interpret different forms of AI explanations and how explainability may influence trust, usability, and the integration of AI systems into clinical practice.

The technical study aims to investigate explainable AI methods for survival prediction models based on both tabular clinical data and medical imaging data. A central objective is to evaluate the consistency, interpretability, and clinical relevance of different explainability approaches across both linear and non-linear models in a healthcare context.

2

Background

This chapter presents the theoretical background for the study, focusing on the use of artificial intelligence in healthcare and its role in clinical decision support. It introduces key concepts in survival analysis, machine learning, and explainable AI, which together form the basis for the modelling approaches and the analysis of how such systems are interpreted in clinical practice.

2.1 Artificial Intelligence in Healthcare

The use of AI in healthcare has evolved from early rule-based systems toward modern data-driven machine learning approaches. While earlier systems relied on manually defined clinical rules, advances in machine learning and the increasing availability of healthcare data have enabled AI models to automatically learn predictive patterns directly from large clinical datasets, medical images, and electronic health records (EHR) [6]. More recently, deep learning has further expanded the capabilities of AI by enabling automatic feature extraction from high-dimensional data, particularly in medical imaging applications [2].

Today, AI is playing an increasingly important role in clinical decision support systems, where machine learning models assist clinicians by analysing patient data and generating diagnostic and prognostic insights [1]. Particularly promising applications have emerged in areas such as medical imaging and risk prediction.

In medical imaging, deep learning models have demonstrated strong performance in tasks including image classification, lesion detection, and identification of pathological abnormalities [7]. These models can automatically learn complex feature representations directly from imaging data, making them especially well suited for radiological imaging applications.

Machine learning methods are also increasingly applied in clinical risk prediction, including survival analysis, mortality prediction, and prediction of disease progression [1]. By analysing structured clinical data, these models can identify patterns associated with patient outcomes and support clinical decision-making.

2.2 Machine Learning and Deep Learning

Machine learning (ML) and deep learning (DL) form the methodological foundation for the modelling approaches used in this study. These methods enable the analysis of both structured clinical data and medical imaging, supporting tasks such as risk prediction and outcome analysis.

2.2.1 Machine Learning in Healthcare

ML refers to a category of algorithms that improve their performance on a specific task by learning from data, rather than relying on explicitly programmed rules [8]. In healthcare, these methods are frequently applied to structured data such as electronic health records, laboratory measurements, and demographic variables to support tasks such as risk prediction and outcome modelling [1].

A common approach within ML is supervised learning, where models are trained on labelled data consisting of input features and corresponding outcomes [8]. The objective is to optimise a function that can predict outcomes for new, unseen data. In clinical applications, this often involves estimating patient risk or predicting time-to-event outcomes.

2.2.2 Deep Learning

Deep learning is a subset of machine learning based on artificial neural networks with multiple layers [2]. These models learn complex, non-linear relationships directly from data, enabling them to capture patterns that are difficult to model using traditional methods.

Model training is typically performed by optimising a loss function using gradient-based methods and backpropagation, where the model's weights are iteratively adjusted to minimise the discrepancy between predicted and observed outcomes [8].

A key advantage of deep learning is its ability to automatically learn feature representations from raw data, reducing the need for manual feature engineering [2]. This has contributed to the widespread use of deep learning in healthcare applications such as medical imaging and risk prediction.

2.2.3 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a class of deep learning models specifically designed for image and spatial data. They use convolutional layers to automatically extract hierarchical features from input images, allowing the model to detect patterns such as edges, shapes, and more complex structures [2].

CNNs utilise local receptive fields and weight sharing to preserve spatial relationships within the data while significantly reducing the number of parameters compared to fully connected networks [9]. This makes them highly effective for high-dimensional medical images.

CNNs are commonly applied to medical imaging tasks, including image classification, segmentation, and detection of pathological findings. Their ability to learn directly from raw image data makes them particularly suitable for imaging-based prediction tasks, where image-derived features are used to estimate patient risk [2].

2.2.4 Challenges with Healthcare Data

Healthcare data presents several challenges for machine learning models. Clinical data is often heterogeneous, originating from multiple sources such as EHR systems, imaging modalities, and laboratory measurements. In addition, missing data is common, which can affect model performance and reliability [1].

Bias and imbalance in healthcare datasets may also cause models to perform unevenly across different patient populations, raising concerns about fairness and generalisability [10]. Furthermore, healthcare data is subject to strict privacy regulations, which can limit data availability and sharing.

These challenges highlight the importance of careful data handling, model evaluation, and interpretation when developing AI systems for healthcare applications, particularly in high-stakes settings such as clinical decision support.

2.3 Survival Analysis in Healthcare

Survival analysis is a fundamental framework for modelling time-to-event data in healthcare, where both the occurrence and timing of clinical events are of interest. It provides statistical methods for handling censored observations and analysing time-dependent risk, and serves as the foundation for both traditional survival models and modern machine learning-based approaches.

2.3.1 Core Concepts in Survival Analysis

In medical research, survival analysis is used to model the time until a specific event occurs, such as death, disease recurrence, or disease progression [11]. Unlike traditional predictive models, it explicitly incorporates the temporal dimension, allowing both the likelihood and timing of events to be analysed.

Survival data often contains censored observations, meaning that the event of interest has not been observed for all individuals during the study period [12]. In such cases, each observation consists of a time variable T and an event indicator δ, where $\delta = 1$ denotes an observed event and $\delta = 0$ indicates a censored observation. Survival analysis methods incorporate this partial information, allowing censored observations to contribute to estimation without introducing bias.

Two fundamental quantities in survival analysis are the survival function and the hazard function. The survival function is defined as:

$$S(t) = P(T > t), \tag{2.1}$$

where T denotes the random variable representing survival time and t represents a specific point in time. The survival function represents the probability that the event has not occurred by time t [13].

The hazard function describes the instantaneous risk of experiencing the event at time t, given survival up to that time, and can be expressed as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}, \tag{2.2}$$

where Δt is a small time interval [3]. It should be interpreted as a rate rather than a probability, describing how risk evolves over time.

2.3.2 Traditional Survival Analysis Methods

Several statistical methods have been developed to analyse time-to-event data. A commonly used non-parametric approach is the Kaplan-Meier estimator, which provides an estimate of the survival function based on observed event times while accounting for censoring [13]. Kaplan-Meier curves are frequently used to visualise survival probabilities over time and to compare survival between patient groups.

For modelling the relationship between explanatory variables and event risk, the Cox proportional hazards (CoxPH) model is one of the most widely applied methods [3]. The CoxPH model expresses the hazard function as:

$$\lambda(t \mid x) = \lambda_0(t) \exp(\beta^T x) \tag{2.3}$$

In this formulation, $\lambda_0(t)$ represents the baseline hazard function, corresponding to the instantaneous risk of the event at time t when all covariates are zero. The term $\exp(\beta^T x)$ models the relative effect of the covariates on the hazard, where β is a vector of regression coefficients associated with the covariates x. The exponentiated regression coefficients correspond to hazard ratios, describing how changes in the covariates influence the relative event risk.

This formulation implies the proportional hazards assumption, meaning that the effect of each covariate on the hazard is multiplicative and remains constant over time. The CoxPH model plays a central role in modern survival modelling and forms the foundation for several machine learning-based survival approaches.

2.3.3 Machine Learning for Survival Analysis

With the increasing availability of large clinical datasets, machine learning methods have been developed to extend classical survival analysis models. Neural network-based approaches, such as DeepSurv [14], replace the linear predictor of the CoxPH model with a learned non-linear function while retaining the Cox partial likelihood as the training objective. Similar convolutional extensions have also been applied to medical imaging applications.

In these models, training is performed using the negative Cox partial likelihood loss, defined as:

$$\mathcal{L} = -\sum_{i:\delta_i=1} \left(h_\theta(x_i) - \log \sum_{j \in R_i} e^{h_\theta(x_j)} \right), \tag{2.4}$$

where $h_\theta(x_i)$ denotes the predicted risk score for patient i, δ_i is the event indicator, and R_i represents the risk set at time t_i, consisting of all individuals still under observation and at risk of experiencing the event at that time.

Instead of predicting absolute survival times, the network learns to assign relative risk scores by maximising the agreement between predicted risk rankings and observed event ordering across comparable patient pairs. This approach allows for flexible modelling of complex, high-dimensional data while handling censored observations and preserving the ranking-based objective of the CoxPH model [14].

2.3.4 Model Evaluation

Evaluation of survival models differs from standard predictive modelling due to censoring and the time-dependent nature of the data. A commonly used metric is the concordance index (C-index), which measures the model's ability to correctly rank individuals according to their risk [15].

The C-index represents the proportion of comparable patient pairs for which the predicted risk agrees with the observed ordering of event times. A higher C-index indicates better discriminative performance. A value of 0.5 corresponds to random prediction, whereas a value of 1.0 indicates perfect discrimination. Since the metric is based on relative ordering rather than absolute predictions, it is particularly suitable for comparing survival models that produce relative risk scores.

2.4 The Black-Box Problem in AI

The increasing use of complex machine learning models in healthcare has introduced significant challenges related to interpretability, transparency, and clinical trust [16]. Many deep learning models operate as "black boxes," meaning that the relationship between the input data and the resulting predictions is often difficult to interpret or explain [17].

These challenges largely arise from the high-dimensional and non-linear nature of modern AI systems. In deep learning models, predictions are generated through multiple layers of transformations, producing complex internal feature representations that are not directly interpretable by humans [4].

Unlike traditional statistical models, where model parameters have clear meanings, complex machine learning models often function as high-dimensional function approximators [4]. Although this flexibility enables strong predictive performance, it

also reduces transparency and makes it difficult to understand how individual input variables contribute to model predictions.

2.4.1 Implications for Healthcare

The lack of transparency in AI systems poses several challenges in clinical settings. Limited interpretability can make it difficult to identify potential biases, detect errors, and assess whether a model is relying on clinically relevant information rather than spurious correlations in the data [18]. These challenges directly affect how AI systems are perceived and used in clinical practice.

Since AI systems in healthcare are typically used as decision support tools rather than autonomous systems, clinicians must be able to interpret and evaluate model outputs to maintain accountability. A lack of transparency may reduce trust, limit usability, and hinder the adoption of AI in real-world healthcare settings [16].

Furthermore, regulatory requirements related to safety and data protection highlight the need for methods that improve interpretability and support the safe integration of AI systems into clinical workflows.

2.5 Explainable AI Methods

To improve transparency and support the clinical interpretation of complex machine learning models, explainable AI methods have emerged to provide insight into otherwise non-interpretable model behaviour [4]. These methods aim to translate raw computational outputs into clinically meaningful insights, allowing healthcare professionals to critically evaluate AI-driven recommendations.

XAI approaches can be broadly divided into intrinsically interpretable models and post-hoc explanation methods applied after training [4]. In intrinsically interpretable approaches, such as the CoxPH model, the relationship between input variables and predicted risk can be interpreted directly through the model parameters. For example, hazard ratios quantify how changes in input variables influence the predicted risk [3]. In contrast, post-hoc methods generate explanations for model predictions without modifying the underlying architecture or training process.

XAI methods can also be divided based on the level of interpretation they provide. Global explanations aim to describe overall model behaviour and identify features that are generally important across the dataset, while local explanations focus on explaining individual predictions for specific samples or patients [4].

In this thesis, post-hoc explanation methods are used to interpret predictions from both tabular and imaging-based models. For tabular data, feature attribution methods are used to quantify how individual variables contribute to model predictions, whereas for imaging data, attribution methods are used to identify and visualise the regions within the input images that most strongly influence the model's predictions.

2.5.1 SHAP

SHapley Additive exPlanations (SHAP) is a widely used post-hoc method for interpreting machine learning models [19]. The method is based on Shapley values from cooperative game theory, where each feature is treated as a contributor to the model prediction.

SHAP represents a prediction as an additive combination of feature contributions:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{2.5}$$

Here, x denotes the input feature vector containing M features, and ϕ_0 represents the base value corresponding to the expected model output over the training dataset. The term ϕ_i denotes the Shapley value for feature i, quantifying its additive contribution to the difference between the prediction $f(x)$ and the base value.

This framework is based on theoretical properties such as consistency and missingness, ensuring that features with no impact receive a Shapley value of zero and that feature contributions are distributed consistently across the input variables.

A key advantage of SHAP is that it provides both local and global interpretability [19]. Individual predictions can be explained at the feature level, while aggregation across samples reveals overall feature importance. This makes SHAP particularly useful in healthcare, where both patient-specific explanations and general insights are needed.

SHAP is well-suited for tabular data, where feature-level contributions are directly interpretable [19]. However, the method can be computationally expensive for complex models, and interpreting the resulting explanations may still require domain expertise.

2.5.2 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a widely used explainability method for interpreting convolutional neural networks, particularly in medical imaging applications [20]. The method generates visual explanations by highlighting the regions of an input image that contribute most strongly to the model's prediction, thereby providing insight into the network's decision-making process.

Grad-CAM computes gradients of the model output with respect to feature maps in a convolutional layer. The resulting heatmap can be expressed as

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right), \tag{2.6}$$

where A^k represents feature map k and α^c_k denotes the corresponding importance weight derived from the gradients of the model output.

A key advantage of Grad-CAM is that it produces spatially interpretable heatmaps that can be overlaid on the original image, making it particularly suitable for clinical applications where visual interpretation is important [20].

However, Grad-CAM produces relatively coarse heatmaps rather than precise pixel-level explanations, since the method relies on convolutional feature maps with reduced spatial resolution [20]. In addition, the resulting explanations may vary depending on which convolutional layer is used for attribution. Despite these limitations, Grad-CAM remains an established method for interpreting deep learning models in medical imaging.

2.5.3 Integrated Gradients

Integrated Gradients is a gradient-based attribution method designed for interpreting predictions of differentiable models, particularly neural networks [21]. The method attributes a model's prediction to its input features by integrating gradients along a path from a baseline input to the actual input.

Formally, the attribution for feature i is computed as:

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{2.7}$$

In this formulation, x represents the input sample and x' denotes a baseline input representing a neutral or uninformative state.

Compared to simple gradient-based methods, Integrated Gradients provides more stable and theoretically grounded attributions and satisfies desirable properties such as sensitivity and implementation invariance [21].

Integrated Gradients can be applied to both imaging and tabular data in neural network-based models, making it suitable for comparing feature attributions across different data modalities.

2.5.4 Occlusion

Occlusion is a perturbation-based attribution method that evaluates the importance of different regions of an input by systematically masking parts of the data and observing the effect on the model's prediction [22].

In imaging applications, this is typically done by sliding a window across the image and replacing each region with a baseline value. The importance of a region can be approximated as:

$$\Delta f = f(x) - f(x'), \tag{2.8}$$

where $f(x)$ is the original model prediction and $f(x')$ is the prediction after occluding part of the input.

Occlusion does not rely on model internals and can therefore be applied to any model, making it a model-agnostic approach to explainability.

A key advantage of occlusion is that it provides intuitive interpretations of feature importance, as it measures the actual impact of removing information from the input [22]. However, the method can be computationally expensive due to the large number of forward passes required, and the resolution of the explanation depends on the size of the occlusion window.

2.6 Human-AI Interaction in Healthcare

In healthcare, AI systems are typically used as decision support tools rather than autonomous systems [23]. This means that model predictions are interpreted and evaluated by clinicians, who remain responsible for the final clinical decisions. As a result, the interaction between human expertise and AI outputs becomes a central aspect of how these systems are used in practice.

In this context, the usefulness of AI models depends not only on predictive performance, but also on how their outputs are presented and understood. Clinicians must be able to interpret model predictions, assess their reliability, and integrate them with clinical knowledge and experience [18]. This places increased importance on transparency and interpretability in AI systems.

Explainability plays a key role in this interaction. The way model explanations are presented can influence clinicians' trust in the system, their ability to use it effectively, and their willingness to adopt it in clinical practice [4]. Poorly understood or misleading explanations may reduce trust or lead to incorrect conclusions, while clear and meaningful explanations can support informed decision-making.

Understanding how clinicians perceive and interpret AI-based decision support systems is therefore essential for the development of usable and trustworthy AI tools in healthcare. This highlights that successful clinical integration of AI systems depends both on predictive performance and on how clinicians interpret and interact with the generated outputs.

To address both the technical and human aspects of explainable AI in healthcare, this thesis combines technical evaluations of explainability methods with a qualitative interview study exploring physicians' perspectives on AI, explainability, and clinical decision-making.

Part I

Qualitative Interview Study with Physicians

3

Methodology

This qualitative study explored physicians' perceptions of Artificial Intelligence and Explainable AI in healthcare. The study focused on how physicians understand and interpret explainable AI and how AI-based systems might integrate into existing decision-making workflows.

3.1 Study Design

The study was conducted using semi-structured interviews with physicians. An interview guide was used to ensure that the same main topics were covered across all interviews while still allowing flexibility for follow-up questions and more in-depth discussion when relevant.

The interviews focused on clinical reasoning in uncertain situations, experiences with digital decision support tools, and general perceptions of AI in healthcare. The study also explored how physicians interpret different forms of AI explanations and how these explanations may influence trust and usability in clinical settings. In addition, it investigated how physicians perceived the relationship between explainability and the accuracy of AI-based systems.

3.2 Interview Guide

An interview guide was developed containing a set of main questions covering the key topics of the study, supported by potential follow-up questions. Transitional explanations were also included to introduce new topics and provide context throughout the interviews. The full interview guide is provided in Appendix A, together with an English translation in Appendix B.

The guide was developed iteratively and revised several times to improve its clarity and structure. Before the main data collection, a pilot interview was conducted to evaluate the clarity, flow, and duration of the interview. Based on feedback from the pilot interview, some questions were clarified and the guide was slightly shortened. The pilot participant was not included in the main data collection.

3.3 Participants and Recruitment

Participants were recruited through professional contacts and existing collaborations. Some physicians were contacted through the professional social media platform LinkedIn, while others were recruited through research collaborations at Sahlgrenska Academy.

To participate in the study, individuals were required to be licensed physicians currently working in clinical practice. These criteria ensured that all participants had recent experience with clinical decision-making in real healthcare settings.

In total, seven physicians participated in the study. The participants represented a range of medical specialties and levels of professional experience, including radiology, oncology, and vascular surgery, as well as early-career physicians in the form of medical interns. This diversity enabled the study to capture perspectives from different clinical contexts and professional backgrounds within healthcare. Although the sample size was limited, seven interviews are consistent with the scope of a qualitative thematic study, where the aim is to generate an in-depth understanding of participants' perspectives rather than statistical generalisability. An overview of the participants, including career stage, specialty, and interview format, is provided in Table 3.1.

Table 3.1: Overview of study participants.

Participant	**Gender**	**Career Level**	**Specialty**	**Interview Format**
P1	Male	Senior Physician	Oncology	In-person
P2	Male	Senior Physician	Vascular Surgery	In-person
P3	Female	Senior Physician	Radiology	In-person
P4	Male	Senior Physician	Radiology	Online
P5	Male	Medical Intern	—	In-person
P6	Male	Medical Intern	—	Online
P7	Male	Medical Intern	—	Online

3.4 Interview Procedure

Before the interviews, participants received a study description, a consent form, and a set of reflection questions intended to familiarise them with the interview topics. These reflection questions addressed topics such as their approach to decision-making in complex clinical cases, their use of digital tools and decision support systems in clinical work, and their previous experiences with AI-based systems.

Out of the seven conducted interviews, tree interviews were carried out online via Zoom, and four interviews were conducted in person. To ensure consistency across interviews, one researcher led all interviews and a second researcher took notes during the sessions. Each interview followed the structure of the interview guide

and lasted approximately one hour. With participants' consent, all interviews were audio-recorded.

During the section concerning explainable AI, participants were shown printed visual materials, including feature importance diagrams and heatmaps applied to medical images. These materials were used to facilitate discussion and gather feedback regarding different explanation formats. All visual materials are included in the interview guide presented in Appendix A.

3.5 Data Documentation and Analysis

During the interviews, notes were taken to document important observations and key discussion points. The audio recordings were transcribed and combined with the interview notes to form the primary dataset for analysis.

The interview data were analysed using a thematic analysis approach. The analysis was guided by the aim of the study and the main themes covered in the interview guide. Both researchers independently reviewed the interview transcripts to identify recurring ideas, patterns, and themes in the material. Attention was given not only to similarities but also to differences in experiences and opinions.

After the independent review process, the researchers compared and discussed their findings to reach agreement on the main themes and interpretations. This process was used to improve consistency and strengthen the credibility of the analysis. The interviews and analysis were conducted in Swedish, with results translated into English for this thesis.

3.6 Ethical Considerations

Participation in the study was voluntary, and participants were informed that they could withdraw at any time and were free to decline any individual question during the interview. Informed consent was obtained from all participants before the interviews were conducted, and audio recording only started once this consent had been confirmed. The recordings were used solely for transcription and analysis within the scope of this thesis project and were not shared beyond the research team. To protect participants' privacy and confidentiality, results are presented at group level, with no information reported in a way that could identify individual participants.

4

Findings

The thematic analysis of the interview data resulted in eight themes, presented in the sections below. The themes are organised to reflect the progression of the interviews: beginning with general clinical decision-making and the use of digital tools, moving through questions of trust and responsibility, and concluding with more specific discussions of explainability formats and the trade-off between accuracy and transparency.

4.1 Clinical Decision-Making and Collaboration

The interviews revealed that physicians perceive clinical decision-making as a complex process with multiple possible approaches rather than a single correct answer. Decision-making was described as particularly challenging in cases that fall outside standard guidelines, such as patients with multiple disease recurrences, significant comorbidities, or conflicting test results. Senior physicians emphasised the difficulty of balancing technical possibilities with the patient's overall health and long-term outcomes, while junior physicians found diffuse findings and non-standardised cases especially difficult to navigate. This complexity was captured by one physician, who described it as follows:

> *"You often weigh a large number of different factors at the same time. It is not yes or no, right or wrong, but rather many simultaneous considerations: comorbidities, technical factors, diagnostic factors. It is the complexity that is the most challenging."*

In navigating this complexity, collaboration emerged as a central part of the decision-making process. Physicians frequently relied on discussions with colleagues to validate and structure their reasoning. Junior physicians described these interactions as an important source of support and guidance, while senior specialists highlighted the value of multidisciplinary conferences for managing complex cases and combining expertise from different clinical fields.

4.2 Digital Decision Support and AI

Participants described using several digital decision support tools in daily clinical work, including medical databases, risk calculators, and radiological support sys-

tems. Although these systems were considered valuable for improving efficiency, participants also expressed frustration regarding poor usability and limited integration between different systems.

AI-based decision support systems were generally viewed with cautious optimism. Participants believed AI could improve efficiency, support more equal care, and assist with repetitive or data-intensive tasks such as screening, documentation, and triage. At the same time, concerns were raised regarding the implementation of new systems, increased administrative workload, and the limited clinical value of poorly integrated systems.

Several physicians also expressed scepticism toward the black box nature of AI systems and emphasised the importance of understanding how AI-generated recommendations are produced. This sentiment was captured by one physician, who when asked about the opportunities of AI in healthcare, responded:

> *"The possibilities are enormous, if those who develop these systems understand what clinicians actually need. Not a black box that simply gives a yes or a no."*

AI was primarily viewed as a supportive tool rather than a replacement for physicians. Participants consistently stated that the final responsibility for clinical decisions must remain with the physician, and it was within this context of professional accountability that questions of trust, risk, and the appropriate boundaries of AI involvement became particularly prominent.

4.3 Trust, Risk, and Clinical Responsibility

Participants identified several risks related to the integration of AI in healthcare. One major concern was that increased reliance on AI could weaken clinical competence. Senior physicians expressed concern that less experienced colleagues might rely too heavily on AI, while younger physicians acknowledged that it could reduce their own critical thinking. This concern was illustrated by one senior physician, reflecting on a scenario where AI replaces the junior primary review of radiological images:

> *"The only way to learn how to read radiological images is to read radiological images. If junior physicians never perform the primary review and are never required to make an assessment, how will they learn, and who will conduct the secondary review in the future?"*

Beyond the risk to clinical competence, concerns were also raised regarding data quality, representativeness, patient safety, and data privacy. Together, these risks underlined that the integration of AI into clinical practice requires careful consideration, not only of technical performance, but of broader consequences for the healthcare system.

For AI systems to be accepted in clinical practice, participants emphasised the importance of seamless integration into existing workflows. AI systems were expected

to simplify clinical work rather than create additional administrative burden, and poorly integrated systems were seen as unlikely to gain physician acceptance regardless of their technical capabilities.

Beyond integration, trust in AI was also shown to be highly contextual and dependent on the physician's own expertise. If an AI system produced a recommendation that differed from their own clinical judgment, participants stated that they would re-evaluate both the AI's reasoning and their own assessment, and in some cases discuss the situation with a colleague. Some junior participants further noted that their trust in AI would depend on their level of expertise within a specific clinical area. In unfamiliar situations they believed they would be more likely to follow an AI recommendation, whereas in familiar scenarios they would place greater trust in their own clinical judgment.

4.4 The Importance of Explainability

The need for explainability was strongly emphasised across all interviews. Participants stated that AI systems should not only provide recommendations, but also explain how those recommendations were generated. Explainability was closely linked to trust, credibility, and the physicians' ability to take responsibility for the final clinical decision.

Many participants compared AI explanations to discussions with colleagues, arguing that clinical recommendations should always be motivated and possible to question. However, physicians emphasised that they did not require detailed technical descriptions of the underlying algorithms. Instead, they wanted clinically understandable explanations showing which factors influenced the recommendation. Usability was considered critical. Explanations were expected to be concise, clear, and adapted to the time constraints of clinical work.

The importance of explanations was considered greatest in complex, uncertain, or high-stakes situations. In simpler cases, or when the AI recommendation aligned with clinical expectations, explanations were seen as less essential.

At the same time, participants highlighted risks associated with persuasive explanations. Several physicians expressed concern that convincing explanations could create a false sense of security and lead clinicians to trust faulty recommendations. This risk was expressed by one physician:

> *"A convincing explanation attached to an incorrect recommendation is a risk, and unfortunately, AI is sometimes wrong."*

4.5 Explainability for Structured Clinical Data

When evaluating explainability methods for clinical data, physicians emphasised the importance of speed and simplicity. Visual explanations such as feature importance diagrams were generally appreciated, provided they were intuitive and did not

contain excessive information.

Participants preferred explanations showing relative influence rather than exact percentages or highly precise numerical values, which were often viewed as clinically irrelevant. A recurring theme was the importance of showing both factors supporting and opposing a recommendation. Physicians also emphasised the value of including the patient's actual clinical values within the explanation to help assess whether the AI reasoning appeared reasonable. Several participants further highlighted that information about the certainty of the decision, such as probability estimates, could help evaluate the reliability of recommendations.

4.6 Visual Explanations for Medical Images

Heatmaps and other image-based explanation methods were generally well received. However, participants emphasised that the original medical image must remain visible, preferably with the explanation overlaid transparently. This makes it easier for the physicians to pinpoint which anatomical structures the AI system had highlighted.

Although visual markers were considered useful for directing attention, physicians noted that heatmaps mainly show where the model focuses rather than how it reasons. Many participants also stated that distinguishing between image regions that had a positive versus a negative impact on the prediction was unnecessary. Instead, they preferred simple visual guidance, such as heatmaps, bounding boxes, or threshold-based markings, that could highlight potentially abnormal areas while allowing physicians to retain responsibility for the final interpretation.

4.7 Global Explanations and Transparency

Participants considered global explanations and information about training data particularly important during the implementation of new AI systems. Understanding the model's general logic and development process was viewed as important for building initial trust.

Transparency regarding the training population emerged as a central theme. Physicians emphasised the importance of knowing the demographics, patient characteristics, and clinical context of the data used to train the model. Several participants compared this information to evaluating evidence in clinical studies.

Differences between experience levels were also observed. Senior physicians generally placed greater importance on understanding the background and validation of the model, whereas some junior physicians assumed that implemented systems had already undergone sufficient technical evaluation.

4.8 Accuracy and Explainability

When discussing the trade-off between accuracy and explainability, many participants preferred a slightly less accurate but explainable model over a highly accurate black box system. Physicians emphasised that transparent explanations make it easier to detect incorrect recommendations and assess whether the AI's reasoning is clinically sound.

Although participants acknowledged the importance of high predictive performance, most considered some degree of explainability necessary for clinical use. Systems that provided no insight into their reasoning were viewed as difficult to trust and challenging to integrate into clinical decision-making. Ultimately, the case for explainability was framed not only in technical terms, but as a matter of professional responsibility and patient rights, as captured by one physician:

> *"I would rather have a system that explains its reasoning than one that is merely accurate. The explanation provides a safety net, it gives you something to review and makes it easier to catch errors. No matter how accurate the system is, you can never trust it blindly, and someone must always be held responsible. Our patients have the right to a human being standing behind every decision."*

Part II

Technical Evaluation of Explainability Methods

5

Data and Preprocessing

5.1 Dataset Overview

The dataset used in this study consists of patients diagnosed with lymphoma and includes both clinical and imaging data. Each patient is represented at the individual level and is identified by a unique ID. The clinical data contain demographic variables such as gender, race, ethnicity, and smoking history, as well as diagnostic variables including histology codes, anatomical site codes, and disease stage. The imaging data consist of 2D PET scans. Additionally, survival-related variables, including days to diagnosis, days to death, and last appointment, were included to facilitate survival analysis.

Several filtering steps were applied before model development. First, only patients with valid IDs in the predefined dataset splits were retained. Second, a lymphoma-specific filter was applied, retaining only patients with ICD-O site codes corresponding to lymphoma diagnoses. To ensure consistency, only one record per patient was used. The earliest diagnosis was selected, and the first available PET scan was retained. Finally, the clinical and imaging datasets were merged, and only patients with complete data for both modalities were included in the final cohort.

5.2 Data Splitting

The final dataset was divided into training, validation, and test sets consisting of 2460, 806, and 836 patients, respectively. The training set was used for model optimisation, the validation set for model selection and early stopping, and the test set was reserved for final performance evaluation.

All preprocessing operations were fitted on the training data to prevent data leakage. The same transformations were then applied to the validation and test sets without any further fitting. This ensures that information from the validation and test data does not influence the training process.

5.3 Clinical Variables

The clinical data used in this study included the following covariates: gender, race, ethnicity, days to diagnosis, smoking history, ICD-O histology code, ICD-O site

code, and best stage. Gender describes the biological sex of the patient, while race and ethnicity provide demographic background information. Days to diagnosis represents the number of days from birth to the recorded diagnosis date. Smoking-related variables indicate the patient's history of smoking or tobacco use. The ICD-O histology code describes the microscopic classification of the tumour, whereas the ICD-O site code identifies the anatomical location of the cancer. Best stage represents the clinically assigned stage of disease progression at diagnosis.

Preprocessing of the clinical variables was performed using a `ColumnTransformer`. The variables were divided into categorical and numerical groups. The categorical variables included gender, race, ethnicity, smoking history, ICD-O histology code, ICD-O site code, and best stage. These variables were first converted to string format and then encoded using one-hot encoding. The numerical variable, days to diagnosis, was converted to numeric format, where invalid values were coerced to missing values (`NaN`). Standardisation was subsequently applied using z-score scaling, resulting in a variable with zero mean and unit variance.

5.4 Imaging Data

For each patient, a single PET image slice was selected for analysis. The images were used in their original form, and no intensity normalisation or data augmentation was applied. All images were kept in their original numeric scale.

The images had different spatial dimensions. To allow batch processing, the images were padded with zeros to match the maximum height and width within each batch. Padding was applied symmetrically around the image center to preserve spatial structure. As the model used in this study expects three-channel input, the single-channel PET images were duplicated across three channels.

5.5 Outcome Definition

The primary outcome for survival analysis was defined using a time-to-event variable and a binary event indicator. The time-to-event represents the duration from the initial diagnosis to either death or the final clinical contact. To ensure the survival timeline begins at the point of diagnosis, this duration was calculated by subtracting the time of diagnosis from both the time of death and the time of the last appointment.

The event indicator serves as a binary status flag for each patient. A value of 1 signifies that a death occurred during the study period, whereas a value of 0 indicates a censored observation, meaning the patient was still alive at the time of their last recorded follow-up.

6

Tabular Model

6.1 Experiments

Using the preprocessed tabular data, two survival modelling approaches were implemented: a Cox Proportional Hazards (CoxPH) model and a neural network-based model. The CoxPH model was used as an interpretable baseline, while the neural network was included to capture potential non-linear relationships between clinical features and survival risk.

6.1.1 Cox Proportional Hazards Model

The CoxPH model was implemented using the Lifelines library [24] and trained using the Cox partial likelihood loss defined in Equation 2.4. The input consisted of tabular clinical features (X), survival times (T), and event indicators (E). Features with no variance were removed prior to training to avoid numerical instability.

To reduce overfitting, L2 regularisation (penalisation) was applied. A hyperparameter search was conducted over penaliser values of 0.01, 0.1, 1, and 10, where the optimal value was selected based on validation C-index. The final model was subsequently retrained on the full training dataset using the selected parameter, producing relative risk scores based on the estimated partial hazard.

6.1.2 Neural Network Model

The neural network-based survival model was implemented following the DeepSurv framework. The model was implemented in PyTorch [25] as a fully connected feedforward network consisting of an input layer matching the number of tabular features, one hidden layer with 64 units and ReLU activation, and a final linear output layer producing a single risk score per patient. The network was trained using the negative Cox partial likelihood loss defined in Equation 2.4.

Training was performed using the Adam optimiser with full-batch optimisation, where the entire training dataset was used in each iteration. Model performance was evaluated at each epoch using both the training and validation sets. Early stopping was applied based on the validation C-index with a patience of 10 epochs, and the model parameters corresponding to the best validation performance were restored after training.

6.1.3 Model Evaluation

Both the CoxPH model and the neural network were evaluated using the concordance index. As the C-index depends only on relative ranking, the predicted risk scores from both models were used directly without calibration, enabling consistent comparison between the linear CoxPH model and the non-linear neural network.

During training, loss and C-index curves were monitored for both the training and validation sets to assess convergence and stability. Final model performance was evaluated on the held-out test set.

6.1.4 Explainability

Post-hoc explainability methods were applied to quantify feature contributions at both global and local levels. Feature importance derived from CoxPH model coefficient magnitudes was compared with attribution-based explanations obtained from SHAP (Equation 2.5) and Integrated Gradients (Equation 2.7), enabling comparison between traditional statistical interpretation and post-hoc explainability approaches across both linear and non-linear models.

Global Explainability: Global explainability analysis was performed to assess overall feature importance across the dataset. For the CoxPH model, global importance was derived from the magnitude of the model coefficients, which correspond to log hazard ratios. SHAP was applied to both the CoxPH model and the neural network, while Integrated Gradients was additionally used for the neural network. This enabled comparison of global attribution patterns across both linear and non-linear modelling approaches.

Local Explainability: Local explainability analysis was performed to interpret individual patient predictions. SHAP and Integrated Gradients were used to generate patient-specific feature attributions for the neural network, while SHAP was applied to the CoxPH model. The explanations were compared to assess consistency between explainability methods. For clarity, only features with an absolute SHAP value greater than 0.01 were included in the local explanation plots.

Feature Grouping and Clinical Mapping: To improve interpretability, raw model inputs were mapped to clinically meaningful categories. No feature selection was applied, and all available variables were retained during both model training and explainability analysis.

Many categorical variables were represented using one-hot encoding and subsequently grouped into higher-level clinical categories, including diagnosis, tumour site, stage, gender, race, ethnicity, and smoking status.

For the explainability analysis, attributions derived from SHAP and Integrated Gradients were aggregated at the level of these grouped categories. This reduced the dimensionality introduced by one-hot encoding and enabled interpretation at the level of clinically relevant concepts rather than individual encoded variables.

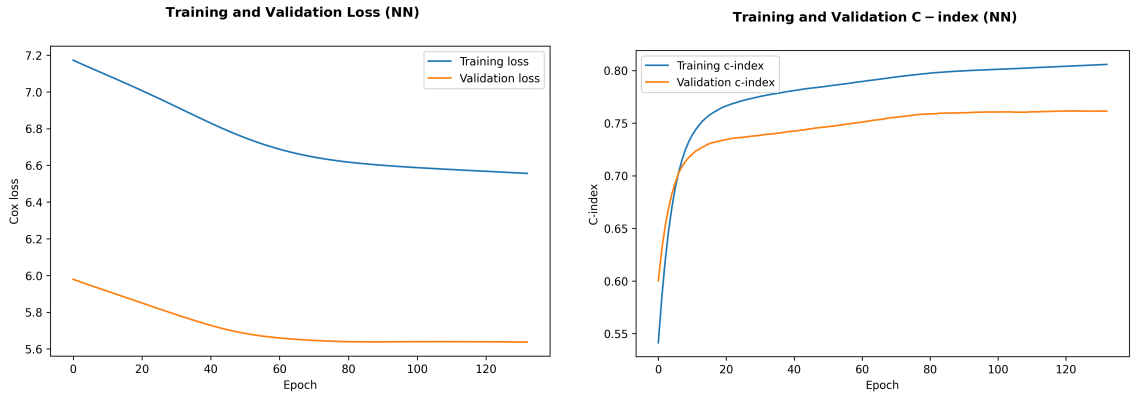
6.2 Results

This section presents the results from the tabular survival analysis experiment. The CoxPH model and the neural network are evaluated using the concordance index, followed by an examination of global and local feature attributions derived from SHAP, Integrated Gradients, and CoxPH coefficients.

6.2.1 Model Performance

The performance of the Cox Proportional Hazards model and the neural network was evaluated using the C-index. The CoxPH model achieved a validation C-index of 0.756 and a test C-index of 0.762, while the neural network achieved a validation C-index of 0.762 and a test C-index of 0.768. Overall, the models demonstrated similar predictive performance, with the neural network showing only a marginal improvement on the test set.

The training curves of the neural network show stable convergence, with a steady decrease in loss and a corresponding increase in C-index. The validation performance follows a similar trend with only a small gap between the training and validation curves, indicating limited overfitting. Performance stabilises after approximately 80 epochs.



(a) Training and validation loss.

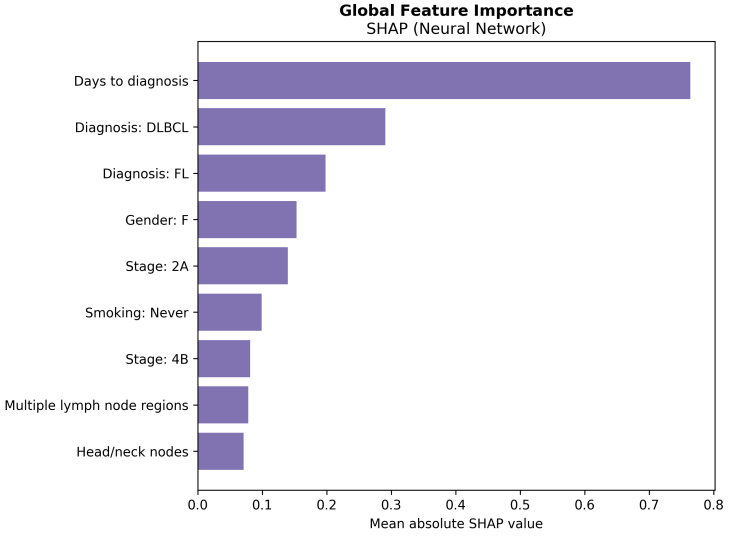
(b) Training and validation C-index.

Figure 6.1: Training dynamics of the neural network model.

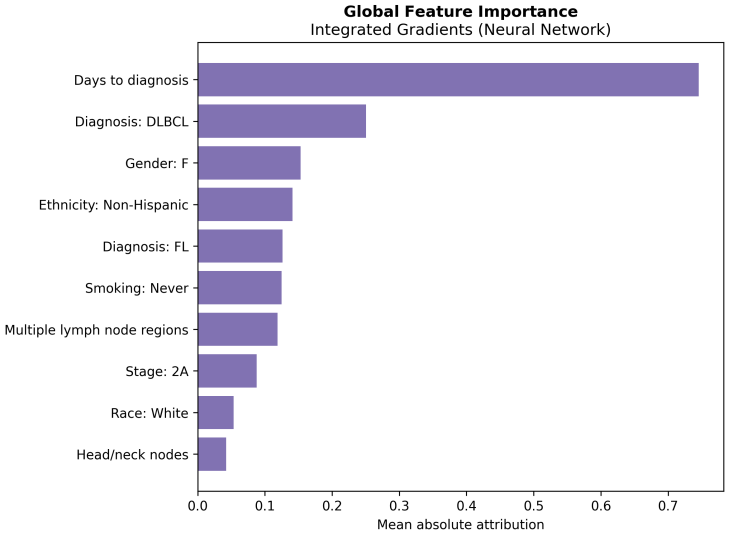
6.2.2 Global Explanations

Global feature importance was analysed using SHAP and Integrated Gradients for the neural network, and using Cox coefficients and SHAP values for the CoxPH model.

For the neural network, both SHAP and IG identified days to diagnosis and Diffuse Large B-Cell Lymphoma (DLBCL) as the most influential features, as shown in Figure 6.2. Additional diagnosis-related variables, including Follicular Lymphoma (FL), also showed moderate importance, together with demographic variables such as gender, ethnicity, and smoking status. The similarity between SHAP and IG indicates consistent feature attribution patterns across methods.



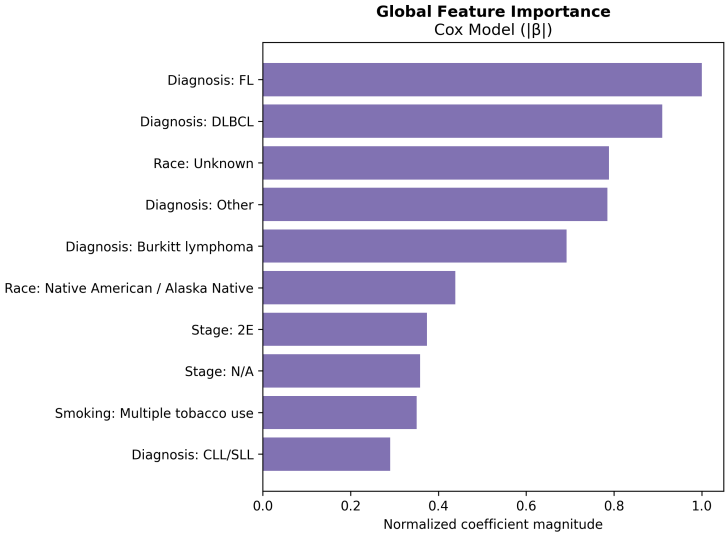
(a) SHAP (NN).



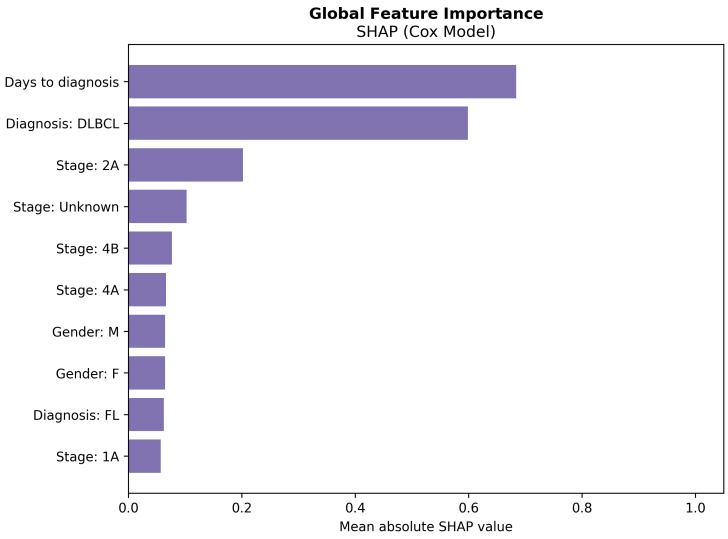
(b) Integrated Gradients (NN).

Figure 6.2: Global feature importance for the neural network using SHAP and Integrated gradients.

For the CoxPH model, global feature importance based on Cox coefficients and SHAP values is shown in Figure 6.3. Both approaches highlight diagnosis-related variables as dominant, particularly follicular lymphoma and Diffuse Large B-Cell Lymphoma. Additional diagnosis-related features, including Burkitt lymphoma and Chronic Lymphocytic Leukemia/Small Lymphocytic Lymphoma (CLL/SLL), also contributed to the predictions, together with staging and demographic variables. Although both approaches highlight diagnosis-related variables, differences in feature ranking and relative importance are observed between SHAP and coefficient-based explanations.



(a) Cox coefficients.



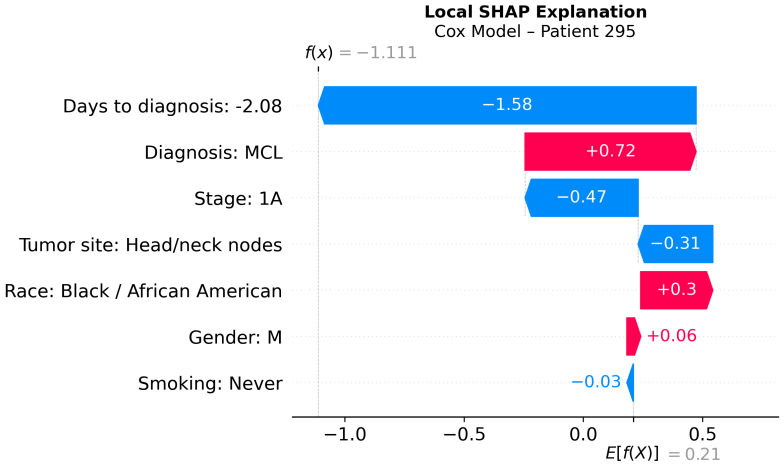
(b) SHAP (CoxPH model).

Figure 6.3: Global feature importance for the CoxPH model based on Cox coefficients and SHAP values.

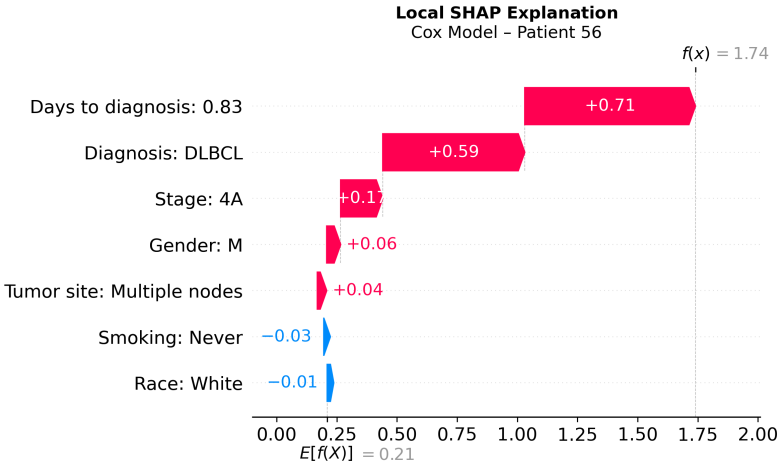
6.2.3 Local Explanations

Local feature attributions were analysed to investigate how individual features contributed to predictions for two representative patients with contrasting risk profiles: a low-risk patient (Patient 295) and a high-risk patient (Patient 56). SHAP was applied to both the CoxPH model and the neural network, while Integrated Gradients was additionally used for the neural network. The results are shown in Figures 6.4, 6.5, and 6.6.

For the CoxPH model (Figure 6.4), the predictions are primarily driven by diagnosis and disease stage. In Patient 295, Mantle Cell Lymphoma (MCL) contributes positively to the predicted risk, while days to diagnosis contributes negatively. In Patient 56, Diffuse Large B-Cell Lymphoma (DLBCL) and higher disease stages further increase the predicted risk. Overall, the contributions are relatively consistent across patients, reflecting the linear and additive structure of the CoxPH model.



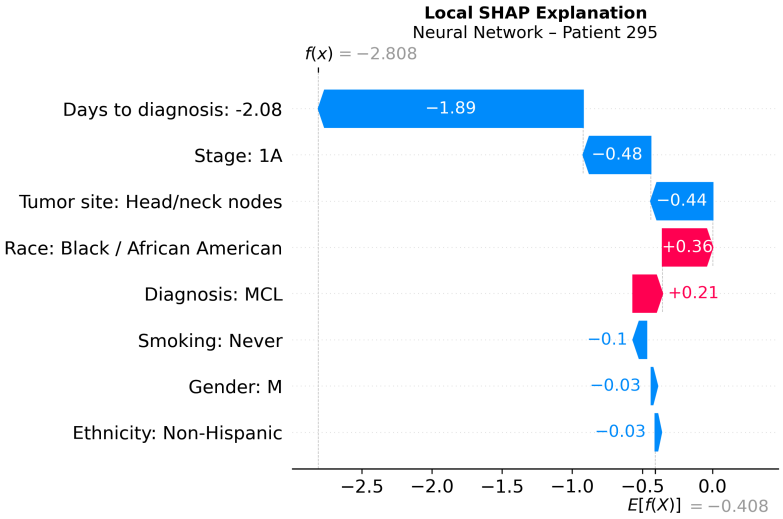
(a) Patient 295 (low risk).



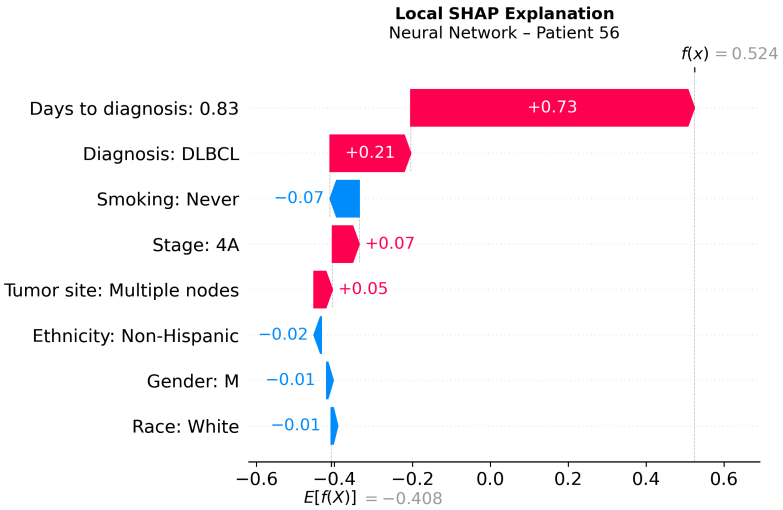
(b) Patient 56 (high risk).

Figure 6.4: Local SHAP explanations for the CoxPH model.

For the neural network, SHAP explanations (Figure 6.5) reveal similar key drivers, with diagnosis and staging variables remaining dominant. However, the magnitude and relative importance of features vary more between the two patients compared to the CoxPH model, indicating greater flexibility in how feature contributions vary across patients.



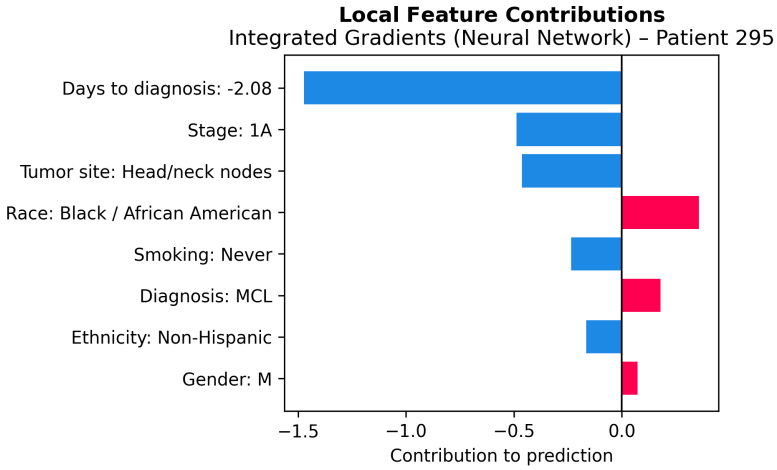
(a) Patient 295 (low risk).



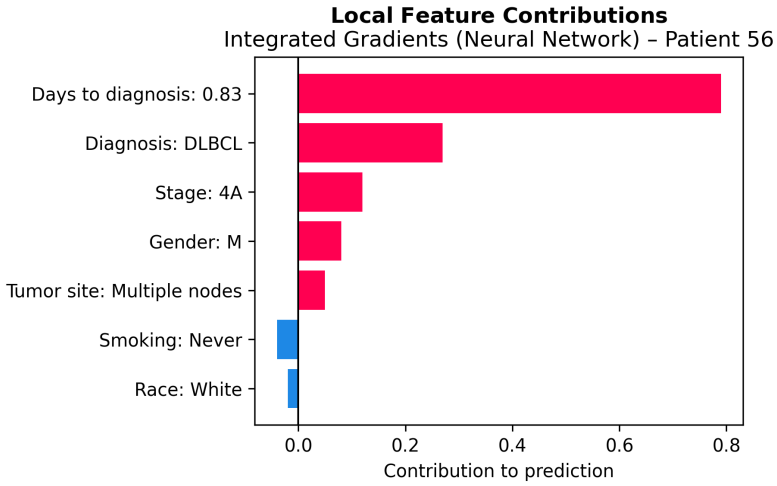
(b) Patient 56 (high risk).

Figure 6.5: Local SHAP explanations for the neural network.

Integrated Gradients explanations (Figure 6.6) are largely consistent with the SHAP results for the neural network. In particular, days to diagnosis and DLBCL remain the most influential features, with similar contribution directions across both patients. While minor differences in magnitude are observed, the overall attribution patterns are aligned between the two methods.



(a) Patient 295 (low risk).



(b) Patient 56 (high risk).

Figure 6.6: Local Integrated Gradients explanations for the neural network.

Overall, both models identify diagnosis and time-related variables as the primary drivers of individual predictions. While the neural network showed greater variability in feature contributions across patients, the overall attribution patterns remained largely consistent between models and explainability methods.

6.3 Discussion

This section discusses the results obtained from both the CoxPH model and the neural network, with a focus on model evaluation and explainability.

6.3.1 Model Evaluation

The results show that both the CoxPH model and the neural network were capable of predicting patient risk with relatively high accuracy, as reflected by the concordance index. The predictive performance of the two models was highly similar, suggesting that much of the predictive signal present in the structured clinical data could already be captured by the linear CoxPH formulation.

The stable validation performance and limited overfitting observed during training support the reliability of the neural network predictions. Establishing this level of predictive stability is an important prerequisite for XAI analysis, since unreliable or unstable models may produce misleading explanations.

6.3.2 Explainability and Reliability

Global Explainability: The global explainability analysis showed strong agreement across methods and models. Diagnosis-related variables, particularly the presence of Diffuse Large B-Cell Lymphoma, together with time-related features such as days to diagnosis, consistently emerged as the most influential predictors. The alignment between SHAP, Integrated Gradients, and CoxPH coefficients suggests that the identified relationships are stable and not merely artefacts of a specific explainability method.

The consistency observed across explanation methods also increases confidence in the reliability of the identified feature patterns. Since SHAP and Integrated Gradients rely on fundamentally different attribution principles, the similarity in their explanations suggests that the identified feature patterns reflect meaningful model behaviour rather than instability related to a single explainability technique.

Local Explainability: At the local level, both models identified similar key drivers for individual predictions. However, some differences were observed in how feature contributions varied between patients. The local explanations generated by the CoxPH model were generally more predictable and easier to interpret, since feature effects followed relatively consistent patterns across individuals. In contrast, the neural network showed greater variability in feature attribution magnitudes across individuals, suggesting increased flexibility in how clinical features contributed to predictions.

Although the neural network may capture more complex relationships, the relatively small performance difference compared to the CoxPH model suggests that simpler, more interpretable models may already provide sufficiently strong predictive performance on structured clinical data.

7

Imaging Model

7.1 Experiments

To complement the tabular model based on clinical variables, a CNN-based framework was developed using imaging data. This model was designed not only to perform survival prediction, but also to enable the use of XAI methods to interpret image-based predictions.

7.1.1 Model Architecture

A CNN was developed to predict patient-specific risk scores from PET imaging data. The model utilised a ResNet-18 architecture [26] pre-trained on ImageNet [27] as its convolutional backbone. To adapt the network for survival analysis, the final classification layer was removed, and the remaining architecture was used for feature extraction. To prevent overfitting and stabilise training, all convolutional layers in the backbone were kept frozen, allowing the model to leverage general image features learned from large-scale data.

The extracted feature vector was passed through two fully connected layers with ReLU activation, ending in a final linear layer that produced a single patient-specific risk score. This architecture integrated convolutional feature extraction into a Cox proportional hazards framework, defined in Equation 2.4, resulting in a continuous value representing a patient's relative risk.

7.1.2 Training Procedure

The model was trained using the Cox partial likelihood loss (Equation 2.4), which enables learning from censored survival data by comparing relative risk between patients within each mini-batch. Optimisation was performed using the Adam optimiser with a learning rate and weight decay of 1×10^{-5}. Training was conducted with a batch size of 32 for up to 100 epochs.

Model performance was evaluated on the validation set after each epoch. Early stopping was applied based on the validation C-index with a patience of 10 epochs, and the model parameters corresponding to the best validation performance were restored after training.

7.1.3 Model Evaluation

Model performance was evaluated using the C-index. During training, loss and C-index curves were monitored for both the training and validation sets to assess convergence and stability. Final model performance was evaluated on the held-out test set.

7.1.4 Explainability

XAI methods were applied to the imaging framework to provide insight into the model's internal decision-making process and highlight the regions within the PET input data contributing most strongly to the predicted risk scores. Three explainability approaches were utilised: Grad-CAM (Equation 2.6), Integrated Gradients (Equation 2.7), and Occlusion (Equation 2.8).

Single-layer Grad-CAM: In the single-layer approach, the LayerGradCam method from the Captum library [28] was applied to the first convolutional layer (conv1) of the ResNet-18 backbone. This layer was selected to prioritise spatial resolution, as early convolutional layers retain detailed anatomical information compared to deeper layers, which capture more abstract features. The attribution process involved a ReLU operation to isolate positive contributions, followed by channel-wise averaging and normalisation to scale all activations between 0 and 1.

Multi-layer Grad-CAM: The multi-layer Grad-CAM approach extended this analysis across the entire CNN hierarchy by extracting features from all convolutional blocks within the ResNet-18 architecture. This approach captures a spectrum of information, ranging from local high-resolution details in early layers to abstract global patterns in deeper layers. Heatmaps were computed separately for each layer and then aggregated by averaging to create a unified representation of importance across the network.

Positive and Negative Grad-CAM Attributions: To provide a more nuanced understanding of the model's logic, Grad-CAM was also computed without the standard ReLU constraint, thereby retaining both positive and negative attribution values. In this framework, positive values signify regions that increase the predicted risk, while negative values identify regions that decrease it. These signed maps were normalised by their maximum absolute value and visualised using a diverging color map, with positive values shown in red and negative values in blue.

Integrated Gradients: Integrated Gradients served as a complementary pixel-level attribution method, implemented using the IntegratedGradients function (equation 2.7) from the Captum library. The algorithm utilised a zero-valued baseline image to compute attributions along an interpolation path toward the actual input. The resulting maps were averaged across input channels, processed with a ReLU operation, and normalised by their maximum value.

Occlusion: Occlusion sensitivity was utilised as an additional attribution method, implemented via the Occlusion function (equation 2.8) in the Captum library. A sliding window approach with a fixed patch size and stride was employed to mask input regions with a baseline value of zero. A patch size of 20 and a stride of 4 were used, balancing spatial resolution with computational feasibility. The resulting attribution maps, representing the change in risk score, were averaged across channels and normalised.

Visualisation: To enhance interpretability and reduce high-frequency noise, a Gaussian filter was applied to create a smoothed version of all heatmaps. Furthermore, a percentile-based threshold was applied to generate a thresholded version where only pixels above the 95th percentile were retained. Finally, a binary mask was generated from the thresholded data to clearly outline the boundaries of the model's focus within the PET scan.

7.2 Results

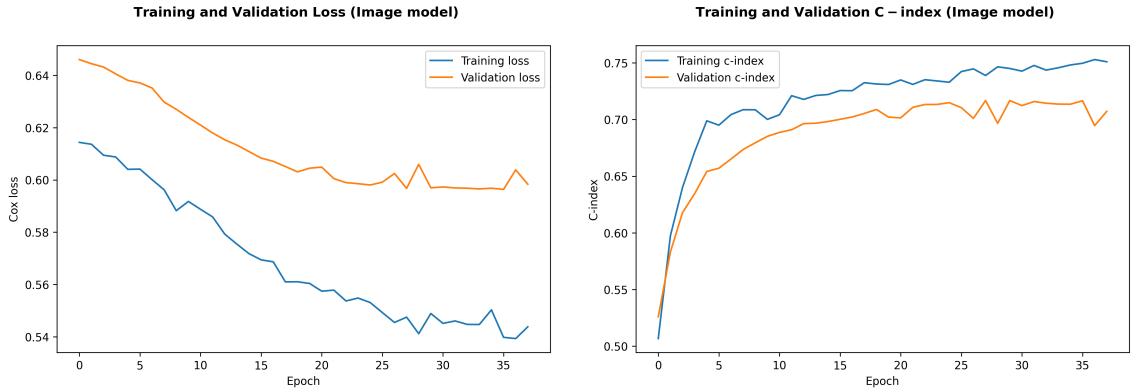
This section presents the results from the image-based survival analysis experiment. The CNN model is evaluated using the concordance index, followed by an examination of spatial feature attributions derived from Grad-CAM, Integrated Gradients, and Occlusion Sensitivity.

7.2.1 Model Performance

The CNN model achieved a validation C-index of approximately 0.72 and a test C-index of around 0.74, indicating strong predictive performance.

Figure 7.1 illustrates the training dynamics of the CNN model. As shown in Figure 7.1a, the training loss decreased consistently throughout the training process, while the validation loss stabilised after approximately 20 epochs, exhibiting minor fluctuations thereafter. While the divergence between the training and validation loss suggests the onset of mild overfitting, the use of early stopping ensured the selection of a model with high generalisability.

Regarding predictive accuracy, Figure 7.1b shows that the C-index for both the training and validation sets increased rapidly during the first 10 epochs. Beyond this point, the improvement rate slowed, maintaining a small and consistent gap between training and validation performance throughout the remainder of the training process.



(a) Training and validation loss.

(b) Training and validation C-index.

Figure 7.1: Training dynamics of the CNN model.

7.2.2 Grad-CAM

Single-layer Grad-CAM attributions for a representative patient are displayed in Figures 7.2 and 7.3.

The raw Grad-CAM output in Figure 7.2 identifies high-intensity pixels that align with regions of high metabolic activity in the original PET scan. While this localised attribution confirms that the model focuses on relevant signal sources, the raw heatmaps contains high-frequency noise which obscure larger anatomical patterns, making clinical interpretation challenging.

To address this, smoothed attribution maps were generated, as shown in Figure 7.3. The Gaussian filter successfully merged the individual activation points into larger, more spatially coherent clusters. The thresholded heatmap and binary mask further isolated these regions by suppressing low-level background activations.

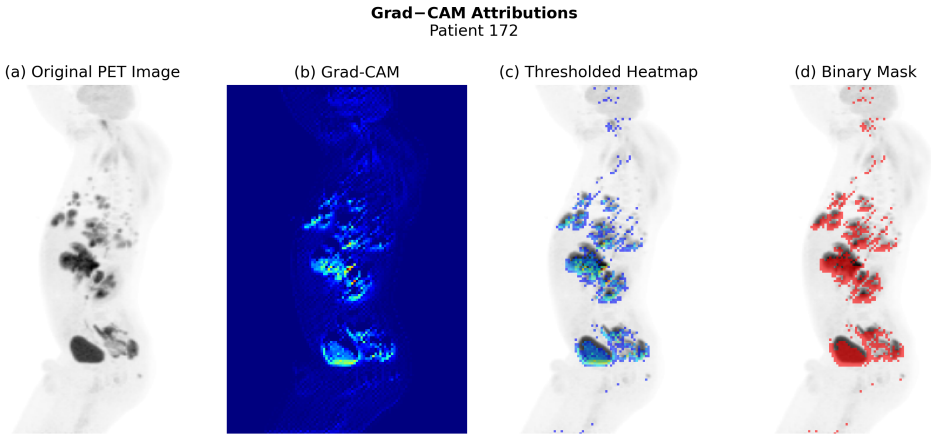


Figure 7.2: Raw Grad-CAM attributions for a representative patient using the first convolutional layer of the CNN.

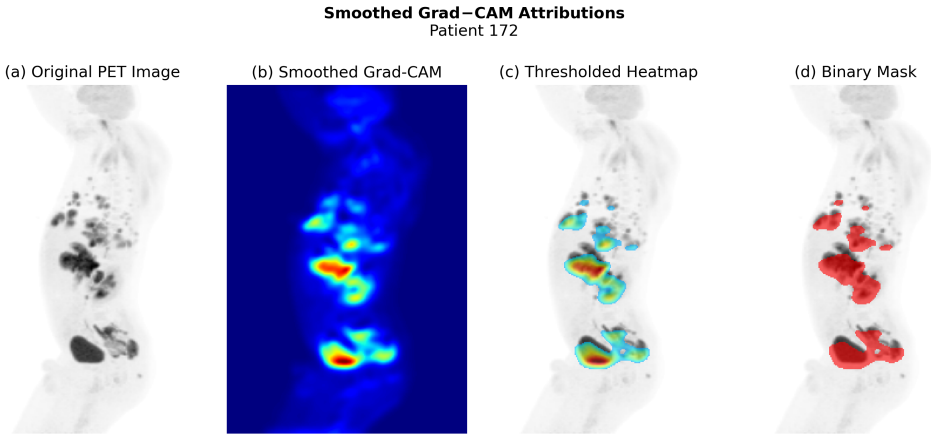


Figure 7.3: Smoothed Grad-CAM attributions for a representative patient using the first convolutional layer of the CNN.

The results for the multi-layer Grad-CAM approach are illustrated in Figures 7.4 and 7.5, with an individual layer breakdown provided in Figure 7.6.

The combined raw multi-layer Grad-CAM in Figure 7.4 highlights significantly larger regions of importance compared to the single-layer analysis. As shown in Figure 7.5, smoothing improves spatial coherence, but the results remain notably diffuse. Although Gaussian filtering reduces pixel-level noise, it does not resolve the issue of widespread activation.

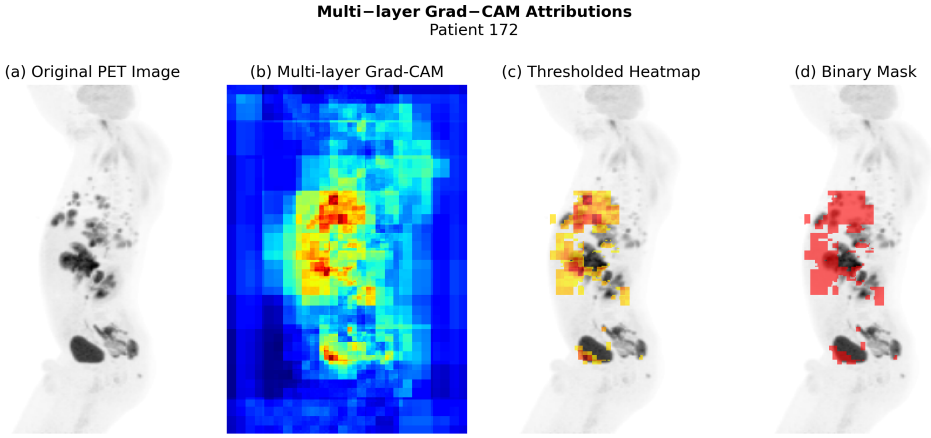


Figure 7.4: Raw multi-layer Grad-CAM attributions for a representative patient, generated by averaging attributions across all convolutional blocks.

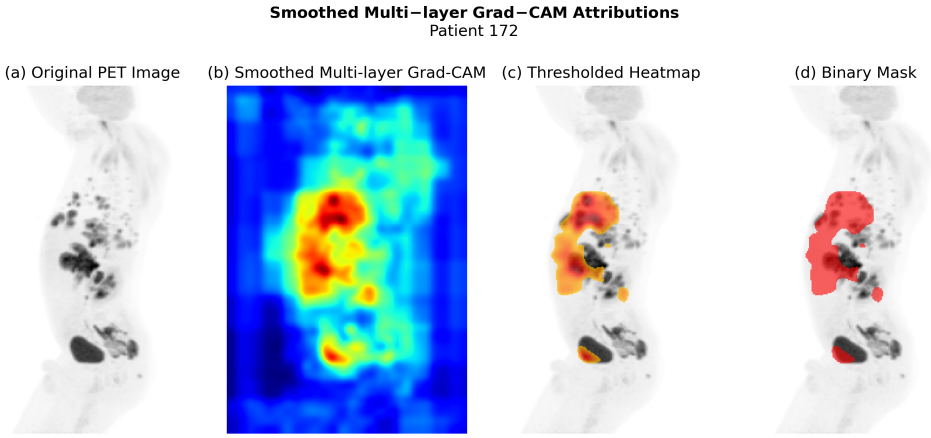


Figure 7.5: Smoothed multi-layer Grad-CAM attributions for a representative patient, generated by averaging attributions across all convolutional blocks.

Figure 7.6 reveals the cause of this diffusion as individual layers focus on distinct and often non-overlapping features within the PET image. While early layers capture fine-grained, localised points, the deeper layers shift their focus toward broader anatomical contexts. Because these individual attributions lack a consistent spatial consensus, averaging them into a single representation results in the smeared effect observed in the final aggregate maps.

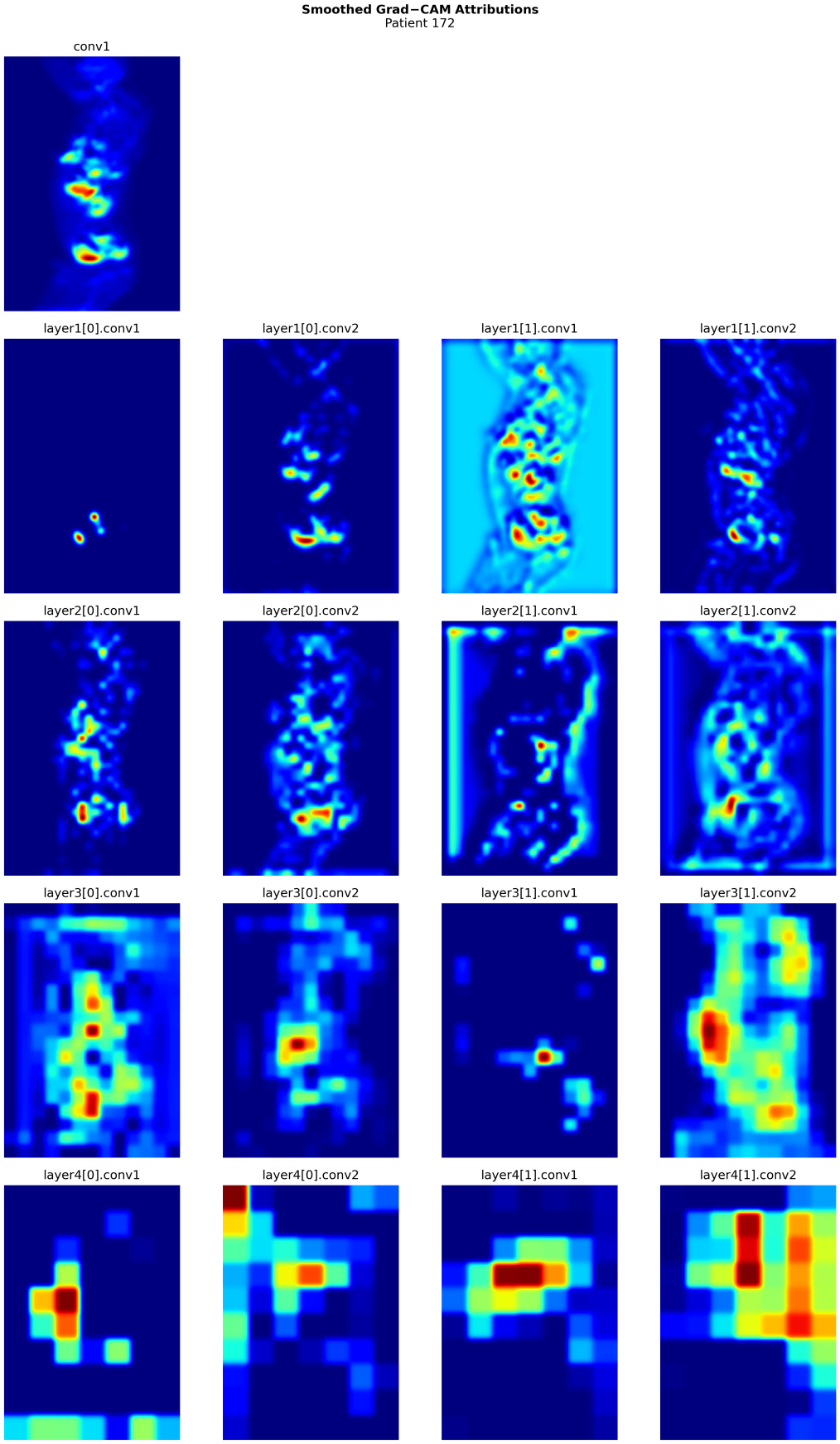


Figure 7.6: Smoothed Grad-CAM attributions for all individual convolutional layers.

Figure 7.7 displays Grad-CAM visualisations generated without a ReLU constraint for three representative patients. Regions increasing the predicted risk are shown in red, while regions decreasing the risk are shown in blue.

For high-risk patients (172 and 280), both positive and negative attributions are concentrated within high-intensity PET regions. In contrast, the low-risk patient (43) exhibits more widespread attributions in lower-intensity areas. No clear difference in the total volume of positive versus negative activations is observed between high- and low-risk cases. Notably, the same anatomical structure may appear as either risk-increasing or risk-decreasing depending on the layer being analysed.

A consistent layer-dependent pattern is observed across all patients. Layers (b) and (e) primarily identify risk-increasing (red) regions, while layers (c) and (d) consistently highlight risk-decreasing (blue) regions, often targeting the same anatomical structures.

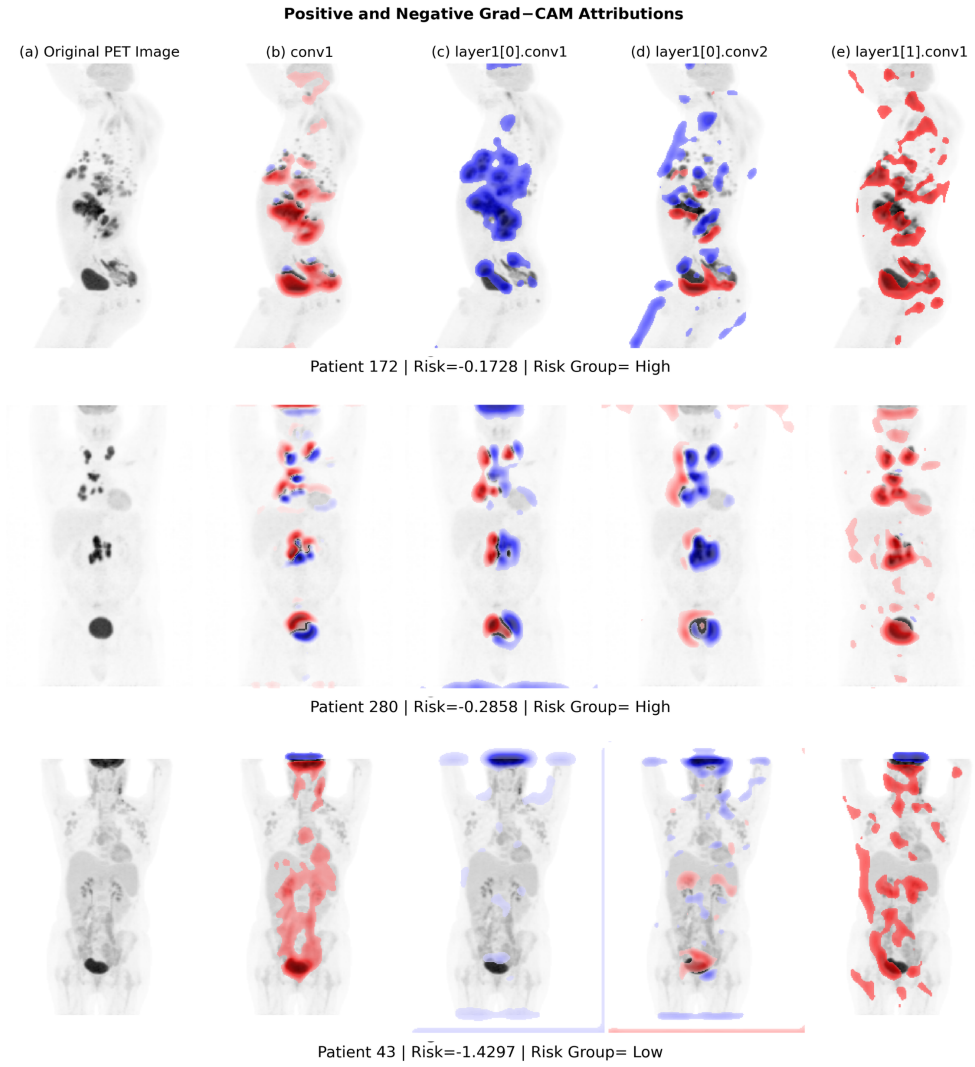


Figure 7.7: Positive and negative Grad-CAM attributions across network layers.

7.2.3 Integrated Gradients

Figures 7.8 and 7.9 present the Integrated Gradients results, which provide a high-resolution map of the features driving the model's predictions.

The raw IG output in Figure 7.8 identifies specific high-intensity pixels within the PET signal as the primary drivers of the risk score. While these influential pixels align with regions of high metabolic activity, the raw visualisation contains significant pixel-level noise, which obscures continuous anatomical structures.

To improve interpretability, smoothed IG results are shown in Figure 7.9. Applying a Gaussian filter aggregates the isolated pixel attributions into more defined, spatially coherent regions. The resulting thresholded heatmap and binary mask effectively isolate the most significant clusters, providing a clearer indication of the specific lesions contributing to the final risk assessment.

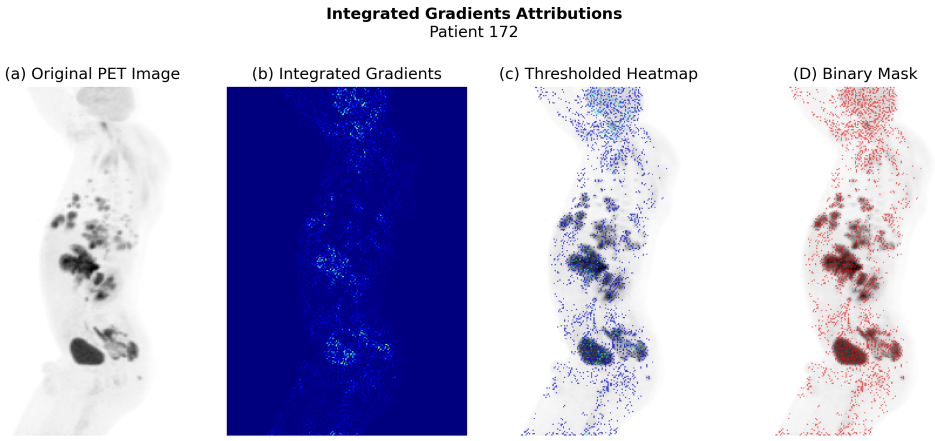


Figure 7.8: Raw Integrated Gradients attributions for a representative patient.

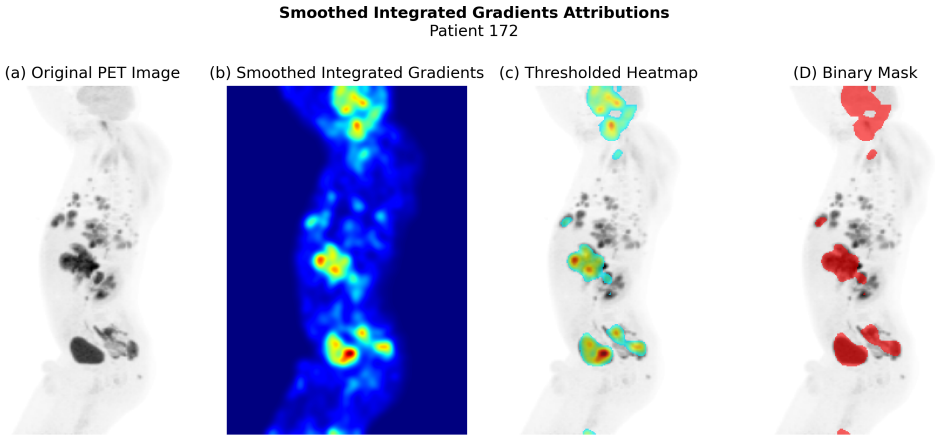


Figure 7.9: Smoothed Integrated Gradients attributions for a representative patient.

7.2.4 Occlusion

Figures 7.10 and 7.11 illustrate the results of the Occlusion sensitivity analysis.

The raw occlusion map in Figure 7.10 identifies two primary regions of importance. However, unlike the previous attribution methods, these regions do not align directly with the high-uptake metabolic signals in the PET scan. Instead of focusing on specific focal points, the model sensitivity appears distributed over a much broader spatial area.

As shown in Figure 7.11, applying a Gaussian filter merges these attributions into more continuous shapes. Even after smoothing, the thresholded heatmap and binary mask remain diffuse, covering large areas of relatively low signal intensity. These results suggest that the Occlusion method captures broader, contextual regions of interest rather than isolating the specific anatomical or metabolic features.

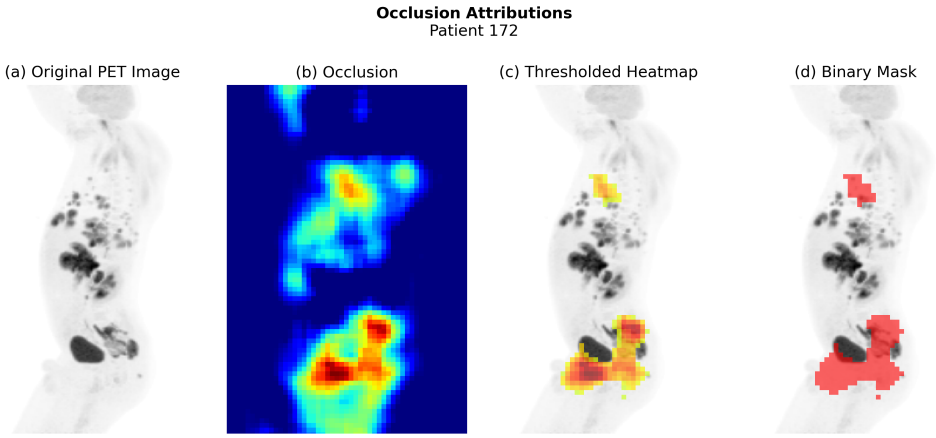


Figure 7.10: Raw Occlusion attributions for a representative patient.

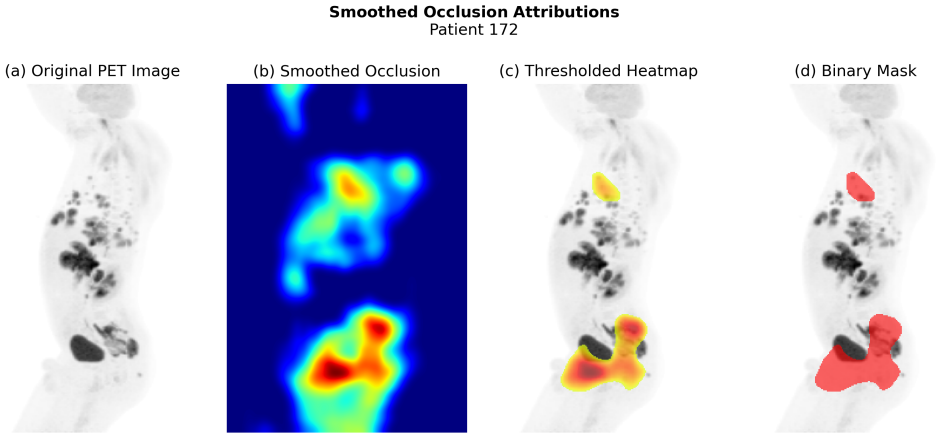


Figure 7.11: Smoothed Occlusion attributions for a representative patient.

7.3 Discussion

This section discusses the results obtained from the CNN-based image survival model, with a focus on model performance and the reliability of the explainability analyses.

7.3.1 Model Performance

The predictive performance of the CNN model serves as the basis for the XAI analysis. The test C-index of 0.74, which slightly exceeds the validation score of 0.72, suggests that the model generalises well to unseen data and captures meaningful anatomical features rather than merely memorising patterns present in the training cohort. This level of accuracy is essential for XAI. Since the model's predictions are fundamentally accurate and stable, the resulting XAI attributions carry the clinical validity necessary to interpret the model's internal decision-making process.

7.3.2 Explainability and Reliability

Single-layer Grad-CAM: The single-layer Grad-CAM results provide intuitive visualisations by utilising the first convolutional layer (conv1), which preserves the spatial resolution necessary to pinpoint anatomical structures driving the prediction. Post-processing proved essential for interpretability, with smoothing creating spatially coherent regions and thresholding filtering noise to highlight the most significant risk predictors. Notably, binary masks proved less informative than heatmaps, as they lack the intensity gradients required to rank pixel importance. The visualisations suggest that the model focuses on clinically relevant image regions while ignoring background noise, increasing confidence in the model's internal representations before moving to more complex multi-layer interpretations.

Multi-layer Grad-CAM: Transitioning from a single-layer analysis to a multi-layer approach introduces significant complexity, as Grad-CAM results vary substantially across the network. Because different layers focus on distinct and often non-overlapping features, simple averaging produces diffuse heatmaps that lack the sharp localisation required for clinical interpretation. In this multi-layer context, standard post-processing techniques like smoothing and thresholding provide only limited improvement. Although these methods effectively reduce pixel-level noise in single-layer explanations, they cannot resolve the underlying lack of spatial agreement between layers.

This inconsistency creates a "layer selection dilemma", where one must decide which layer, or combination of layers, most accurately represents the model's decision-making process. This challenge highlights a major obstacle in standardising XAI for medical applications. The issue extends beyond the trade-off between the global context of deeper layers and the pixel-level precision of early ones. Attribution patterns also vary between neighboring layers within the same architectural block, suggesting that the model's internal focus is highly dynamic.

Positive and Negative Attributions: Removing the ReLU constraint reveals an important internal inconsistency, where the same anatomical region can be interpreted as having opposite influences on risk depending on the layer analysed. This suggests a form of layer specialisation in which different parts of the network interpret identical biological signals in conflicting ways. A region may be marked as risk-decreasing by one layer and risk-increasing by another. This issue reinforces the layer selection dilemma and raises concerns about the reliability of Grad-CAM in clinical settings. Because the explanation is highly sensitive to architectural quirks and choice of layer, these visualisations lack the stability required for clinical transparency.

Integrated Gradients: IG allows for the identification of specific high-intensity pixels influencing the risk score. While raw IG heatmaps are inherently noisier than Grad-CAM due to the lack of spatial smoothing from convolutional layers, post-processing successfully aggregates attributions into interpretable regions.

After smoothing, there is significant overlap between IG and Grad-CAM results, particularly in regions of high metabolic activity. This agreement between two mathematically distinct XAI methods suggests that the model consistently relies on these high-uptake features when generating survival predictions.

Occlusion: Occlusion sensitivity analysis yields results fundamentally different from gradient-based methods, marking larger, more diffuse areas rather than precise high-uptake hotspots. This distributed sensitivity suggests the model may rely on broader spatial context or inter-regional relationships rather than raw feature intensity. However, the clinical utility of Occlusion is limited by its dependence on hyperparameters, specifically patch size and stride. Because the resulting heatmap can be significantly altered by changing the dimensions of the sliding mask, the method lacks the stability required for medical applications.

Comparative Analysis and Clinical Outlook: Across all methods, the gradient-based approaches, Grad-CAM and Integrated Gradients, identified overlapping regions of high metabolic activity, providing mutual support for the biological plausibility of the detected patterns. In contrast, Occlusion captured broader contextual sensitivity and produced explanations that were less anatomically precise.

Taken together, the results reveal a fundamental challenge for post-hoc explainability in medical imaging. The generated explanations were highly sensitive to method-specific choices, and different methods or network layers could produce substantially different interpretations of the same prediction. Such inconsistencies complicate the interpretation of XAI outputs and limit the transparency these methods are intended to provide. For XAI to move beyond initial trust-building toward genuine clinical utility, more standardized evaluation frameworks are needed to assess explanation consistency and determine which attributions are clinically relevant. This challenge connects directly to the broader findings of this thesis and is discussed further in the integrated discussion.

8

Model Comparison

8.1 Experiments

Having evaluated the tabular and imaging models independently, this chapter brings them into direct comparison to assess whether the two modalities capture complementary or overlapping aspects of patient risk. The comparison focuses on predictive performance and risk stratification rather than explainability methods.

8.1.1 Kaplan-Meier Analysis

To assess the ability of the models to stratify patients into risk groups, Kaplan-Meier survival analysis (Section 2.3.2) was performed using predicted risk scores from both the tabular neural network and the convolutional neural network trained on imaging data.

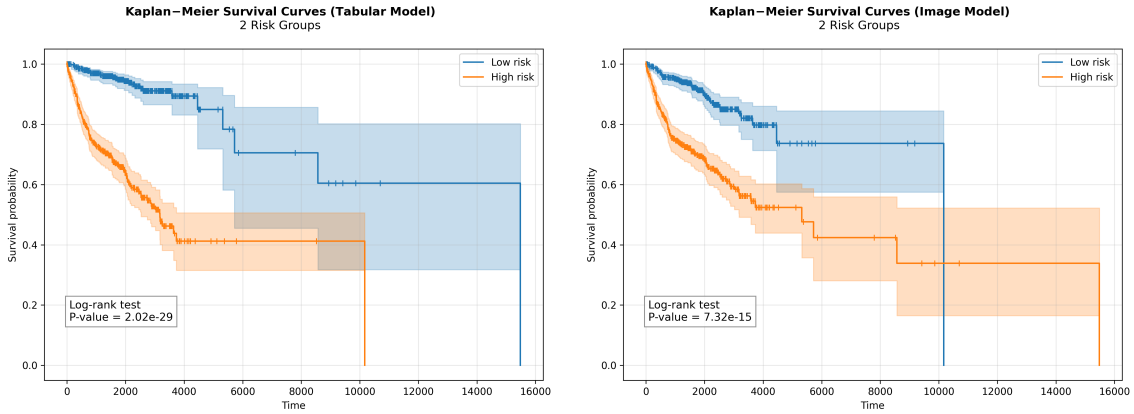
Patients in the test set were stratified into two risk groups (low and high risk) based on the median predicted risk score for each model. Kaplan-Meier survival curves were then generated to visualise differences in survival probability over time, both within each model and in a direct comparison between the two models.

The log-rank test was used to assess statistical differences between risk groups. A significant p-value indicates successful separation of patients into groups with distinct survival outcomes.

8.2 Results

Kaplan-Meier analysis demonstrated significant separation between low- and high-risk groups for both the tabular neural network and the convolutional neural network ($p < 0.001$ for both models), as shown in Figure 8.1.

A direct comparison between the risk stratifications produced by the two models is shown in Figure 8.2. A significant difference was observed between the low-risk groups predicted by the models ($p < 0.01$), whereas no significant difference was observed between the high-risk groups ($p = 0.09$).



(a) Tabular neural network.

(b) Convolutional neural network.

Figure 8.1: Kaplan-Meier survival curves for low- and high-risk groups predicted by the tabular and imaging models.

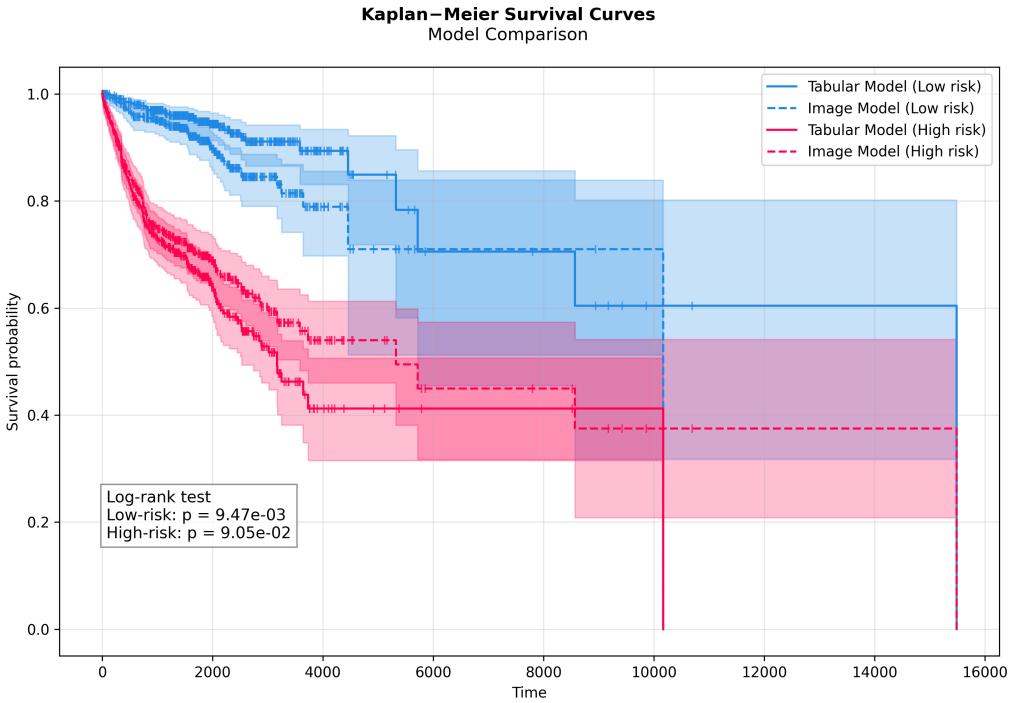


Figure 8.2: Kaplan-Meier survival curves comparing low- and high-risk groups predicted by the tabular and imaging models.

8.3 Discussion

Both models demonstrated the ability to stratify patients into clinically meaningful risk groups. However, the tabular model showed clearer separation between low- and high-risk groups, indicating stronger discriminative performance. The greater overlap observed for the CNN model suggests that image-based features capture risk-related patterns in a less distinct manner, potentially reflecting more complex or heterogeneous signals compared to structured clinical variables.

As shown in the results, both models appeared more consistent in identifying high-risk patients than low-risk patients. One possible explanation is that high-risk patients exhibit stronger and more consistent patterns across both structured clinical variables and imaging features, making them easier to identify regardless of modelling approach. In contrast, low-risk patients may represent a more heterogeneous group, where favourable outcomes arise from different combinations of clinical and biological factors. In such cases, the relative weighting of structured variables versus image-derived features may differ between models, leading to greater disagreement in risk assignment.

Kaplan-Meier analysis evaluates group-level separation rather than exact individual risk prediction. Therefore, the comparison was complemented by C-index evaluation. As shown in Sections 6.2.1 and 7.2.1, the tabular model achieved a higher C-index on the test set (0.77) compared to the CNN (0.74), indicating better overall predictive accuracy. This is consistent with the clearer separation observed in the Kaplan-Meier curves shown in Figure 8.2. Taken together, these findings suggest that while both approaches are effective, the tabular model appears to provide stronger risk stratification.

Part III

Integrated Discussion

9

Discussion

9.1 Overview of the Study and Main Findings

This thesis investigated explainable artificial intelligence in healthcare from both technical and clinical perspectives, combining qualitative interviews with physicians and technical evaluations of explainability methods applied to survival prediction models.

The interview study revealed cautious optimism toward AI-based decision support systems, with physicians emphasising the importance of trust, usability, and seamless workflow integration. The technical evaluation demonstrated that both tabular and imaging-based models achieved strong predictive performance. Kaplan-Meier analysis further confirmed that both models were able to stratify patients into clinically meaningful risk groups, although the tabular model showed clearer separation. At the same time, the variability observed across explainability methods highlighted important challenges regarding the reliability, consistency, and clinical applicability of current post-hoc XAI approaches.

9.2 Performance and Interpretability

Important differences emerged between model types regarding interpretability and explanation stability. The tabular models generally produced more consistent and clinically interpretable explanations, likely because structured clinical variables align more naturally with human reasoning and established medical knowledge. In contrast, the imaging-based models generated visually intuitive heatmaps, but the resulting explanations were often more difficult to validate and showed greater variability across explainability methods.

For the tabular data, the CoxPH model and DeepSurv achieved relatively similar predictive performance, suggesting that much of the predictive signal is already captured by simpler linear relationships. Although the neural network enabled more flexible attribution patterns across patients, it also introduced additional challenges regarding interpretability and consistency. These findings reflect a broader trade-off in explainable AI research, where increased predictive complexity may reduce transparency. This was also reflected in the interview study, where physicians generally preferred explainable systems over highly complex black-box models, and where

many expressed a preference for a slightly less accurate but interpretable model over a high-performing one that offered no insight into its reasoning. In healthcare settings, small improvements in predictive performance may therefore not justify reduced interpretability, particularly in high-stakes clinical decision-making.

9.3 Reliability and Consistency of XAI Methods

An important question in explainable AI is whether explanations themselves can be considered reliable and robust. For the tabular models, SHAP and Integrated Gradients showed relatively similar overall patterns, with the same features consistently appearing among the most influential predictors. Although some differences in magnitude were observed, the overall agreement increases confidence that the identified patterns reflect meaningful model behaviour rather than artefacts of a single explainability technique.

For the imaging models, the explainability results were considerably less consistent. Single-layer Grad-CAM and Integrated Gradients showed some overlap in regions of high metabolic activity, whereas the Occlusion method produced substantially different and more diffuse maps. Furthermore, Grad-CAM explanations varied depending on the selected convolutional layer, and the same anatomical region could simultaneously appear as risk-increasing in one layer and risk-decreasing in another. This internal inconsistency makes it very difficult for a clinician to understand how conflicting signals are reconciled in the final prediction, and raises serious questions about using such explanations as a basis for clinical decisions.

This variability highlights a fundamental challenge for clinical XAI implementation. Current post-hoc explainability methods lack standardisation, and there is no established framework for resolving contradictions between different methods or layers. This connects directly to the concerns expressed by physicians in the interview study, who highlighted that persuasive visualisations could create false confidence if interpreted uncritically. The limitations of current XAI methods must therefore be carefully considered before these tools can be safely integrated into clinical practice.

9.4 Clinical Usability of AI Explanations

The interview study showed that physicians preferred explanations that were concise, intuitive, and adapted to clinical workflows, supporting rapid decision-making without introducing unnecessary complexity or additional workload.

For the tabular models, the explainability approaches aligned relatively well with these preferences. Feature grouping reduced the complexity introduced by one-hot encoding and allowed explanations to be presented at the level of clinically meaningful concepts such as diagnosis, stage, and smoking status. Visualisations clearly showed whether features contributed positively or negatively to the predicted risk, while avoiding excessive numerical detail. Including patients' actual feature values within the explanations further improved interpretability by helping physicians

relate the output directly to the clinical context.

For the imaging models, physicians generally preferred simple visual overlays where the original image remained visible alongside the heatmap. Post-processing methods such as smoothing and thresholding improved visual clarity, though simplification may also reduce nuance and obscure uncertainty in the attribution maps. Notably, several physicians stated that they primarily wanted a simple indication of where to focus their attention before performing their own interpretation, suggesting that simpler approaches such as bounding boxes or threshold-based markings may in some situations be more practically useful than complex heatmaps. This highlights an important distinction between technically detailed and clinically useful explanations, and implies that the design of XAI outputs should be guided by clinical workflow requirements rather than technical complexity alone.

9.5 Human-AI Collaboration

Beyond the format and usability of explanations, the interview study raised a more fundamental question about the role of AI in clinical practice. Physicians consistently perceived AI as a supportive tool rather than a replacement for clinical judgement. Effective human-AI collaboration was considered more appropriate than full automation, particularly in complex and uncertain clinical situations.

This framing also defines what explainability needs to achieve. Rather than providing complete technical transparency, explanations primarily need to support clinician oversight and enable calibrated trust. Physicians did not require detailed descriptions of the underlying algorithms. Instead, they wanted recommendations to be clinically motivated and open to inspection, much like advice from a colleague. This allowed physicians to assess whether an AI-generated recommendation appeared reasonable before deciding whether to accept or override it.

However, the findings also highlight risks that come with greater AI integration. Overdependence on automated recommendations may gradually weaken independent clinical reasoning and increase the risk of automation bias. These findings reinforce the importance of designing AI systems that support active clinical reflection rather than passive acceptance, and that make the boundaries of the system's competence clearly visible to the user.

9.6 Methodological Strengths and Limitations

A major strength of this thesis is the mixed-methods design, which combined technical analyses of AI models with qualitative interviews with physicians. This made it possible to study explainable AI from both a technical and a clinical perspective, and to identify points of alignment and imbalance between what XAI methods currently produce and what clinicians actually need. A further strength is that several explainability methods were evaluated across both tabular and imaging models, allowing comparison across different approaches and data modalities.

Several limitations should also be considered. In the technical study, the dataset size was relatively limited, which may have restricted the ability of neural networks to fully learn non-linear relationships. The models were evaluated on a single dataset without external validation, limiting the generalisability of the results to other clinical populations. For the imaging analysis, only a single 2D PET slice per patient was used, meaning that potentially important three-dimensional spatial information was not captured. Correlated clinical features may also have influenced how attribution methods distributed feature importance, and no multimodal fusion model combining clinical and imaging data was implemented.

A further limitation concerns the explainability methods themselves. Post-hoc XAI methods provide approximations of model behaviour rather than direct representations of true internal decision-making, and the generated attributions should not be interpreted as evidence of causal relationships. As demonstrated particularly in the imaging models, explanations were sensitive to both method selection and implementation choices. Due to time constraints, the technical explainability results were not systematically evaluated together with physicians, which would have provided valuable additional insight into the clinical relevance of the generated explanations.

The interview study also has limitations. The number of participants was relatively small and represented only a limited selection of medical specialties. Some participants were already involved in AI-related projects or had a particular interest in AI, which may have influenced their perspectives on explainable AI systems. This could introduce some degree of selection bias and may limit the generalisability of the findings to the broader clinical population.

9.7 Future Perspectives

The findings of this thesis point to several directions for future research. From a technical perspective, multimodal models that combine structured clinical variables with medical imaging data in a single framework represent a natural next step, particularly given the finding that the two modalities showed complementary strengths in risk stratification. Larger datasets and external validation across different clinical populations are also needed to improve generalisability. The layer selection dilemma identified in the Grad-CAM analysis highlights a specific problem in imaging XAI and future work should investigate principled methods for selecting or aggregating layer-wise attributions in a way that produces stable and clinically interpretable results. Uncertainty estimation, which was raised as important by physicians in the interview study, also warrants further investigation in the context of survival prediction.

From a clinical perspective, the gap between technically detailed and clinically useful explanations identified in this thesis underscores the need for closer collaboration between clinicians and AI developers throughout the design process. Prospective evaluation of AI explanations in real clinical environments, ideally involving physician assessment of explanation quality alongside model performance, would provide insight that technical evaluations alone cannot. Increased focus on AI literacy and

education for healthcare professionals will also become important as AI systems become more deeply integrated into clinical workflows. Finally, regulatory and ethical frameworks governing transparency, accountability, and the conditions under which AI recommendations may be acted upon will play an increasingly role in determining how explainable AI systems are implemented in practice.

10

Conclusion

This thesis investigated explainable AI in healthcare from both a technical and a clinical perspective, examining how XAI methods perform in survival prediction for lymphoma patients and how physicians perceive and interpret AI explanations in clinical practice.

The two parts of the study converge on a common finding: the value of an AI explanation depends less on technical complexity and more on whether it supports meaningful clinical oversight. From a technical perspective, simpler and more transparent models performed competitively with more complex alternatives, while explanation consistency across methods was stronger for structured clinical data than for imaging data. From a clinical perspective, physicians did not require complete transparency into model internals. Instead, they wanted recommendations that were clinically motivated, inspectable, and actionable within the time constraints of real clinical practice.

Current XAI methods appear to fall short precisely at this intersection. Although explanations can be generated technically, they are not always clinically interpretable. In addition, explanations may vary substantially across methods and implementation choices, limiting their reliability and making them difficult to act upon in practice.

Together, the findings highlight both the potential and the current limitations of XAI in clinical decision support. While explainability methods can support trust, oversight, and clinical reasoning, their variability and sensitivity to implementation choices mean that they cannot yet be relied upon as a definitive basis for clinical decisions. Realising the full potential of explainable AI in healthcare will require not only more robust and standardised technical methods, but also ongoing collaboration between AI developers and clinicians to ensure that explanations are designed for the people who will ultimately use them.

Bibliography

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019. DOI: 10.1056/NEJMra1814259.

[2] G. Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017. DOI: 10.1016/j.media.2017.07.005.

[3] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B*, vol. 34, no. 2, pp. 187–220, 1972. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. DOI: 10.1016/j.inffus.2019.12.012.

[5] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[6] F. Jiang et al., "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017. DOI: 10.1136/svn-2017-000101.

[7] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017. DOI: 10.1038/nature21056.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. DOI: 10.1038/nature14539.

[10] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care-addressing ethical challenges," *New England Journal of Medicine*, vol. 378, no. 11, pp. 981–983, 2018. DOI: 10.1056/NEJMp1714229.

[11] D. G. Altman and J. M. Bland, "Time to event (survival) data," *BMJ*, vol. 317, no. 7156, pp. 468–469, 1998. DOI: 10.1136/bmj.317.7156.468.

[12] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part i: Basic concepts and first analyses," *British Journal of Cancer*, vol. 89, no. 2, pp. 232–238, 2003. DOI: 10.1038/sj.bjc.6601118.

[13] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. DOI: 10.1080/01621459.1958.10501452.

[14] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, 2018. DOI: 10.1186/s12874-018-0482-1.

[15] F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, vol. 15, no. 4, pp. 361–387, 1996. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

[16] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 195, 2019. DOI: 10.1186/s12916-019-1426-2.

[17] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018. DOI: 10.1145/3236386.3241340.

[18] A. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," *Nature Medicine*, vol. 25, no. 1, pp. 44–48, 2019.

[19] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[20] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision (ECCV)*, 2014.

[23] E. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44–56, 2019. DOI: 10.1038/s41591-018-0300-7.

[24] C. Davidson-Pilon, "Lifelines: Survival analysis in python," *Journal of Open Source Software*, vol. 4, no. 40, p. 1317, 2019. DOI: 10.21105/joss.01317.

[25] A. Paszke, S. Gross, F. Massa, et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8024–8035.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[27] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[28] N. Kokhlikyan et al., *Captum: A unified and generic model interpretability library for pytorch*, 2020. arXiv: 2009.07896 [cs.LG].

A

Original Interview Guide (Swedish)

The qualitative interviews were conducted using the semi-structured interview guide provided below. This appendix contains the original Swedish version used during the interviews.

Läkares perspektiv på förklarbar AI i kliniskt beslutsfattande: En kvalitativ intervjustudie i svensk vårdkontext

Intervjumall

Intervjudeltagare:

Förnamn Efternamn
Yrkesroll
Organisation
mail@mail.com

Bakgrund och klinisk kontext

Intervjun kommer att handla om ditt kliniska arbete, dina erfarenheter av beslutsfattande samt dina tankar kring AI-baserade beslutsstöd.

Vi kommer att börja med några korta frågor om din yrkesroll. Därefter går vi vidare till kliniskt beslutsfattande, digitala system och slutligen AI-baserade beslutsstöd samt olika typer av förklaringar som sådana system kan ge.

Kan du börja med att berätta om din yrkesroll och din kliniska kontext?

Möjliga följdfrågor:

- **Vilken specialitet arbetar du inom?**
- **I vilken vårdmiljö är du verksam?**
- **Hur länge har du arbetat kliniskt?**

Kliniskt beslutsfattande

Nu skulle vi vilja prata om hur du resonerar i ditt kliniska beslutsfattande, särskilt i situationer som upplevs som komplexa eller osäkra.

I vilka situationer upplever du att kliniskt beslutsfattande är som mest utmanande?

Kan du beskriva hur du brukar gå tillväga när du står inför ett beslut i ett komplext eller osäkert patientfall?

Möjliga följdfrågor:

- **Vilka resurser brukar du luta dig mot i sådana situationer?**
- **Hur brukar du resonera när du diskuterar ett fall med en kollega?**
- **Hur värderar du olika typer av information när de pekar åt olika håll?**

Digitala beslutsstöd

Nu när vi har pratat om hur du resonerar i ditt kliniska beslutsfattande skulle vi vilja gå vidare till att diskutera vilken roll digitala system spelar i ditt arbete. I den här delen fokuserar vi inte specifikt på AI, utan mer generellt på olika typer av tekniska hjälpmedel som stödjer eller påverkar dina beslut.

Vilka digitala system eller tekniska hjälpmedel använder du i ditt arbete som på något sätt påverkar eller stödjer ditt beslutsfattande?

Möjliga följdfrågor:

- **På vilket sätt påverkar dessa system ditt beslutsfattande?**
- **Vad fungerar bra med de digitala system du använder idag?**
- **Vad fungerar mindre bra?**

AI-baserade beslutsstöd

I vår studie fokuserar vi på AI som beslutsstöd, det vill säga system som analyserar patientdata, exempelvis provsvar, bilddiagnostik eller annan klinisk information, och ger någon form av rekommendation som stöd i kliniska beslut. Det kan till exempel handla om system som ger behandlingsrekommendationer, föreslår diagnoser eller gör riskbedömningar.

När du hör ordet ”AI” inom vården, vad tänker du på då?

Möjliga följdfrågor:

- **Vilka möjligheter ser du?**
- **Vilka utmaningar eller problem ser du?**

Har du någon erfarenhet av AI-baserade beslutsstöd i ditt kliniska arbete?

Möjliga följdfrågor:

- **Kan du beskriva hur systemet fungerade och hur det påverkade ditt beslutsfattande?**

Vad skulle krävas för att du ska känna dig trygg i att använda ett AI-baserat beslutsstöd?

I vilka situationer skulle du vara mer eller mindre benägen att ta hjälp av en AI-rekommendation?

Om en AI-rekommendation skulle skilja sig från din egen kliniska bedömning, hur tror du att du skulle resonera?

Förklarbar AI

Vissa AI-baserade beslutsstöd ger inte bara en rekommendation utan också en förklaring till hur systemet har kommit fram till sin bedömning. Det kan till exempel handla om vilka faktorer som har störst betydelse för resultatet eller hur systemet generellt sett resonerar.

Vi skulle nu vilja höra hur du ser på sådana förklaringar.

Ser du ett behov av förklaringar när AI-baserade beslutsstöd används kliniskt?

Möjliga följdfrågor:

- **I vilka situationer skulle en förklaring vara särskilt viktig?**
- **Finns det situationer där en förklaring skulle vara mindre viktig?**

Ser du några möjliga risker eller nackdelar med att visa förklaringar?

Vi kommer nu att visa några exempel på hur ett system kan förklara sin bedömning. För varje exempel är vi intresserade av din spontana reaktion och dina tankar.

Förklaringar baserade på patientegenskaper

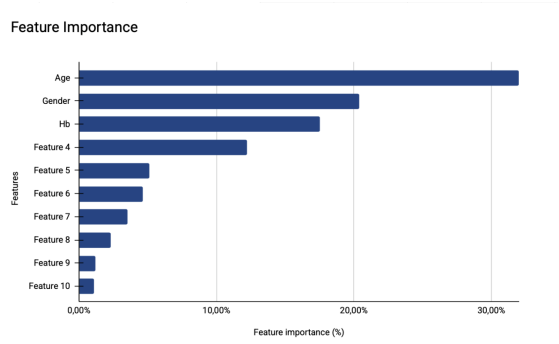
Föreställ dig att du använder ett AI-baserat beslutsstöd som analyserar kliniska data för att ge en riskbedömning eller behandlingsrekommendation. Bedömningen baseras på olika patientegenskaper, till exempel sjukdomshistorik, vitalparametrar och annan klinisk information. Utöver själva rekommendationen kan systemet också visa vilka faktorer som har störst betydelse och hur de påverkar resultatet.

Diagram 1 och 2

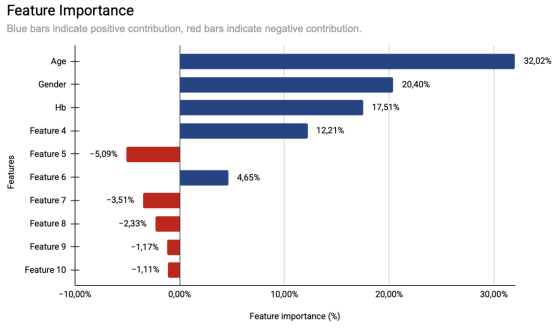
Diagram 1 och 2 visar hur olika patientegenskaper påverkar modellens bedömning för en specifik patient.

Diagram 1 visar framför allt hur olika faktorer förhåller sig till varandra, det vill säga vilka faktorer som har störst respektive minst påverkan.

Diagram 2 visar dessutom det exakta procentuella bidraget från varje egenskap och i vilken riktning egenskapen påverkar bedömningen. Blå staplar representerar faktorer som ökar sannolikheten för modellens bedömning, medan röda staplar representerar faktorer som talar emot systemets bedömning.



(a) Diagram 1



(b) Diagram 2

Vad tänker du spontant när du ser dessa två förklaringar?

Hur lätt eller svårt tycker du att det är att tolka dessa diagram?

Tycker du att det är värdefullt att se det procentuella bidraget från varje faktor?

Tycker du att det är värdefullt att se i vilken riktning egenskaperna påverkar bedömningen?

Diagram 3
Diagram 3 visar också hur olika patientegenskaper påverkar modellens bedömning, men visualiserar detta genom att visa hur olika faktorer flyttar bedömningen uppåt eller nedåt i förhållande till ett basvärde.

Modellen utgår från ett basvärde som representerar den genomsnittliga prediktionen i populationen.

Varje pil representerar en egenskap, till exempel ålder eller tumörstadium. Röda pilar visar faktorer som ökar sannolikheten för låg överlevnad, medan blå pilar visar faktorer som ökar sannolikheten för hög överlevnad.

Pilarnas längd anger hur stor påverkan varje faktor har, och bidragen summeras till modellens slutliga prediktion.

Diagram 3

Vad tänker du spontant när du ser den här förklaringen?

Hur lätt eller svårt tycker du att det är att tolka diagrammet?

Diagram 4
Diagram 4 visar en alternativ förklaring till hur modellen har kommit fram till sin bedömning.

Till vänster visas modellens prediktion i procent, det vill säga sannolikheten för hög respektive låg överlevnad.

I mitten visas vilka faktorer som påverkar modellens bedömning. Den blå sidan representerar faktorer som talar för hög överlevnad, medan den orangea sidan representerar faktorer som talar för låg överlevnad.

Till höger visas de faktiska värdena för patientens egenskaper.

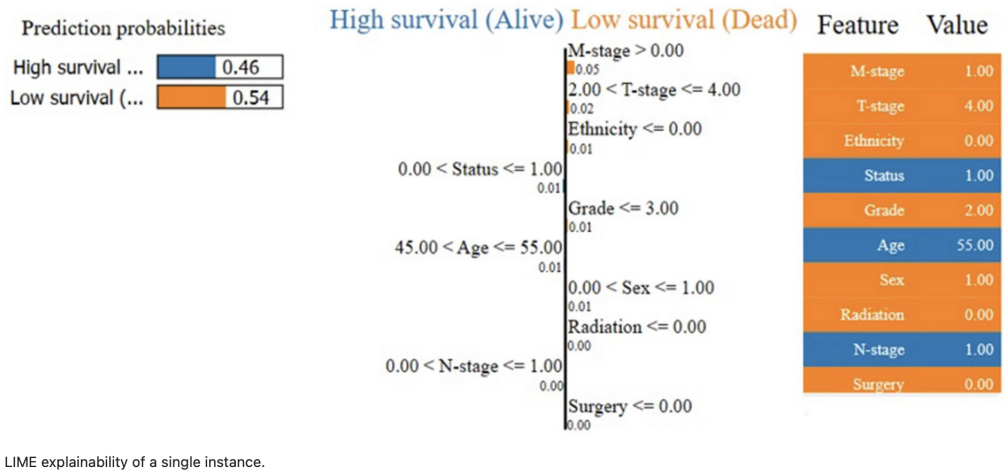


Diagram 4

Vad tänker du spontant när du ser den här förklaringen?

Hur lätt eller svårt tycker du att det är att tolka diagrammet?

Tycker du att det är värdefullt att se modellens prediktion i procent?

Tycker du att det är värdefullt att se patientens faktiska värden i ett diagram?

Förklaringar baserade på medicinska bilder

Vi kommer nu att lämna de tidigare förklaringarna som fokuserade på patientegenskaper och i stället titta på förklaringar baserade på medicinska bilder.

Föreställ dig att du använder ett AI-baserat beslutsstöd som analyserar medicinska bilder, till exempel röntgen-, CT- eller MR-bilder, och ger en bedömning, såsom tumörförekomst eller behandlingsrespons. Systemet analyserar bilden och ger en rekommendation baserad på bildens innehåll. Utöver själva bedömningen kan systemet också markera de områden i bilden som har störst betydelse för resultatet.

Bild 1

Bild 1 visar tre visualiseringar relaterade till bedömningen av en medicinsk bild.

I den första bilden är ett område av intresse manuellt markerat av en kliniker.

I den andra bilden visas en så kallad värmekarta, där färgerna indikerar vilka delar av bilden som har störst betydelse för systemets bedömning.

I den tredje bilden markerar systemet det område som identifieras som mest relevant för beslutet.

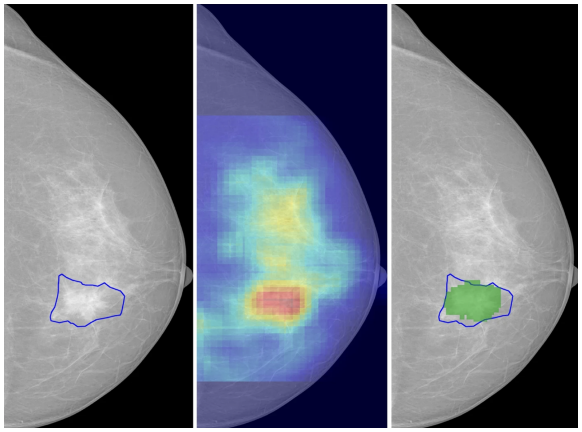


Bild 1

Vad tänker du spontant när du ser den här förklaringen?

Tycker du att bilderna hjälper dig att förstå systemets bedömning?

Bild 2

Bild 2 visar ytterligare ett exempel på hur ett AI-baserat beslutsstöd kan visualisera sin bedömning vid analys av röntgenbilder.

Den första bilden visar den ursprungliga röntgenbilden.

I den andra bilden visas samma originalbild tillsammans med en värmekarta som markerar de områden i bilden som har störst betydelse för bedömningen.

Den tredje bilden visar en liknande visualisering där de viktigaste områdena är markerade, men där originalbilden inte längre ligger i bakgrunden.

Den sista bilden visar en visualisering där röda pixlar indikerar områden som ökar sannolikheten för modellens bedömning, medan blå pixlar indikerar områden som talar emot bedömningen.

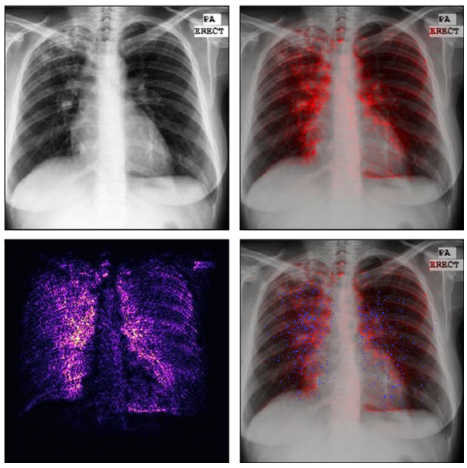


Bild 2

Vad tänker du spontant när du ser den här förklaringen?

Tycker du att bilderna hjälper dig att förstå systemets bedömning?

Tycker du att det är värdefullt att se en visualisering utan originalbilden i bakgrunden?

Tycker du att det är värdefullt att se vilka pixlar som bidrar positivt respektive negativt till systemets bedömning?

Förklaringar baserade på liknande patientfall

Vi skulle nu vilja prata om en annan typ av förklaring som bygger på jämförelser med tidigare patientfall.

Utöver själva rekommendationen visar systemet exempel på tidigare patienter med liknande kliniska egenskaper och deras utfall. Likheten kan till exempel baseras på diagnos, sjukdomsstadium eller annan relevant information.

Det skulle till exempel kunna handla om ett system som föreslår en behandling och samtidigt visar liknande patientfall, vilken behandling patienterna fick och deras utfall.

Vad tänker du spontant om en sådan förklaring?

Finns det situationer där sådana jämförelsefall skulle vara användbara?

Finns det situationer där en sådan typ av förklaring skulle kunna vara missvisande eller problematisk?

Generella Modellförklaringar

Vi skulle nu vilja avsluta med att prata om mer övergripande modellförklaringar.

Till skillnad från tidigare exempel, som handlade om bedömningen för en specifik patient, handlar detta om hur modellen fungerar generellt. Denna typ av information fokuserar alltså inte på den enskilda patienten, utan på hur systemet fungerar i stort.

Den här listan visar exempel på vad generella modellförklaringar skulle kunna innebära.

1. **Vilka egenskaper som generellt har störst betydelse**
   (t.ex. ålder, BMI och systoliskt blodtryck är de tre viktigaste egenskaperna)

2. **Hur olika egenskaper påverkar modellens prediktion**
   (t.ex. en ökad ålder innebär en högre predicerad risk)

3. **En enklare modell som efterliknar den komplexa modellen**
   (t.ex. ett beslutsträd som efterliknar modellen)

4. **Interaktioner mellan egenskaper**
   (t.ex. kombinationen av hög ålder och högt BMI leder till betydligt högre predicerad risk än varje variabel separat)

5. **Tolkbara beslutsregler från modellen**
   (t.ex. om ålder > 60 och systoliskt blodtryck $> 140 \rightarrow$ hög risk)

6. **För vilka patientgrupper modellen har utvecklats eller validerats**
   (t.ex. modellen är tränad på patienter mellan 40–80 år)

Vilken information i listan upplever du som mest relevant?

Vilken information i listan upplever du som mindre relevant?

Hur viktig tycker du att sådan övergripande information är?

Finns det något utöver informationen i listan som du skulle vilja veta om systemet?

Avslutande Reflektion

Vi skulle nu vilja avsluta intervjun med några frågor om träffsäkerhet hos AI-baserade beslutsstöd.

Med träffsäkerhet menar vi hur ofta systemets rekommendation faktiskt är korrekt. En hög träffsäkerhet skulle till exempel kunna innebära att systemet ger en korrekt bedömning i 96 % av fallen, medan en lägre träffsäkerhet skulle kunna innebära att systemet ger en korrekt bedömning i 82 % av fallen.

Om ett system har en mycket hög träffsäkerhet men inte ger någon förklaring, hur skulle det påverka din tillit till systemet?

Om ett system i stället har en något lägre träffsäkerhet men ger en tydlig förklaring, hur skulle det påverka din tillit till systemet?

Avslutningsvis

Finns det något annat du skulle vilja lyfta kring AI-baserade beslutsstöd eller förklaringar?

B

Translated Interview Guide (English)

This appendix contains an English translation of the interview guide presented in Appendix A.

Physicians' perspectives on explainable AI in clinical decision-making: A qualitative interview study in a swedish healthcare context

Interview Guide

Interview participant:

First & Last name
Professional role
Organisation
mail@mail.com

Background and Clinical Context

The interview will focus on your clinical work, your experiences with decision-making, and your thoughts on AI-based decision support.

We will begin with a few brief questions about your professional role. We will then move on to clinical decision-making, digital systems, and finally AI-based decision support and the different types of explanations such systems can provide.

Can you start by telling us about your professional role and clinical context?

Possible follow-up questions:

- **Which specialty do you work in?**
- **In what healthcare setting do you work?**
- **How long have you been working clinically?**

Clinical Decision-Making

We would now like to talk about how you typically reason when making clinical decisions, particularly in complex or uncertain situations.

In which situations do you find clinical decision-making to be the most challenging?

Could you describe how you typically approach a decision in a complex or uncertain patient case?

Possible follow-up questions:

- **What resources do you typically rely on in such situations?**
- **How do you typically reason when discussing a case with a colleague?**
- **How do you weigh different types of information when they point in different directions?**

Digital Decision Support

Now that we have talked about how you reason in your clinical decision-making, we would like to move on to discuss the role that digital systems play in your work. In this section, we are not focusing specifically on AI, but more generally on different types of technical tools that support or influence your decisions.

Which digital systems or technical tools do you use in your work that in some way influence or support your decision-making?

Possible follow-up questions:

- **In what way do these systems influence your decision-making?**
- **What works well with the digital systems you use today?**
- **What works less well?**

AI-Based Decision Support

In our study, we focus on AI as decision support. This refers to systems that analyse patient data, such as test results, medical imaging, or other clinical information, and provide some form of recommendation to support clinical decisions. This could include systems that give treatment recommendations, suggest diagnoses, or perform risk assessments.

When you hear the word "AI" in healthcare, what comes to mind?

Possible follow-up questions:

- **What opportunities do you see?**
- **What challenges or problems do you see?**

Do you have any experience with AI-based decision support in your clinical work?

Possible follow-up questions:

- **Could you describe how the system worked and how it influenced your decision-making?**

What would be required for you to feel comfortable using an AI-based decision support system?

In which situations would you be more or less likely to rely on an AI recommendation?

If an AI recommendation were to differ from your own clinical assessment, how do you think you would reason

Explainable AI

Some AI-based decision support systems not only provide a recommendation but also an explanation of how the system arrived at its assessment. This could include, for example, which factors had the greatest influence on the outcome, or how the system reasons in general.

We would now like to hear your views on such explanations.

Do you see a need for explanations when AI-based decision support is used in clinical practice?

Possible follow-up questions:

- **In which situations would an explanation be particularly important?**
- **Are there situations where an explanation would be less important?**

Do you see any potential risks or disadvantages in showing explanations?

We will now show some examples of how a system can explain its assessment. For each example, we are interested in your spontaneous reaction and thoughts.

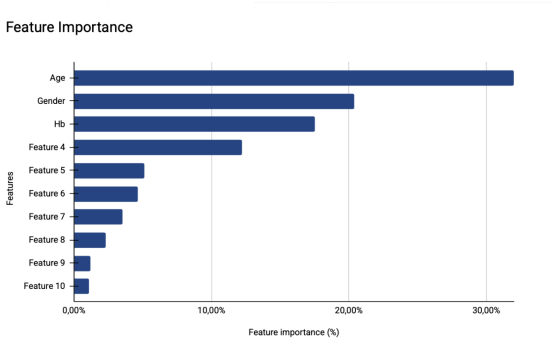
Explanations Based on Patient Characteristics

Imagine that you are using an AI-based decision support system that analyses clinical data to provide a risk assessment or treatment recommendation. The assessment is based on different patient characteristics, such as medical history, vital parameters, and other clinical information. In addition to the recommendation itself, the system can also show which factors had the greatest influence and how they affected the outcome.

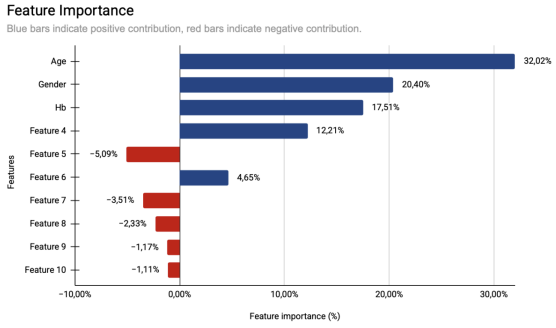
Diagram 1 and 2
Diagrams 1 and 2 show how different patient characteristics influence the model's assessment for a specific patient.

Diagram 1 mainly shows how different factors relate to each other, meaning which factors had the greatest and the least influence.

Diagram 2 additionally shows the exact percentage contribution from each characteristic and the direction in which each characteristic influences the assessment. Blue bars represent factors that increase the probability of the model's assessment, while red bars represent factors that argue against it.



(a) Diagram 1



(b) Diagram 2

What are your initial thoughts when you see these two explanations?

How easy or difficult do you think it is to interpret these diagrams?

Do you find it valuable to see the percentage contribution from each factor?

Do you find it valuable to see in which direction each feature influenced the assessment?

Diagram 3

Diagram 3 also shows how different patient characteristics influence the model's assessment, but visualises this by showing how different factors shift the assessment upwards or downwards relative to a baseline value.

The model starts from a baseline value representing the average prediction in the population.

Each arrow represents a characteristic, such as age or tumour stage. Red arrows indicate factors that increase the probability of low survival, while blue arrows indicate factors that increase the probability of high survival.

The length of the arrows indicates how large an influence each factor has, and the contributions are summed to produce the model's final prediction.

Diagram 3

What are your initial thoughts when you see this explanation?

How easy or difficult do you think it is to interpret the diagram?

Diagram 4

Diagram 4 presents an alternative explanation of how the model arrived at its assessment.

On the left, the model's prediction is displayed as percentages, indicating the probability of high vs. low survival.

In the middle, the factors that influenced the model's decision are shown. The blue side represents factors that support high survival, while the orange side represents factors that support low survival.

On the right, the patient's actual values for each characteristic are shown.

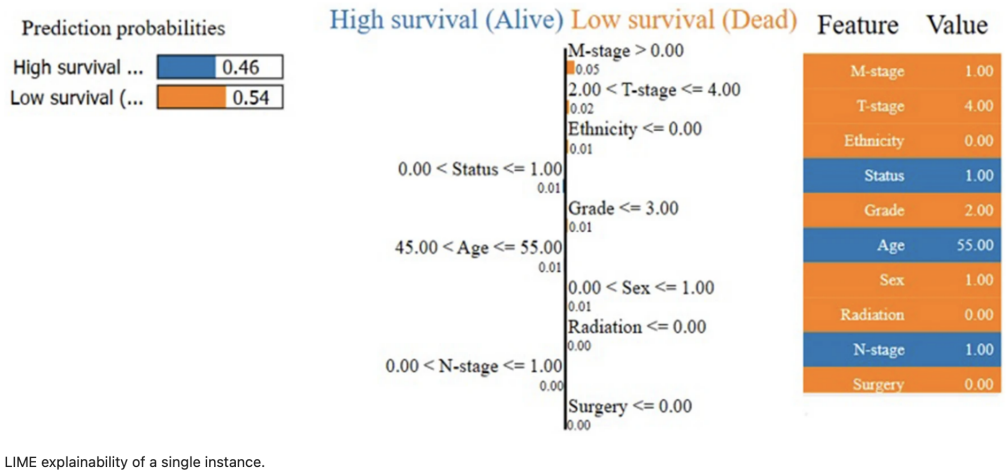


Diagram 4

What are your initial thoughts when you see this explanation?

How easy or difficult do you think it is to interpret the diagram?

Do you find it valuable to see the model's prediction as a percentage?

Do you find it valuable to see the patient's actual values within the diagram?

Explanations Based on Medical Images

We will now leave the previous explanations that focused on patient characteristics and instead look at explanations based on medical images.

Imagine that you are using an AI-based decision support system that analyses medical images, such as X-ray, CT, or MRI images, and provides an assessment such as tumor presence or treatment response. The system analyses the image and provides a recommendation based on its content. In addition to the assessment itself, the system can also highlight the areas in the image that had the greatest influence on the outcome.

Image 1

Image 1 shows three visualisations related to the assessment of a medical image.

In the first image, a region of interest has been manually marked by a clinician.

In the second image, a so-called heatmap is shown, where colours indicate which parts of the image had the greatest influence on the system's assessment.

In the third image, the system marks the area identified as most relevant for the decision.

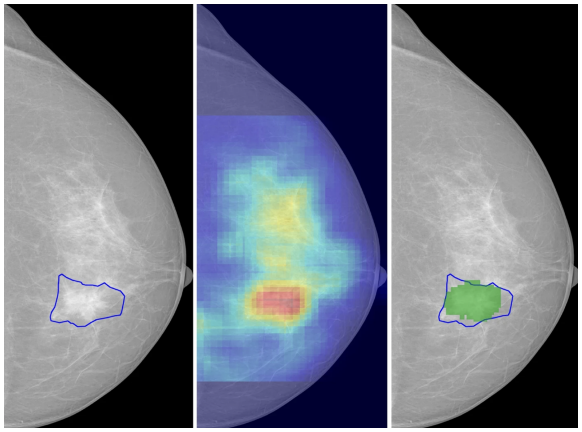


Image 1

What are your initial thoughts when you see this explanation?

Do you think the images help you understand the system's assessment?

Image 2

Image 2 provides another example of how an AI-based decision support system can visualise its assessment when analysing X-ray images.

The first image shows the original X-ray.

The second image shows the same original image together with a heatmap highlighting the areas of the image that had the greatest influence on the assessment.

The third image shows a similar visualisation where the most important areas are marked, but where the original image is no longer visible in the background.

The final image shows a visualisation where red pixels indicate areas that increase the probability of the model's assessment, while blue pixels indicate areas that argue against it.

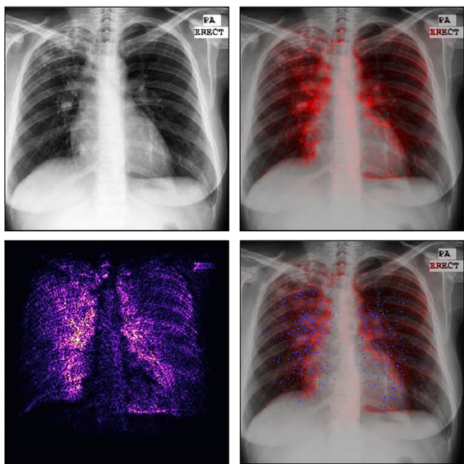


Image 2

What are your initial thoughts when you see this explanation

Do you think the images help you understand the system's assessment?

Do you find it valuable to see a visualisation without the original image in the background?

Do you find it valuable to see which pixels contribute positively and negatively to the system's assessment?

Explanations Based on Similar Patient Cases

We would now like to talk about a different type of explanation that is based on comparisons with previous patient cases.

In addition to the recommendation itself, the system shows examples of previous patients with similar clinical characteristics and their outcomes. The similarity could be based on, for example, diagnosis, disease stage, or other relevant information.

For example, this could involve a system that proposes a treatment and simultaneously shows similar patient cases, the treatment those patients received, and their outcomes.

What are your initial thoughts about this type of explanation

Are there situations in which such comparison cases would be useful?

Are there situations in which such a type of explanation could be misleading or problematic?

General Model Explanations

We would now like to finish by talking about more general model explanations.

Unlike the previous examples, which concerned the assessment for a specific patient, this is about how the model functions in general. This type of information therefore does not focus on the individual patient, but on how the system works overall.

The following list shows examples of what general model explanations could include.

1. **Which features generally have the greatest importance**
   (e.g. age, BMI, and systolic blood pressure are the three most important features)

2. **How different features influence the model's prediction**
   (e.g. increasing age leads to a higher predicted risk)

3. **A simpler model that approximates the complex model**
   (e.g. a decision tree that mimics the model)

4. **Interactions between features**
   (e.g. the combination of high age and high BMI leads to a much higher predicted risk than each feature on its own)

5. **Interpretable decision rules from the model**
   (e.g. if age > 60 and systolic blood pressure $> 140 \rightarrow$ high risk)

6. **Which patient groups the model was developed or validated for**
   (e.g. the model was trained on patients aged 40–80)

Which information in the list do you find most relevant?

Which information in the list do you find less relevant?

How important do you think this type of general information is?

Is there anything beyond the information in the list that you would like to know about the system?

Final Reflection

We would now like to conclude the interview with some questions about the accuracy of AI-based decision support systems.

By accuracy, we mean how often the system's recommendation is actually correct. A high accuracy might mean, for example, that the system provides a correct assessment in 96 % of cases, while a lower accuracy might mean it provides a correct assessment in 82 % of cases.

If a system has a very high accuracy but provides no explanation, how would that affect your trust in the system?

If a system instead has a somewhat lower accuracy but provides a clear explanation, how would that affect your trust in the system?

Finally

Is there anything else you would like to add regarding AI-based decision support or explanations?